A revised genome annotation of the model cyanobacterium *Synechocystis* based on start and stop codon-enriched ribosome profiling and proteogenomics

Lydia Hadjeras[1#], Vanessa Krauspe[2#], Rick Gelhausen[3#], Benjamin Heiniger[4,5], Philipp Spät[6], Viktoria Reimann[2], Garance Jaques[4], Paul Minges[2], Raphael Bilger[2], Maren Gerstner[2], Boris Maček[6], Christian H. Ahrens[4,*], Rolf Backofen[3,7], Cynthia M. Sharma[1] and Wolfgang R. Hess[2,*]

[1]University of Würzburg, Institute of Molecular Infection Biology (IMIB), University of Würzburg, Germany;

[2]Genetics and Experimental Bioinformatics, Faculty of Biology, University of Freiburg, Germany;

[3]Bioinformatics Group, Department of Computer Science, Technical Faculty, University of Freiburg, Germany;

[4]Molecular Ecology, Agroscope and SIB Swiss Institute of Bioinformatics, 8046 Zurich, Switzerland;

[5]PhD Program in Systems Biology, Zurich Life Science Graduate School

[6]Department of Quantitative Proteomics, Interfaculty Institute for Cell Biology, University of Tübingen, D-72076 Tübingen, Germany;

[7]Signalling Research Centre CIBSS, University of Freiburg, Germany

[#]Co-sharing first authors

*Corresponding authors: Christian H. Ahrens: christian.ahrens@agroscope.admin.ch, Wolfgang R. Hess: wolfgang.hess@biologie.uni-freiburg.de

Lead contact: Wolfgang R. Hess, wolfgang.hess@biologie.uni-freiburg.de

Running title: TIS- and TTS-Ribo-seq in *Synechocystis* sp. PCC 6803

ABSTRACT

Cyanobacteria are important primary producers and are used in biotechnology as microbial cell factories due to their ability to use solar light for oxygenic photosynthesis. *Synechocystis* sp. PCC 6803 is a popular model cyanobacterium, yet there are ambiguities in the precise coding regions of many genes, and numerous genes encoding small proteins have remained undetected. Here we present the results of a ribosome profiling (Ribo-seq) analysis involving inhibitors that stall ribosomes at translation initiation and termination sites (TIS- and TTS-Ribo-seq), combined with a proteogenomic reevaluation and reannotation of its entire genome. We report evidence for the translation of 3,050 annotated genes based on proteogenomics (83%), of 3,492 based on Ribo-seq (95.2%), and of 3,009 supported by both methods (82%). The data suggested both novel protein-coding genes and corrections for annotated ones. We validated 15 novel small proteins translated from antisense RNAs, from intergenic and intragenic regions and identified 69 novel, mostly small proteins based on proteogenomics. With *slr0489, slr1079* and *slr1082* we identified three genes with ~300 nt long intragenic out-of-frame coding regions and show that both the internal and host reading frames are translated. The resulting proteins interact with each other, resembling certain defense or toxin-antitoxin systems. Our data illustrate the enormous value of consolidating genome annotations in the context of integrated experimental data and suggest that genome annotations in general need to be extended and revised. All of our data can be accessed via an intuitive and interactive genome browser platform at https://www.bioinf.uni-freiburg.de/~ribobase/.

INTRODUCTION

Oxygenic photosynthesis powers photoautotrophic organisms such as cyanobacteria enabling them to build complex organic molecules from inorganic carbon. As primary producers, cyanobacteria have a pivotal ecological significance[1] and considerable value in biotechnological applications[2–4]. Among cyanobacteria, *Synechocystis* sp. PCC 6803 (from here: *Synechocystis* 6803), the third organism to have its complete genome sequenced[5], is considered one of the most advanced models. However, despite careful annotation attempts, severe annotation ambiguities still exist.

Cyanobacteria, like other prokaryotes, harbor a higher share of genes encoding small proteins compared to eukaryotes due to their more compact genome architecture[6]. Importantly, small protein-coding genes are frequently missed during standard genome annotation which aims to minimize the over-prediction of spurious short ORFs. Moreover, standard mass spectrometry-based approaches fail to comprehensively detect this class of proteins[7], making small proteins a systematically underestimated component of the total proteome. Small proteins (here defined as ≤70 amino acids (aa)) are often considered too small to exert enzymatic activity. Nevertheless, many small proteins were demonstrated to interact with larger proteins or multi-protein complexes and have important functions[8]. Examples are the One-Helix Proteins in the assembly or stabilization of photosynthetic pigment-protein complexes in plant chloroplasts and cyanobacteria[9,10], small structural proteins in photosystem (PS) I[11], from which PsaJ (40 aa) is essential for the stable assembly and structural integrity of the entire complex[12]. Due to the extensive previous work on the photosynthetic apparatus, cyanobacteria provide a paradigm for small protein functions: Nineteen well-characterized proteins smaller than 50 aa are associated with the photosynthetic machinery alone[13] . Globally, small proteins operate as modulators of larger proteins' enzymatic activities[8,14–16], such as the 48 aa protein AtpΘ, which inhibits the futile reverse reaction of ATP synthase under unfavorable conditions[17]. Lastly, many small proteins are toxic, including those that are part of type I toxin–antitoxin systems[18], or antimicrobial peptides (AMPs)[19].

Ribosome profiling (Ribo-seq) detects translated mRNA segments on a genome-wide level[20]. Following cellular lysis and nuclease treatment, intact RNA associated with proteins – including the mRNA engaged with ribosomes during translation, is separated from cell debris by sucrose gradient ultracentrifugation followed by recovery of the associated RNA, cDNA synthesis and sequencing[21]. The analysis of RNA associated with the bacterial 70S ribosomes yields ribosome footprints with a length of 15–40 nt, peaking around 24 nt for *E. coli* [22]. Through comparison with an RNA-seq library generated in parallel from total RNA harvested from the same culture and fragmented into similar size, the coverage by sequenced ribosome footprints then allows to infer ribosome occupancy, open reading frames (ORFs) boundaries, and translational efficiency. While Ribo-seq was initially developed to study translational regulation[23–25], the discovery of extensive, previously unrecognized, coding potential

in genomes of diverse organisms, such as yeast[26], vertebrates including humans[27,28], plants[29], bacteria[30–32], and viruses[33] was another major benefit. The identification of potential short ORF encoded proteins (SEPs) catalogued by Ribo-seq in diverse organisms is gaining momentum. After the adaptation of the Ribo-seq workflow, this has only recently been extended to a range of bacteria and archaea[32,34–36].

The antibiotic retapamulin, an inhibitor of the initiation of translation in bacteria, has revolutionized the view on the small proteome through Ribo-seq. Retapamulin (a pleuromutilin antibiotic) stalls ribosomes in treated cells at the translation initiation site (TIS), while elongating ribosomes finish the translation and run off the mRNA releasing the nascent peptide. Thus, retapamulin-assisted TIS-Ribo-seq results in defined peaks for translated proteins at their respective translational start sites, which allows to detect both unknown internal proteins, alternative versions of known proteins and unannotated, novel small proteins[30,37]. Further improvements have been obtained by the introduction of Apidaecin, a proline-rich antimicrobial peptide, which arrests translation close to the stop codon[38] and leads to the accumulation of ribosome footprints close to the 3′ ends of coding regions[39], and thus can be used for global mapping of stop codons via translation termination site (TTS)-Ribo-seq[32,40].

Combining TIS- and TTS-Ribo-seq data can thus yield the precise borders of a translated sequence from start to stop[32,40]. These recent methodological breakthroughs applied alone have already yielded unprecedented insight not only into the small proteome of bacteria, but also suggested an unexpectedly high number of proteins translated from regions within annotated larger genes and even in antisense orientation[15,30–32,34,35,40,41]. However, thus far these approaches have only been applied to a small number of model species, and even more rarely as powerful combination of TIS-Ribo-seq and TTS-Ribo-seq. Given the large diversity among prokaryotes, substantial insight can be expected from describing the small and alternative proteome of species from diverse taxonomic ranks.

Previously collected Ribo-seq data for wild type and carbon-starved cells of *Synechocystis* 6803 showed allocation of ribosomes in carbon depletion[42] and helped to characterize the protein turnover at different time points of the light-dark cycle[43]. To our knowledge, there are no TIS- or TTS-Ribo-seq datasets at all and no Ribo-seq datasets for any other cyanobacterium available.

Altogether, 165 genes encoding small proteins ≤70 aa are annotated (Kazusa (KZS) annotation) in the *Synechocystis* 6803 genome (4.5% of all genes). We have recently characterized five additional small proteins NblD, AtpΘ, AcnSP, SliP4 and NirP1 that are functionally relevant in the energy and nutrient metabolism of *Synechocystis* 6803[9,17,44–46] demonstrating the immense value of fully unraveling the composition and functional relevance of the small and alternative proteome.

Here, we report the results of TIS-, TTS-Ribo-seq and proteogenomic analyses in the model cyanobacterium *Synechocystis* 6803. We present evidence for the translation of 3,492 and 3,050 annotated genes based on Ribo-seq and proteogenomics, and of 1,069 and 58 previously unknown ORFs, including 935 SEPs of 70 aa or less. We detected an instance of ribosome frameshifting and small proteins translated from intragenic out-of-frame alternative reading frames and from antisense RNAs (asRNAs). We integrated our Ribo-seq data with the annotations from several reference databases, performed a reanalysis of previously generated proteomic datasets[47], and *ab initio* gene predictions. This integrated view of several annotations overlaid with experimental evidence allowed to consolidate many of the genome annotation differences. The proteogenomic approach, based on a search database that captures these differences and novel predictions[48] also validated some Ribo-Seq candidates and identified additional small and larger novel proteins. The consolidated data provide a significantly updated and extended annotation of the *Synechocystis* 6803 genome that is available for the community through an easily accessible interactive web-based genome-browser at https://www.bioinf.uni-freiburg.de/~ribobase/. The here developed generic approach provides an optimal basis to better exploit the genome sequence information; we expect it to be highly beneficial for many other model organisms and research communities.

RESULTS

Ribo-seq and retapamulin-based TIS-Ribo-seq captures the translatome of *Synechocystis* 6803

To globally measure translation of annotated ORFs and to discover potential new short ORF encoded proteins (SEPs), we applied Ribo-seq detecting actively translated mRNA segments to *Synechocystis* 6803. In parallel to this standard Ribo-seq approach, we added retapamulin to stall ribosomes on the start codon positions (set-up in **Figure 1A**) for TIS-Ribo-seq. Cells were rapidly harvested to halt ribosomes along ORFs and preserve polysomes, and lysed, followed by converting polysomes into monosomes and trimming of ribosome footprints by MNase. Polysome profiles generated by sucrose density gradient centrifugation showed successful capture of translating ribosomes as well as their controlled digestion into monosomes (**Figure 1A**). Following the results of our pre-experiments (**Figure S1**), we used a concentration of 10 µg/mL retapamulin in 100 mL *Synechocystis* 6803 wild type cultures for TIS-Ribo-seq. We found that retapamulin treatment collapsed polysomes to monosomes even without micrococcal nuclease (MNase) digestion (**Figure 1A**), as reported for *E. coli*[30], validating our conditions for TIS profiling.

The mapped reads from the standard Ribo-seq analysis revealed the transcriptome-wide ribosome occupancy. Comparison to the RNA-seq coverage inferred in parallel allowed us to calculate the translation efficiency (TE) for a particular transcript (ratio Ribo-seq/RNA-seq coverage). The comparison of TEs between different annotated gene classes showed a much higher mean TE for protein-coding sequences than for sRNAs or 5′ and 3′ UTR sequences (**Figure 1B**).

Typical examples for the obtained read profiles are shown for gene *psaM* encoding the small PSI subunit XII protein (**Figure 1C**) and the tricistronic *cpcBAC2* genes encoding the phycobilisome subunits CpcB, CpcA and the rod linker polypeptide CpcC2 (**Figure 1D**). The profiles show transcriptome coverage starting at the transcriptional initiation sites (TSS) and ending within the 3′UTR downstream the coding region. In contrast, the Ribo-seq read coverage is confined to the coding sections. As reported for MNase-generated Ribo-seq libraries in other bacteria[22,32], the mRNA fragments protected by ribosomes showed a much cleaner, more defined cutoff at the 3′ ends, where the ribosomes' footprints ended compared to the 5′ ends,

where the footprints started (**Figure S2**). As examples, we show the sharp 3′ ends observed 19 nt downstream the respective annotated start codons of *psaM* (**Figure 1C**) and the *cpcB, cpcA* and *cpcC2* genes (**Figure 1D**). To generalize these observations, we performed a meta-analysis of footprints enriched in TIS-Ribo-seq near annotated start codons (ATG, TTG, GTG). This revealed an average length of the footprints of retapamulin-stalled ribosomes of 32 nt, with the vast majority of their 3′ ends 19 nt downstream the annotated start codons (**Figure 1E**). A parallel metagene analysis focusing on short ORFs (sORFs; ≤50 nt) was conducted to confirm that translation initiation patterns in these shorter sequences did not differ substantially from those observed in longer ORFs (**Figure S2**).

Minor peaks for the footprint 3′ ends were observed 2 nt and 4 nt downstream the annotated start codons. As an example, we have zoomed in the region around the start codon of *rpl34*, a house-keeping gene encoding the large ribosomal subunit protein L34. For control, the 3′ ends of cDNA fragments generated from standard Ribo-Seq (in the absence of retapamulin), are plotted for all three examples (**Figure 1C, D, E**).

The nature and conservation of possible ribosome binding sites for the initiation of translation in cyanobacteria is not known. Therefore, we searched for potentially conserved nucleotides 30 nt upstream and 10 nt downstream of the start codons, focusing on those genes contributing to the read density at the specific offsets (+2,+4,+19) and filtered them using a 10 read cutoff. This revealed a clear preference for ATG over GTG as start codon, and a slight preference for 2 preceding pyrimidine residues. A purine-rich motif that would indicate a likely ribosome binding site was found in only a subset of genes (**Figure 1F)**.

Overall, metagene analysis and single-gene inspection suggested successful set-up of TIS-Ribo-seq profiling in *Synechocystis* 6803, an average footprint length of retapamulin-stalled ribosomes of 32 nt, with their 3′ ends mapping with an offset of 2 nt and 4 nt, but mainly 19 nt downstream the annotated start codons. A purine-rich motif suitable as ribosome binding site was observed only for a subset of genes (**Figure 1F**).

Establishing TTS-Ribo-seq in a cyanobacterium

While the location of stop codons of annotated protein-coding genes is uncontroversial, their comprehensive analysis and precise localization is highly valuable for accurately calling previously overlooked translated reading frames. Translation termination site (TTS)-Ribo-seq is a method to infer stop codon positions based on the accumulation of ribosome footprints in the presence of inhibitors of termination[32,39,40]. This method has not been established yet for cyanobacteria. To facilitate TTS-Ribo-seq, we used the inhibitor Api137 that traps the release factors during translation termination, impeding the natural process of peptide liberation[38]. The experimental set-up is shown in **Figure 2A**. In contrast to the polysome depletion in retapamulin-treated cultures (**Figure 1A**), the gradient profile of Api-treated cells did not change much in terms of polysome abundance (**Figure 2A**). Remaining disome peaks were also observed quite frequently in samples digested with MNase, even without Api137. We reasoned that this was most probably due to missing space between two adjacent ribosomes which would be necessary for the digestion of mRNAs, an effect that appeared pronounced in the presence of Api-137 stalling ribosomes at termination codons.

The 3′ ends of sequenced footprints from treated samples compared to the Ribo-seq control were mapped for the TTS-Ribo-seq analysis. As examples, we show the sharp 3′ ends observed 14 nt downstream of the respective stop codons of *norf1* (**Figure 2B**), and the dicistronic *slr1648* and *ssr2754* genes encoding the carotenoid oxygenase SynDiox2[49], and a YefM-type antitoxin[50] (**Figure 2C**). Note that the inclusion of TIS-Ribo-seq data allowed the precise definition of the *norf1* reading frame (**Figure 2B**) that had not been annotated in any of the used reference genome annotations (KZS, GBK, RefSeq; see Methods). This gene has meanwhile been identified to encode the 48 aa ATP synthase inhibitor AtpΘ[17].

Meta-analysis of footprints enriched in TTS-Ribo-seq from isolated monosomes revealed that the majority of their 3′ ends mapped 14 nt downstream of the respective stop codons (**Figure 2D**). For illustration, we show the stop codon region of *petB* encoding the cytochrome b6 protein involved in photosynthetic electron transport (**Figure 2D**). The analysis of cDNA 3′ ends enriched in TTS-Ribo-seq from prepared disomes revealed peaks mapping 14 nt downstream of the respective stop codons as

well (**Figure 2E**), exemplified for *psbN* encoding the 43 aa photosystem II protein PsbN.

To summarize, the 3′ ends of the footprints obtained from Api137-stalled ribosomes were observed on average 14 nt downstream of the respective stop codons. Metagene analysis and single-gene inspection suggested successful set-up of TTS-Ribo-seq profiling in *Synechocystis* 6803.

Validation by proteogenomic analysis and database visualization

The combination and filtering of all TIS- TTS- and classical Ribo-seq data yielded 2,711 instances of potentially novel protein-coding regions or suggested corrections of annotated genes (1,069 unannotated, 1,190 Internal-Out of-Frame (IOF), 305 Internal-In-Frame (IIF), 88 longer and 59 shorter proteoforms). The majority of these instances were short protein-coding genes (**Figure S3, Table S1, S2**). After manual inspection of 808 instances from this list (30%) and based on the Ribo-seq coverage, we considered 148 of those to be correct (18%). Among these 148 candidates, we found 80 unannotated CDS, 46 IOF, 9 IIF, 6 N-terminal extensions and 7 possible N-terminal truncations. 114/148 (77%) encode proteins of 70 or less aa.

Because manual curation is subjective, we attempted to also leverage the large amount of proteomics data acquired for *Synechocystis* 6803[47]. For this, we created integrated proteogenomics databases (iPtgxDBs) based on the *Synechocystis* 6803 chromosome and plasmid sequences and searched publicly available mass spectrometry datasets[47] against them (**Figure S4**). For the initial Ribo-seq data mapping, we used an in-house curated genome sequence based on the KZS annotation that contained all previously mapped sRNAs, transcriptional units, 5′ and 3′ UTRs that were defined in previous comparative differential RNA-seq analyses over ten different growth conditions[51]. Moreover, to distinguish differences in gene calling and annotation, we integrated the annotations according to NCBI RefSeq 2022 (ID: GCF_000009725.1) (2022) and GenBank (2016). Despite being based on the same genome sequence, these differed substantially (**Figure S5**). Furthermore, to capture the protein-coding potential comprehensively[48], we also included two *ab initio* gene predictions (Prodigal 2.6.3[52] and ChemGenome 2.1[53]), and an *in-silico* ORF prediction

that considers the main alternative start codons. These were hierarchically integrated into a standard iPtgxDB[48] and a smaller custom iPtgxDB (see Methods) that are accessible via the platform at https://www.bioinf.uni-freiburg.de/~ribobase/.

Altogether, from the list of 165 KZS-annotated small proteins (**Table S3**), 138 (83.6%) were considered expressed in our dataset based on the Ribo-seq data (RIBO-WT-avg_TE was ≥0.288), of which 97 had support from TIS-Ribo-seq. In addition, peptide evidence was observed in the proteogenomic searches for 63 of the 165 SEPs (38.2%) (**Table S3**), all of which had some form of Ribo-seq support. Previously, Baers et al. had validated 47 small proteins[54]. Our proteogenomic dataset missed 6 of those (*ssl0312, ssr0854, ssl0511, hliD, ssl5096* and *MYO_RS19000*), which however, were all covered by the Ribo-seq data. Thus, based on our combined data, 138 annotated SEPs are translated, almost tripling the previously available experimental evidence for 47 annotated and validated SEPs. Among the 27 remaining sORFs lacking evidence for translation, 12 are pseudogenes, mainly fragments of transposase genes.

To identify novel proteins or those unique to one of the annotations, we performed searches against the standard and custom iPtgxDB. Overall, 11 proteins were identified that were annotated in KZS and/or GenBank, but not in the latest RefSeq annotation, i.e., examples of CDS that got wrongly removed from the genome annotation (**Figure S6**). Searches against both iPtgxDBs identified 7 of these 11 proteins as novel (compared to RefSeq) that were annotated by KZS and GBK: 4 of these were classified as pseudogenes in RefSeq and 3 were removed (one being 4 amino acids shorter in KZS compared to GBK). Both iPtgxDBs uniquely identified one KZS-only annotated protein missing from RefSeq, while the custom iPtgxDB identified 3 more GBK-only annotated proteins, two of which were annotated as pseudogenes in RefSeq. The other 58 proteins identified through searches of the iPtgxDBs were entirely new (**Figure S6**). The custom iPtgxDB uniquely uncovered 22 proteins with Ribo-seq evidence, while the standard iPtgxDB uniquely found 27 *in-silico* ORFs (not contained in the custom iPtgxDB), and 6 were predicted by both. Finally, two predictions from ChemGenome and one from Prodigal were only found using the much larger standard iPtgxDB, again illustrating the value of integrating several annotations[48].

TIS- and TTS-Ribo-seq permit start codon re-annotation and the detection of ribosome frameshifting

We observed that the annotation of reading frames and their respective start codons frequently differs between the considered annotation files, although these were based on identical genome sequences. This can cause substantial problems if proteins of interest are expressed only in a truncated version or contain unannotated N-terminal extensions. These problems can be resolved by considering the TIS- and TTS-Ribo-seq data and consolidating the different genome annotations on a case-by-case basis.

We illustrate this here for the gene *sll0553* encoding an alpha/beta fold hydrolase, that is widely conserved in cyanobacteria. The annotated longest gene model in GBK is 1,038 nt, encoding a 345 aa homolog with a GTG start codon (WP_010873707.1), while the corresponding gene model in the KZS annotation suggests a 306 aa protein. In contrast to both annotations, the Ribo-seq coverage and TIS-Ribo-seq mapping data indicated a reading frame encoding a 298 aa protein. DNA and protein sequence alignments show that the here reported start codon is conserved among different *Synechocystis* strains and matches the annotations of homologs in *Synechocystis* strains PCC 6714 (hereafter *Synechocystis* 6714) and 7339 (**Figure 3A**). We further noticed that the protein length of 298 aa is supported by 19 different peptides from the proteogenomic analysis, while no peptides were found that support the longer N termini annotated in the GBK and KZS annotations. Hence, based on the combined evidence, the gene model for *sll0553* should be truncated by 47 codons.

As a second example, we show the gene *sll0872,* annotated by both GBK and KZS to encode a 130 aa protein (BAL28778.1). However, the very closely related homolog D082_22070 in *Synechocystis* 6714 with its annotated 68 aa is much shorter. Here, our Ribo-seq data indicate that likely the annotations in both strains are incorrect. Ribo-seq coverage and a TIS signal 19 nt downstream of another suitable start codon suggest that Sll0782 is a 101 aa protein in *Synechocystis* 6803 (**Figure 3B**), consistent with 9 detected peptides, while none was found matching the annotated 29 codons of the N-terminus before the here defined start. Sequence alignments show that a start codon at a matching position exists also in *Synechocystis* 6714, for which consequently the gene model for D082_22070 should be extended by 33 aa, yielding

101 aa proteins in both strains (**Figure 3B**). Thus, the data suggest that the *sll0872* gene model should be truncated by 29 codons.

The predictive power of TIS- and TTS-Ribo-seq profiling is furthermore evident in the region around genes *slr1220* and *slr1222* encoding unknown proteins. TIS-Ribo-seq in combination with Ribo-seq coverage supported an N-terminal extension of Slr1222 by 116 aa and initiation from a rare TTG codon, yielding a 321 aa protein **(Figure S7A**).

In many different bacteria, translation of *prfB* encoding release factor 2 (RF2 or PrfB) includes ribosomal frameshifting[55–57]. This *prfB* ribosomal frameshifting has not yet been demonstrated in cyanobacteria. However, the identification of two peptides by mass spectrometry provided proteogenomic support for the translation of an 26 aa SEP located upstream of *prfB*[47], while the annotated *prfB* gene (*sll1865*) appears truncated and starts from a rare TTG codon (**Figure 3C**). Here, we noticed abundant signals in both the Ribo-seq and RNA-seq coverage exactly upstream of the likely pausing site (**Figure 3C**). The encoded release factor 2 mediates termination at UAA and UGA codons. Almost all described frameshifting instances in other bacteria involve CUU UGA as the shift site/stop codon. A cytosine base 3´ of the UGA stop codon was shown in *E. coli* to make UGA a relatively weak termination codon, contributing to the frameshifting[58]. These bases – CUU UGA C – are also found in the C-terminal region of the 26 aa sORF here. In the presence of high PrfB concentrations, most ribosomes terminate at the UGA stop codon and only the 26 aa peptide is formed. If the concentration of PrfB is low, the CUU codon can detach from the anticodon of peptidyl-$tRNA_{Leu}$, mediating a shift from the +1 to the +2 frame[57]. Further described elements involved in the frameshifting are a short spacer upstream of the shift site (often UAU;[57], here CUAU), upstream of which an internal Shine-Dalgarno sequence is located, consistent with the TIS signal at the pausing site. All of these elements are also present here **(Figure 3C**). However, direct proof for the ribosomal frameshifting comes from an identified peptide (LGQTQDYLDLPG) that directly spans the frameshifting site **(Figure 3C**). This data corroborates the presence of this naturally occurring frameshift in cyanobacteria and provides a view of how such a site appears in ribosome profiling data. The abundant Ribo-seq coverage and TIS signals at the

shift site provide direct evidence for ribosome stalling upstream of the termination codon and re-initiation at the internal SD element as evidenced from the TIS signal.

We conclude that the generated Ribo-seq data are useful to spot ambiguities concerning functional start codons and support gene models that should be truncated or N-terminally extended. They furthermore illustrate an instance of ribosomal frameshifting in cyanobacteria.

Ribosome profiling aids the discovery of novel toxin-antitoxin candidates

Our ribosome profiling data allowed us to identify several novel protein-coding genes. Names were assigned conferring to their location in non-coding areas according to the location of the forward or reverse strand (*ncr* or *ncl*), their transcriptional unit (TU), or their location within an annotated gene. Two previously identified transcripts, Ncl1450 and Ncr1470, were thought to be two non-coding sRNAs complementary to each other[51]. Their origin from the same locus in forward and reverse orientation, their upregulation under low carbon[51], as well as the arrangement of *ncl1450* resembling an excludon[59] with regard to adjacent genes, prompted further investigation. Intriguingly, Ribo-seq coverage, TIS, and TTS signals indicated the translation of both Ncl1450 and Ncr1470 (**Figure 4A**). Fusions of the reading frames with the P_{petE} promoter led to their expression detected in western blots (**Figure 4B**). Ncl1450 homologs are found in several other cyanobacteria (**Figure 4C**), including *Pleurocapsa*, *Stanieria* (both 53% aa identity) and *Cyanothece* (50% aa identity), while *ncr1470* is only found in *Synechocystis* strains (**Figure 4E**). Ncl1450 is a 36 aa protein that was predicted as containing a single transmembrane domain by DeepTMHMM[60]. Corresponding fractionation experiments into soluble and membrane fractions and comparison with the membrane protein PetC supported a membrane localization for Ncl1450 (**Figure 4D**). The combination of a gene encoding a single transmembrane protein with an asRNA resembles the architecture of type I toxin-antitoxin systems[61]. We conclude that Ncl1450 and Ncr1470, previously considered as two non-coding sRNAs complementary to each other, encode SEPs of 36 and 40 aa and appear as a derivative of a type I toxin-antitoxin system.

An even more compelling case is provided by *ncr1610* previously considered as an sRNA[51]. Both TIS- and TTS-Ribo-seq data suggested that the *ncr1610* gene contains two sORFs encoding small proteins of 71 and 34 aa (**Figure 4F**). The coding potential of sORF1 was verified by the detection of 4 peptides from the search against the iPtgxDB and immunologically by FLAG tagging and western blot detection (**Figure 5A**). Homologs to *ncr1610-sORF1* exist throughout the cyanobacterial phylum, in Pseudomonadota, Chloroflexota and other bacteria (**Figure S7B**), while no conservation was found for *ncr1610-sORF2.* The *ncr1610-sORF1* start codon overlaps the stop codon of the preceding gene *slr1421* indicating a tight genetic association between the two genes. Searches using Foldseek[62] suggested that Slr1421, a thus far uncharacterized DUF4258 protein, is a structural homolog of MqsR (**Figure 4G**), an mRNA interferase toxin[63]. Therefore, the Ncr1610-sORF1 protein is suggested as the corresponding antitoxin.

Toxin-antitoxin proteins are often small, which makes the detection of the corresponding loci particularly difficult. Ribosome profiling aids in the discovery of novel elements of such systems, or entirely new ones. This is illustrated here by the antitoxin candidate Ncr1610-sORF1, and the *ncl1450/ncr1470* locus.

ORFBounder identification of unannotated sORFs

Existing tools for sORF identification from ribosome profiling data have significant limitations for TIS-Ribo-seq and TTS-Ribo-seq analysis. While HRIBO[64] incorporates DeepRibo for ORF detection, this tool is designed for conventional Ribo-seq data and not optimized for TIS or TTS datasets. The only analysis software available at the time for TIS-Ribo-seq data was RETscript[30], which had been applied to TIS peak calling in *E. coli*, but was difficult to adapt to other organisms and lacked any support for TTS data. To address these limitations, we developed ORFBounder, a customizable and efficient tool specifically designed for calling TIS and TTS peaks from the respective Ribo-seq data types, with a particular focus on identifying small ORFs. ORFBounder identifies translation sites based on ribosomal footprint enrichment following retapamulin treatment near potential start codons for TIS data. Using optimal read lengths and P-site offsets determined by metagene analysis (**Figure 1E**), start codons were detected by TIS-Ribo-seq for 100/165 of the previously annotated sORFs (KZS

annotation), which overlapped with the 98 ORFs that had significant Ribo-seq coverage (TE ≥0.288, **Table S3)**. All predictions made by ORFBounder were corroborated by HRIBO and confirmed through manual curation.

The power of combining the Ribo-seq and ORFBounder analysis with proteogenomic searches was well illustrated in the assignment of start codons. There was a single ATT start codon annotated (KZS), for gene *ureF/slr1899* encoding urease accessory protein F, while in the GBK and RefSeq annotations, *ureF* was annotated to start from a GTG. However, ORFBounder specified yet another, more distal TTG as start with a very good score, leading to a 243 aa protein, instead of the 217 aa (GBK and RefSeq) and 180 aa (KZS) forms. This N-terminally extended form was directly supported by a peptide MDDNLLPLLSGSAQLK identified in the standard iPtgxDB search with a spectral count of 5. It serves as direct proof for the N terminus, as the genomically encoded sequence would have been LDDNLLPLLSGSAQLK.

Validation of SEPs

Based on the DeepRibo and ORFBounder predictions, we selected 18 sORFs for experimental validation in addition to the Ncl1450/Ncr1470 pair and Ncr1610-sORF1 discussed above. Six of these sORFs were intergenic, three were short internal out-of-frame ORFs (IOFs), three were asRNA-located sORFs, and three were intragenic IIFs in the same reading frame as the main gene. All sORFs were fused to C-terminal 3xFLAG tags and if these could not be detected, we repeated the experiment using a SPA tag. Using this strategy, 15 of the 18 candidates were validated by western blotting (**Figure 5A**, **Table 1**).

Validated intergenically encoded SEPs

The genes *sml0012* and *norf2* were suggested before to encode hypothetical proteins[13,51,65], but have never been validated. The gene *sml0012* encodes a 45 aa polypeptide and is located upstream of *aroQ*, the 3-dehydroquinate dehydratase involved in the shikimate/chorismate pathway, essential for the synthesis of aromatic amino acids. We validated it by western blotting (**Figure 5A**) and detection of a single peptide in the iPtgxDB search (**Figure S8A**). Homologs of Sml0012 are widespread in the cyanobacterial phylum, sharing 38 to 88% aa identity (**Figure S8A**). Sml0012

might function as an attenuator peptide for *aroQ* translation, as the closely spaced arrangement of both genes in the same operon is conserved. Translation of the 67 aa protein Norf2 was well supported by western blotting, TIS- and TTS-Ribo-seq data, and the detection of 4 peptides in the iPtgxDB search (**Figure S8B**). There is no homolog for *norf2* in any other cyanobacterium, but some closely related proteins exist in gamma proteobacteria (**Figure S8B**). The region surrounding the *norf2* gene in the *Synechocystis* 6803 genome has been characterized as a genomic island[66], consistent with its paraphyly. The *norf2* gene overlaps by three codons with the downstream located gene *slr1062* encoding a PemK-type toxin[50] suggesting Norf2 is possibly the cognate antitoxin.

We also found SEPs on the ~100 kb plasmid pSYSA encoding the three distinct CRISPR-Cas systems of *Synechocystis* 6803[67]. There are substantial numbers of accessory proteins encoded with type III CRISPR-Cas gene cassettes[68], and more such genes might have been overlooked. Indeed, there is an sORF close to the 3′ end of the CRISPR3 array encoding a 46 aa SEP, here called Ssr70935, supported by strong Ribo-seq, TIS- and TTS-Ribo-seq coverage (**Figure S8C**) and by western blotting (**Figure 5A**). Ssr70935 is a very basic protein due to the presence of 9 lysine and 8 glutamine residues (isoelectric point (pI) of 12.04), a hallmark of nucleic acid-binding proteins.

We validated the translation of three previously defined sRNAs, Ncl1350, Ncr1420 and NsiR7[51]. Ncl1350 and Ncr1420 encode a 57 and a 52 aa SEP, respectively. Homologs of similar size exist in many different cyanobacteria (**Figure S9** and **S10**). The transcript level of *ncl1350* was previously found to be high and almost constitutive under ten different conditions. The transcription of *ncl1350* is linked with that of the downstream gene *sll0376* encoding a core protein of the cyanoglobule lipid droplet[69], to which it may be functionally related. Translation of Ncl1420 was validated by six different peptides identified in the iPtgxDB. Transcription of *ncl1420* is highest during exponential growth and very low under adverse conditions suggesting involvement in vegetative growth. In contrast, NsiR7 is upregulated under nitrogen starvation[70] and encodes a 51 aa SEP. Putative homologs of NsiR7 were identified only in a small number of other *Synechocystis* strains, which share the presence of cysteine residues at conserved positions (**Figure S10B)** suggesting redox sensitivity.

SEPs translated from asRNAs

Three asRNA-located SEPs were chosen for validation. One of them, *rsbU*-asORF, was found on an asRNA to *slr2031* encoding a homolog to the protein phosphatase RsbU[71]. The asRNA to *slr2031* was detected earlier and divergent transcript accumulation with the *slr2031* mRNA was observed under high light compared to standard conditions and cultivation in darkness[65]. In the Δ*slr2031* mutant, pigment contents (Chl and phycocyanin) were significantly higher compared to wild type[71]. This finding was consistent with reports that spontaneous mutations can be observed in this gene[72] and that deletion of a 154 bp DNA segment encoding part of the N-terminal domain of Slr2031 led to a glucose-resistant substrain[73] that also displayed higher pigment contents and showed an increased high-light tolerance[74]. The RNA-seq and Ribo-seq signals are stronger for *rsbU*-asORF than for the gene *slr2031* in sense orientation (**Figure 5B**). Within this asRNA, we found two sORFs in-frame with each other, encoding a 62 or a 38 aa peptide. Support from TTS-Ribo-Seq is excellent for both, since they end with the same stop codon. TIS-Ribo-Seq support is very clear and at the optimal distance of 19 nt for the shorter sORF. However, a peptide was found in the custom iPtgxDB search, directly supporting the long variant (**Figure 5B**) and expression of the FLAG tagged protein was confirmed by western blotting (**Figure 5A**). Therefore, it is also possible that the 38 aa variant is an IOF to the longer sORF. Corresponding contiguous ORFs exist in some other cyanobacteria including possible start codons at the corresponding positions for the 62 and 38 aa sORFs in *Synechocystis* 6803 (**Figure 5B**). RsbU-type phosphatases such as Slr2031 commonly belong to a partner switching system that consists of four modules. These are RsbU, a phosphorylatable anti-sigma antagonist, a kinase anti-sigma factor and an alternative sigma factor[75]. This already complex system is made even more complex by the asRNA and the herein encoded SEP.

We also validated the coding potential in two more asRNAs, in as_csx18 to gene *slr7091,* encoding a CRISPR-associated Csx18 protein **(Figure S11),** and in as_psbC to gene *sll0851,* encoding the photosystem II protein PsbC/CP43 **(Figure S12)**. For as_csx18 and its cognate gene in sense orientation, *slr7091,* both RNA-seq, Ribo-seq, and TIS-Ribo-seq coverage supported the expression of both DNA strands. The coding potential within *as_csx18* was well supported by the characteristic TIS-Ribo-seq signals 3′ of the start codon at position +19 nt. Our TIS-Ribo-seq signals also

indicated that the *slr7091* ORF encodes a 96 aa protein, which is shorter than the annotation as a 122 aa protein in the KZS, RefSeq and GBK databases, but consistent with the shorter Prodigal prediction (**Figure S11A**). The translation of the *as_csx18* sORF was validated by western blot, detecting the SPA-tagged asCsx18 protein (**Figure S11B**). Homologs of the asCsx18 protein could be encoded on asRNAs to the respective *csx18* gene in *Synechocystis* 6714, and possibly further cyanobacteria (**Figure S11C**). The asRNA overlapping the 3′ end of the *psbC* coding region originates from a TSS at position 1,347,388 on the forward strand of the chromosome[51]. Thus, this TSS is located 86 nt downstream of the *psbC* stop codon, while the asRNA overlaps >300 nt of the *psbC* coding region, suggesting a regulatory function. Indeed, both transcripts are divergently regulated. While the *as_psbC* transcript is maximum under adverse conditions, *psbC* mRNA accumulation is maximum under optimal growth conditions and exponential growth[51]. Translation of the *as_psbC*-encoded 28 aa SEP was supported by Ribo-Seq, TIS-Ribo-seq, TTS-Ribo-seq data and western blotting **(Figure S12A, B)**. The asPsbC protein in *Synechocystis* 6803 and several possible homologs in other cyanobacteria are very basic (pI of ~12) and very rich in alanine **(Figure S12C)**.

Translation of SEPs from alternative start sites within coding regions

In addition to SEPs encoded in intergenic regions or within asRNAs, we also found evidence for the translation of SEPs from alternative start sites within protein-coding genes Internal-Out of-Frame (IOF). Three examples were selected for validation, *der*-IOF, *crtR*-IOF and *slr1923*-IOF. The *der*-IOF encodes an only 25 aa SEP, located within the *der* gene (*slr1974*) for translation initiation factor IF-3 (**Figure 5C**). Although translation of the 42 aa SEP originating from the IOF within the *crtR* gene and of the SEP originating from the IOF within *cvrA* was well supported (**Figure S13**), their reading frames appear not as widely conserved. Lastly, we picked three examples of IIF-encoded SEPs (within *sll0871, slr0601* and *slr1397*), but these could not be validated.

To summarize, we validated 15 of 18 selected sORFs in addition to the Ribo-seq support by western blots, and five were further supported by the proteogenomic detection of peptides (**Table 1**). From these 15 validated instances, one was annotated as a hypothetical protein (*sml0012*), one was a previously predicted protein-coding

gene based on transcriptomic evidence (*norf2*), while the others were known only as sRNAs or asRNAs, or are antisense intragenic IOF-encoded SEPs. These results indicated the protein-coding potential of widely different types of loci.

A family of genes encoding two proteins by internal out-of-frame translation

Thus far we have reported on SEPs encoded in intergenic regions, antisense transcripts, or within coding regions. However, the Ribo-seq data also uncovered the existence of a hitherto uncharacterized family of genes that encoded a second, relatively long protein IOF. In **Figure 6** we show that a 112 aa protein is encoded within the coding sequence for the 151 aa protein Slr0489. From here on, we call these proteins Slr0489L for the longer, annotated gene product, and Slr0489S for the shorter protein, encoded IOF. Reads from the standard Ribo-seq analysis cover the entire annotated *slr0489* gene, but the coverage is visibly higher for the internal segment encoding Slr0489S (**Figure 6A)**. There are TIS-Ribo-seq signals for both reading frames at the characteristic position 19 nt 3′ of the respective start codons, and TTS-Ribo-seq signals mapped 14 nt 3′ of the respective stop codons. These signals are strikingly higher for the internal segment. Ultimate proof for the simultaneous expression of both proteins was obtained by the proteogenomic detection of 6 different peptides for Slr0489L and 5 for Slr0489S (**Figure 6B, C**).

A search for related genes yielded two additional loci in *Synechocystis* 6803, *slr1079* and *slr1082*. These genes harbor IOF reading frames as well, and Ribo-seq data supported their expression (**Figure S14**) and the detection of peptides in the iPtgxDB search directly supported the presence of Slr1079L, Slr1079S and Slr1082S (**Figure S14D, E**). Notably, the internally encoded proteins Slr0489S, Slr1079S and Slr1082S are closely related to each other, as are Slr0489L, Slr1079L and Slr1082L (**Figure 6B, C**). However, the respective longer and shorter proteins encoded within each of these loci are not related in sequence or structure. While Slr0489S has a slightly acidic pI of 5.56, Slr0489L has a pI of 9.38. Moreover, Slr0489L shares the presence of a signal peptide and a transmembrane domain with Slr1079L and Slr1082L (**Figure 6B**). The presence of these three genes does not follow phylogeny — the closely related strain *Synechocystis* 6714 lacks a related gene, while in the only distantly related

Synechococcus sp. PCC 7003, a close homolog exists (75% identical and 87% conserved aa with Slr0489, sequence ID WP_065714473.1).

To search for interacting proteins, we expressed a FLAG-tagged version of Slr0489L from a plasmid vector and performed coimmunoprecipitation (co-IP) analysis compared to a separate co-IP for sfGFP-3xFLAG as control. The most enriched protein was Slr0489S, suggesting that the internally encoded smaller and the longer 2TM proteins interact with each other ($\log_2$-fold enrichment of 10.6 for Slr0489L and 8.8 for Slr0489S, **Figure 6D**). However, the overexpression of *slr0489* inherently leads to a higher level of both, Slr0489L and Slr0489S, as they are both encoded on the same mRNA. Therefore, we constructed a cell line separately overexpressing Slr0489S fused to a C-terminal 3xFLAG tag (Slr0489S-3xFLAG) and performed co-IP. In this case, a $\log_2$-fold enrichment of 8.3 was observed for Slr0489S and of 6.4 for Slr0489L (**Figure 6E**). We noticed that in both cases two other highly enriched proteins were the CRISPR3 type III-Bv system Cmr5 protein Sll7086[76], and the minor pilin Slr2018[77] suggesting a connection to complexes close to the cell membrane. We conclude that the two different proteins produced from the *slr0489* locus interact with each other. In parallel, we performed the same type of co-IP experiments with Slr1079 (**Figure S14**). The co-IP using Slr1079L as bait yielded Slr1079S as the most enriched protein ($\log_2$-fold enrichment of 12.1 for Slr1079L and 6.7 for Slr1079S, **Figure S14B**), while the co-IP using Slr1079S as bait yielded Slr1079L as the most enriched protein ($\log_2$-fold enrichment of 10.1 for Slr1079S and of 8.8 for Slr1079L, **Figure S14C**).

To measure possible physiological effects of the Slr0489L/S system, we constructed a *Synechocystis* 6803 Δ*slr0489* deletion mutant. In this mutant, we then overexpressed the native *slr0489* gene with intact reading frames for both Slr0489L and Slr0489S. Alternatively, we expressed only Slr0489S, or only Slr0489L. For the latter, we introduced a *slr0489* gene with a single point mutation substituting the ATG start codon by an ACG codon for threonine, thus preventing the translation of Slr0489S. In a growth experiment lacking the inducer rhamnose, all gene variants had no or only negligible growth effects compared to wild type (**Figure 6F**). In this line only, clear growth retardation was observed in the presence of the inducer rhamnose, while all other strains grew similarly over a 5-day time course (**Figure 6F**). We conclude that

the sole expression of Slr0489L had a toxic effect on cell growth, which was neutralized in the presence of Slr0489S.

DISCUSSION

Genome consolidation effort in *Synechocystis* 6803

Synechocystis 6803 is an important model organism, which is reflected by the fact that its genome was the overall third to be fully sequenced 30 years ago[5]. Nevertheless, similar to other model systems, precise determination of coding regions has remained challenging up to this date, visible in differences in the available annotations of the identical genome sequence, such as the ones used here, deriving from NCBI RefSeq (2022), GenBank (2016) and the KZS 2009 annotation. These annotations differ both in terms of the overall numbers of coding sequences predicted and their precise start sites (**Figure S5, S6**). As we show in this study, several reading frames should be extended or truncated (**Figure 3**), which resolves ambiguities concerning functional start codons. Nevertheless, we are aware that ultimate proteomics proof requires in many instances specialized N-terminomics approaches[78], beyond the scope of this work. We have integrated both the different reference annotations and complemented them by *ab initio* gene predictions using Prodigal 2.6.3[52], ChemGenome 2.1[53], and an *in-silico* ORF prediction integrated into a standard and a custom integrated proteogenomic search database (iPtgxDB), which are both accessible through a single platform. The generation of a database comprising different annotations is generic; it enables researchers to jointly consolidate the genome annotation of their favorite model organism in the context of overlaid experimental data. The addition of data from multiple conditions and sources forms the basis to more comprehensively discover the full coding potential of genome sequences. To facilitate access to the entire set of data, we developed a JBrowse 2 instance[79], a web-accessible, fully interactive genome browser provided at https://www.bioinf.uni-freiburg.de/~ribobase/.

While SEPs mediate multiple important functions in all domains of life, they frequently have escaped annotation. Here, the combination of TIS- and TTS-Ribo-seq approaches with a proteogenomic analysis identified hundreds of candidate genes, of which several were experimentally validated to be present (**Table 1, Figures 4, 5, S7–**

S13). Our discovery of many new SEPs is in line with the recent prediction of extensive microprotein families in the Enterobacteriaceae[16]. More than 140 novel small proteins (in this case even ≤50 aa) have been manually validated over the span of roughly ten years in *E. coli*[80], recent work in *Synechocystis* 6803 provided evidence for 516 SEPs <100 aa[81], both consistent with our findings. We have previously characterized the *Synechocystis* 6803 SEPs AtpΘ as a regulator of ATP synthase[17], NblD as a factor involved in the nitrogen starvation-induced degradation of the phycobilisomes[45], AcnSP as a regulator of aconitase activity[44] and NirP1 as regulating nitrite reductase[46] demonstrating the physiological relevance of SEPs. In extension, many of the newly identified candidates in this study warrant further consideration. These include the *slr1421-ncr1610-sORF1* locus (**Figure 4F and G**), or the *ncl1450/ncr1470* locus, which are related to toxin-antitoxin systems. Our data indicate that a SEP evolved in the asRNA Ncr1470, which can be detected in some but not in the majority of related loci in other cyanobacteria possessing homologs of *ncl1450* (**Figure 4C, E**). The *ncl1450* gene and its overlapping *ncr1470* are located in an arrangement between *slr0841* and *slr0842* resembling an excludon[59]. The Ncl1450/Ncr1470 pair might constitute an associated type I toxin-antitoxin system, possibly with a stabilizing role[82]. Secondarily, the Ncr1470 sORF may have evolved in an originally non-coding asRNA.

CRISPR-associated hidden genes

We present two instances of previously unknown CRISPR-associated sORFs and corrected the ORF for gene *slr7091/csx18* **(Figure S8C, S11A)**. The here defined gene *ssr70935* encodes a 46 aa protein. It is located directly between the CRISPR array of a type III-Bv system[76], and the genes *slr7094* and *slr7095* with which it is cotranscribed as part of a tricistronic operon. Homologs of *ssr70935* in other cyanobacteria are also located adjacent to CRISPR-Cas loci, indicating a genetic association. There are several examples for the relevance of small proteins encoded in atypical sORFs associated with CRISPR-Cas systems. In the archaeal CRISPR-Cas type I-B system of *Haloferax volcanii*, the Cas11b is encoded by an internal in-frame translation within the *cas11* gene[83]. In a study on the type I-D system of the here investigated *Synechocystis* 6803, cryo-EM structures of the purified complex revealed an additional small subunit Cas11d[84]. Initially, no gene could be assigned to this

protein until it was discovered that it originated from an IIF translation within the gene of the large subunit *cas10d*[84]. Therefore, the here identified asCsx18 protein translated from an asRNA further illustrates the coding capacity of CRISPR-Cas systems to be extended by translation of ORFs from unorthodox locations. While currently neither the precise function of Csx18 nor of the asCsx18 protein is known, we noticed that Csx18 proteins are generally associated with type III-D and III-B CRISPR systems and likely adaptation-associated[68]. This would be consistent with the location of *slr7091* and its asRNA just upstream of the *cas1* and *cas2* adaptation module and next to the CRISPR3 effector genes on plasmid pSYSA in *Synechocystis* 6803[76].

Genes-in-genes

Our work not only provided evidence for several novel sORFs, but also uncovered the existence of a hitherto uncharacterized, novel family of 3 genes encoding a second, relatively long protein out-of-frame, which may be considered as genes-in-genes (**Figure 6, Figure S14**). This genetic arrangement resembles some bacteriophage loci encoding two-component spanins. Spanins are lysis proteins produced by phages that are required to disrupt the outer membrane during lysis of gram-negative host cells. Two-component spanins consist of an integral inner membrane i-spanin, and an outer membrane o-spanin, which is a lipoprotein[85]. In bacteriophage lambda, the reading frame Rz1 encoding the o-spanin is embedded within the reading frame Rz encoding the i-spanin[85]. The lambda phage i-spanin possesses a transmembrane domain and is located in the inner membrane. The C-terminal domain of the i-spanin, which extends into the periplasm, consists of alpha helical segments. In contrast, the o-spanin possesses a proline-rich region, is connected to the outer membrane, lacks a transmembrane domain and any other detectable structure[86]. Therefore, there are striking parallels to the here described *slr0489, slr1079* and *slr1082* loci. In addition to the genetic arrangement, the three proteins Slr0489L, Slr1079L and Slr1082L share the presence of a signal peptide and a transmembrane domain (**Figure 6B**), while Slr0489S, Slr1079S and Slr1082S encoded by the embedded ORFs possess a conserved proline-rich motif at positions 6–8 (**Figure 6C**), recognized as a membrane fusion motif in o-spanins[85].

However, there are also important differences. These three loci do not belong to any bacteriophage or prophage genome and their overexpression did not cause the cells to lyse. However, we did observe that the expression of Slr0489L without Slr0489S led to growth retardation (**Figure 6F**). Thus, the tight genetic connectivity between Slr0489L and Slr0489S first had to be disentangled to observe this effect, resembling a toxin-antitoxin effect in which binding of Slr0489S negates the deleterious effect of Slr0489L on our cultures. This resembles rather a defense-associated toxin–antitoxin relationship, as in the Panoptes defense system[87,88], providing an exciting topic for further research.

Outlook and limitations of this work

There are two previously published instances of alternative proteins produced from single genes in *Synechocystis* 6803. These are the IIF-encoded CRISPR-Cas subunit Cas11d[84] and the two alternative forms of ferredoxin:NADP oxidoreductase[89]. Translation of the corresponding genes *slr7011* and *slr1643/petH* can be seen in the Ribo-seq data, but not the initiation of translation at the internal sites. This points at a sensitivity limit and that we rather under- than overestimate the number of alternative ORFs. Part of this may relate to the fact that we have established TIS- and TTS-Ribo-seq and performed that analysis under one condition only. The shorter ferredoxin:NADP oxidoreductase proteoform is mainly made under heterotrophic conditions[89], which were not included here. Nevertheless, we provide information about 2,711 instances of potentially novel protein-coding sections or suggested corrections of annotated genes. This number includes more than 1,000 candidates each for SEPs from loci considered non-coding, or IOF-produced SEPs and we validated several instances related to the CRISPR-Cas systems, genes encoding core proteins of photosystem II, or regulatory factors. The functional analysis of these SEPs as well as of the instances of longer genes-in-genes is an exciting topic for further research.

DATA AVAILABILITY

Ribo-seq, TIS and TTS profiling datasets were deposited in the Gene Expression Omnibus (GEO) Database, with accession number GSE246336 (direct access: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE246336). The mass spectrometry proteomics data generated in this study have been deposited at the ProteomeXchange Consortium via the PRIDE[90] partner repository with the dataset identifier PXD060427. The datasets described in this study can be viewed with an interactive online JBrowse 2 instance at https://www.bioinf.uni-freiburg.de/~ribobase/.

ACKNOWLEDGEMENTS

This work was supported by the German Research Foundation (DFG) priority program SPP2002 “Small Proteins in Prokaryotes, an Unexplored World” (grants HE2544/12-1/2 to WRH, SH580/7-1/2 to CMS, BA 2168/21-1/2 to ROB), by grants HE 2544/22-1 to WRH and DFG_MA 49184-1 to BM and by Germany's DFG-funded Excellence Strategy (CIBSS - EXC-2189 - Project ID 390939984) to ROB. CHA acknowledges funding from the Swiss National Science Foundation for BH and GJ (grant 197391). The computational resources were supported by the de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) and ELIXIR-DE (Forschungszentrum Jülich and W-de.NBI-001, W-de.NBI-004, W-de.NBI-008, W-de.NBI-010, W-de.NBI-013, W-de.NBI-014, W-de.NBI-016, W-de.NBI-022). We thank Ann-Janine Imsiecke, Stephanie Färber and Philipp Kible (Würzburg, Germany) for the technical assistance in establishing the Ribo-seq workflow for *Synechocystis* 6803, Paulo Oliveira (Porto, Portugal) for the *tolC* transporter mutants and Joëlle Sasse Schläpfer (Zurich, Switzerland) for critical reading.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest.

AUTHOR CONTRIBUTIONS

WRH, CMS and ROB designed and supervised the Ribo-seq part of the project and secured funding. CHA supervised the proteogenomics part of the project. VK and PM constructed and cultivated strains. LH and VK designed and established Ribo-seq and TIS/TTS methods for *Synechocystis* 6803. RG performed bioinformatic processing of the Ribo-seq, TIS and TTS data, developed custom analysis tools and scripts, and created a JBrowse 2 instance for the project. VK, LH and WRH performed the analysis of Ribo-seq, TIS and TTS data including the manual labeling of the implied novel ORFs and sORFs and curation of gene annotations. PS and BM performed mass spectrometry, BH, GJ and CHA performed the proteogenomic analysis to identify known and novel ORFs/sORFs including iPtgxDB-based protein searches informed by Ribo-seq and TIS/TTS data, and contributed the generic approach for genome annotation consolidation. VK, VR and MG constructed strains and performed western blot analyses. RAB performed cultivation experiments with *slr0489* and *slr1079*. VK, LH, CHA and WRH wrote the manuscript with input from all authors. All authors read and agreed on the manuscript.

REFERENCES

1. Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., Karl, D.M., Li, W.K.W., Lomas, M.W., Veneziano, D., et al. (2013). Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. Proc Natl Acad Sci USA *110*, 9824–9829. https://doi.org/10.1073/pnas.1307701110.

2. Hagemann, M., and Hess, W.R. (2018). Systems and synthetic biology for the biotechnological application of cyanobacteria. Curr Opin Biotechnol *49*, 94–99.

3. Vijay, D., Akhtar, M.K., and Hess, W.R. (2019). Genetic and metabolic advances in the engineering of cyanobacteria. Curr. Opin. Biotechnol. *59*, 150–156. https://doi.org/10.1016/j.copbio.2019.05.012.

4. Klähn, S., Opel, F., and Hess, W.R. (2024). Customized molecular tools to strengthen metabolic engineering of cyanobacteria. Green Carbon.

5. Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., et al. (1996). Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. *3*, 109–136. https://doi.org/10.1093/dnares/3.3.109.

6. Zhang, J. (2000). Protein-length distributions for the three domains of life. Trends Genet. TIG *16*, 107–109. https://doi.org/10.1016/s0168-9525(99)01922-8.

7. Ahrens, C.H., Wade, J.T., Champion, M.M., and Langer, J.D. (2022). A practical guide to small protein discovery and characterization using mass spectrometry. J. Bacteriol. *204*, e0035321. https://doi.org/10.1128/JB.00353-21.

8. Burton, A.T., Zeinert, R., and Storz, G. (2024). Large roles of small proteins. Annu. Rev. Microbiol. *78*, 1–22. https://doi.org/10.1146/annurev-micro-112723-083001.

9. Alvarenga-Lucius, L.V., Linhartová, M., Schubert, H., Maaß, S., Becher, D., Hess, W.R., Sobotka, R., and Hagemann, M. (2023). The high-light-induced protein SliP4 binds to NDH1 and photosystems facilitating cyclic electron transport and state transition in *Synechocystis* sp. PCC 6803. New Phytol. *239*, 1083–1097.

10. Komenda, J., and Sobotka, R. (2016). Cyanobacterial high-light-inducible proteins--Protectors of chlorophyll-protein synthesis and assembly. Biochim. Biophys. Acta *1857*, 288–295. https://doi.org/10.1016/j.bbabio.2015.08.011.

11. Fromme, P., Melkozernov, A., Jordan, P., and Krauss, N. (2003). Structure and function of photosystem I: interaction with its soluble electron carriers and external antenna systems. FEBS Lett. *555*, 40–44.

12. Xu, W., Tang, H., Wang, Y., and Chitnis, P.R. (2001). Proteins of the cyanobacterial photosystem I. Biochim. Biophys. Acta BBA - Bioenerg. *1507*, 32–40. https://doi.org/10.1016/S0005-2728(01)00208-0.

13. Baumgartner, D., Kopf, M., Klähn, S., Steglich, C., and Hess, W.R. (2016). Small proteins in cyanobacteria provide a paradigm for the functional analysis of the bacterial micro-proteome. BMC Microbiol. *16*, 285. https://doi.org/10.1186/s12866-016-0896-z.

14. Brandenburg, F., and Klähn, S. (2020). Small but smart: On the diverse role of small proteins in the regulation of cyanobacterial metabolism. Life *10*, 322. https://doi.org/10.3390/life10120322.

15. Gray, T., Storz, G., and Papenfort, K. (2022). Small proteins; Big questions. J. Bacteriol. *204*, e0034121. https://doi.org/10.1128/JB.00341-21.

16. Fesenko, I., Sahakyan, H., Dhyani, R., Shabalina, S.A., Storz, G., and Koonin, E.V. (2025). The hidden bacterial microproteome. Mol. Cell *85*, 1024–1041.

17. Song, K., Baumgartner, D., Hagemann, M., Muro-Pastor, A.M., Maass, S., Becher, D., and Hess, W.R. (2022). AtpΘ is an inhibitor of F0F1 ATP synthase to arrest ATP hydrolysis during low-energy conditions in cyanobacteria. Curr. Biol. *32*, 136-148.e5. https://doi.org/10.1016/j.cub.2021.10.051.

18. Jurėnas, D., Fraikin, N., Goormaghtigh, F., and Van Melderen, L. (2022). Biology and evolution of bacterial toxin–antitoxin systems. Nat. Rev. Microbiol. *20*, 335–350.

19. Santos-Júnior, C.D., Torres, M.D.T., Duan, Y., Río, Á.R. del, Schmidt, T.S.B., Chong, H., Fullam, A., Kuhn, M., Zhu, C., Houseman, A., et al. (2024). Discovery of antimicrobial peptides in the global microbiome with machine learning. Cell *187*, 3761-3778.e16. https://doi.org/10.1016/j.cell.2024.05.013.

20. Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223. https://doi.org/10.1126/science.1168978.

21. Ingolia, N.T., Hussmann, J.A., and Weissman, J.S. (2019). Ribosome profiling: global views of translation. Cold Spring Harb. Perspect. Biol. *11*, a032698. https://doi.org/10.1101/cshperspect.a032698.

22. Mohammad, F., Green, R., and Buskirk, A.R. (2019). A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. eLife *8*, e42591. https://doi.org/10.7554/eLife.42591.

23. Gerashchenko, M.V., Lobanov, A.V., and Gladyshev, V.N. (2012). Genome-wide ribosome profiling reveals complex translational regulation in response to oxidative stress. Proc. Natl. Acad. Sci. *109*, 17394–17399. https://doi.org/10.1073/pnas.1120799109.

24. Zupanic, A., Meplan, C., Grellscheid, S.N., Mathers, J.C., Kirkwood, T.B.L., Hesketh, J.E., and Shanley, D.P. (2014). Detecting translational regulation by change point analysis of ribosome profiling data sets. RNA *20*, 1507–1518. https://doi.org/10.1261/rna.045286.114.

25. Chotewutmontri, P., and Barkan, A. (2021). Ribosome profiling elucidates differential gene expression in bundle sheath and mesophyll cells in maize. Plant Physiol. *187*, 59–72. https://doi.org/10.1093/plphys/kiab272.

26. Xiao, Z., Huang, R., Xing, X., Chen, Y., Deng, H., and Yang, X. (2018). De novo annotation and characterization of the translatome with ribosome profiling data. Nucleic Acids Res. *46*, e61. https://doi.org/10.1093/nar/gky179.

27. Ingolia, N.T., Lareau, L.F., and Weissman, J.S. (2011). Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell *147*, 789–802. https://doi.org/10.1016/j.cell.2011.10.002.

28. Bazzini, A.A., Johnstone, T.G., Christiano, R., Mackowiak, S.D., Obermayer, B., Fleming, E.S., Vejnar, C.E., Lee, M.T., Rajewsky, N., Walther, T.C., et al. (2014). Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. EMBO J. *33*, 981–993. https://doi.org/10.1002/embj.201488411.

29. Hsu, P.Y., Calviello, L., Wu, H.-Y.L., Li, F.-W., Rothfels, C.J., Ohler, U., and Benfey, P.N. (2016). Super-resolution ribosome profiling reveals unannotated translation events in Arabidopsis. Proc. Natl. Acad. Sci. *113*, E7126–E7135. https://doi.org/10.1073/pnas.1614788113.

30. Meydan, S., Marks, J., Klepacki, D., Sharma, V., Baranov, P.V., Firth, A.E., Margus, T., Kefi, A., Vázquez-Laslop, N., and Mankin, A.S. (2019). Retapamulin-assisted ribosome profiling reveals the alternative bacterial proteome. Mol. Cell *74*, 481-493.e6. https://doi.org/10.1016/j.molcel.2019.02.017.

31. Orr, M.W., Mao, Y., Storz, G., and Qian, S.-B. (2020). Alternative ORFs and small ORFs: shedding light on the dark proteome. Nucleic Acids Res. *48*, 1029–1042. https://doi.org/10.1093/nar/gkz734.

32. Froschauer, K., Svensson, S.L., Gelhausen, R., Fiore, E., Kible, P., Klaude, A., Kucklick, M., Fuchs, S., Eggenhofer, F., Yang, C., et al. (2025). Complementary Ribo-seq approaches map the translatome and provide a small protein census in the foodborne pathogen *Campylobacter jejuni*. Nat. Commun. *16*, 3078. https://doi.org/10.1038/s41467-025-58329-w.

33. Weingarten-Gabbay, S., Bauer, M.R., Stanton, A.C., Yu, Y., Freije, C.A., Welch, N.L., Boehm, C.K., Klaeger, S., Verzani, E.K., López, D., et al. (2025). Pan-viral ORFs discovery using massively parallel ribosome profiling. Science *388*, 1218–1224. https://doi.org/10.1126/science.ado6670.

34. Hadjeras, L., Heiniger, B., Maaß, S., Scheuer, R., Gelhausen, R., Azarderakhsh, S., Barth-Weber, S., Backofen, R., Becher, D., Ahrens, C.H., et al. (2023). Unraveling the small proteome of the plant symbiont *Sinorhizobium meliloti* by ribosome profiling and proteogenomics. microLife *4*, uqad012. https://doi.org/10.1093/femsml/uqad012.

35. Hadjeras, L., Bartel, J., Maier, L.-K., Maaß, S., Vogel, V., Svensson, S.L., Eggenhofer, F., Gelhausen, R., Müller, T., Alkhnbashi, O.S., et al. (2023).

Revealing the small proteome of *Haloferax volcanii* by combining ribosome profiling and small-protein optimized mass spectrometry. microLife *4*, uqad001. https://doi.org/10.1093/femsml/uqad001.

36. Tufail, M.A., Jordan, B., Hadjeras, L., Gelhausen, R., Cassidy, L., Habenicht, T., Gutt, M., Hellwig, L., Backofen, R., and Tholey, A. (2024). Uncovering the small proteome of *Methanosarcina mazei* using Ribo-seq and peptidomics under different nitrogen conditions. Nat. Commun. *15*, 8659.

37. Weaver, J., Mohammad, F., Buskirk, A.R., and Storz, G. (2019). Identifying small proteins by ribosome profiling with stalled initiation complexes. mBio *10*. https://doi.org/10.1128/mBio.02819-18.

38. Florin, T., Maracci, C., Graf, M., Karki, P., Klepacki, D., Berninghausen, O., Beckmann, R., Vázquez-Laslop, N., Wilson, D.N., Rodnina, M.V., et al. (2017). An antimicrobial peptide that inhibits translation by trapping release factors on the ribosome. Nat. Struct. Mol. Biol. *24*, 752–757. https://doi.org/10.1038/nsmb.3439.

39. Mangano, K., Florin, T., Shao, X., Klepacki, D., Chelysheva, I., Ignatova, Z., Gao, Y., Mankin, A.S., and Vázquez-Laslop, N. (2020). Genome-wide effects of the antimicrobial peptide apidaecin on translation termination in bacteria. eLife *9*, e62655. https://doi.org/10.7554/eLife.62655.

40. Stringer, A., Smith, C., Mangano, K., and Wade, J.T. (2022). Identification of novel translated small open reading frames in *Escherichia coli* using complementary ribosome profiling approaches. J. Bacteriol. *204*, e00352-21. https://doi.org/10.1128/JB.00352-21.

41. Vazquez-Laslop, N., Sharma, C.M., Mankin, A., and Buskirk, A.R. (2020). Identifying small open reading frames in prokaryotes with ribosome profiling. J. Bacteriol. *204*, e00294-21. https://doi.org/10.1128/JB.00294-21.

42. Karlsen, J., Asplund-Samuelsson, J., Thomas, Q., Jahn, M., and Hudson, E.P. (2018). Ribosome profiling of *Synechocystis* reveals altered ribosome allocation at carbon starvation. Msystems *3*, e00126-18.

43. Karlsen, J., Asplund-Samuelsson, J., Jahn, M., Vitay, D., and Hudson, E.P. (2021). Slow protein turnover explains limited protein-level response to diurnal transcriptional oscillations in cyanobacteria. Front. Microbiol. *12*, 657379.

44. de Alvarenga, L.V., Hess, W.R., Hagemann, M., and Hagemann, M. (2020). AcnSP - a novel small protein regulator of aconitase activity in the cyanobacterium *Synechocystis* sp. PCC 6803. Front. Microbiol. *11*, 1445.

45. Krauspe, V., Fahrner, M., Spät, P., Steglich, C., Frankenberg-Dinkel, N., Maček, B., Schilling, O., and Hess, W.R. (2021). Discovery of a novel small protein factor involved in the coordinated degradation of phycobilisomes in cyanobacteria. Proc. Natl. Acad. Sci. USA *118*, e2012277118. https://doi.org/10.1073/pnas.2012277118.

46. Kraus, A., Spät, P., Timm, S., Wilson, A., Schumann, R., Hagemann, M., Maček, B., and Hess, W.R. (2024). Protein NirP1 regulates nitrite reductase and nitrite excretion in cyanobacteria. Nat. Commun. *15*, 1911. https://doi.org/10.1038/s41467-024-46253-4.

47. Spät, P., Krauspe, V., Hess, W.R., Maček, B., and Nalpas, N. (2023). Deep proteogenomics of a photosynthetic cyanobacterium. J. Proteome Res. *22*, 1969–19. https://doi.org/10.1021/acs.jproteome.3c00065.

48. Omasits, U., Varadarajan, A.R., Schmid, M., Goetze, S., Melidis, D., Bourqui, M., Nikolayeva, O., Québatte, M., Patrignani, A., Dehio, C., et al. (2017). An integrative strategy to identify the entire protein coding potential of prokaryotic genomes by proteogenomics. Genome Res. *27*, 2083–2095. https://doi.org/10.1101/gr.218255.116.

49. Chen, Q., van der Steen, J.B., Arents, J.C., Hartog, A.F., Ganapathy, S., de Grip, W.J., and Hellingwerf, K.J. (2018). Deletion of *sll1541* in *Synechocystis* sp. strain PCC 6803 allows formation of a far-red-shifted holo-proteorhodopsin in vivo. Appl. Environ. Microbiol. *84*, e02435-17. https://doi.org/10.1128/AEM.02435-17.

50. Kopfmann, S., Roesch, S.K., and Hess, W.R. (2016). Type II toxin-antitoxin systems in the unicellular cyanobacterium *Synechocystis* sp. PCC 6803. Toxins *8*, 228.1-228.23. https://doi.org/10.3390/toxins8070228.

51. Kopf, M., Klähn, S., Scholz, I., Matthiessen, J.K.F., Hess, W.R., and Voß, B. (2014). Comparative analysis of the primary transcriptome of *Synechocystis* sp. PCC 6803. DNA Res. *21*, 527–539. https://doi.org/10.1093/dnares/dsu018.

52. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

53. Mishra, A., Siwach, P., Singhal, P., and Jayaram, B. (2019). ChemGenome2.1: An ab initio gene prediction software. In Gene Prediction: Methods and Protocols, M. Kollmar, ed. (Springer), pp. 121–138. https://doi.org/10.1007/978-1-4939-9173-0_7.

54. Baers, L.L., Breckels, L.M., Mills, L.A., Gatto, L., Deery, M.J., Stevens, T.J., Howe, C.J., Lilley, K.S., and Lea-Smith, D.J. (2019). Proteome mapping of a cyanobacterium reveals distinct compartment organization and cell-dispersed metabolism. Plant Physiol. *181*, 1721–1738. https://doi.org/10.1104/pp.19.00897.

55. Craigen, W.J., Cook, R.G., Tate, W.P., and Caskey, C.T. (1985). Bacterial peptide chain release factors: conserved primary structure and possible frameshift regulation of release factor 2. Proc. Natl. Acad. Sci. *82*, 3616–3620. https://doi.org/10.1073/pnas.82.11.3616.

56. Craigen, W.J., and Caskey, C.T. (1987). The function, structure and regulation of E. coli peptide chain release factors. Biochimie *69*, 1031–1041. https://doi.org/10.1016/0300-9084(87)90003-4.

57. Baranov, P.V., Gesteland, R.F., and Atkins, J.F. (2002). Release factor 2 frameshifting sites in different bacteria. EMBO Rep. *3*, 373–377. https://doi.org/10.1093/embo-reports/kvf065.

58. Tate, W.P., Poole, E.S., Horsfield, J.A., Mannering, S.A., Brown, C.M., Moffat, J.G., Dalphin, M.E., McCaughan, K.K., Major, L.L., and Wilson, D.N. (1995). Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. Biochem. Cell Biol. *73*, 1095–1103. https://doi.org/10.1139/o95-118.

59. Sesto, N., Wurtzel, O., Archambaud, C., Sorek, R., and Cossart, P. (2013). The excludon: a new concept in bacterial antisense RNA-mediated gene regulation. Nat. Rev. Microbiol. *11*, 75–82. https://doi.org/10.1038/nrmicro2934.

60. Hallgren, J., Tsirigos, K.D., Pedersen, M.D., Armenteros, J.J.A., Marcatili, P., Nielsen, H., Krogh, A., and Winther, O. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. Preprint at bioRxiv, https://doi.org/10.1101/2022.04.08.487609.

61. Shore, S.F.H., Leinberger, F.H., Fozo, E.M., and Berghoff, B.A. (2024). Type I toxin-antitoxin systems in bacteria: from regulation to biological functions. EcoSal Plus *12*, eesp-0025-2022. https://doi.org/10.1128/ecosalplus.esp-0025-2022.

62. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2024). Fast and accurate protein structure search with Foldseek. Nat. Biotechnol. *42*, 243–246. https://doi.org/10.1038/s41587-023-01773-0.

63. Yu, V., Ronzone, E., Lord, D., Peti, W., and Page, R. (2022). MqsR is a noncanonical microbial RNase toxin that is inhibited by antitoxin MqsA via steric blockage of substrate binding. J. Biol. Chem. *298*, 102535. https://doi.org/10.1016/j.jbc.2022.102535.

64. Gelhausen, R., Svensson, S.L., Froschauer, K., Heyl, F., Hadjeras, L., Sharma, C.M., Eggenhofer, F., and Backofen, R. (2021). HRIBO: high-throughput analysis of bacterial ribosome profiling data. Bioinformatics *37*, 2061–2063. https://doi.org/10.1093/bioinformatics/btaa959.

65. 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. USA *108*, 2124–2129. https://doi.org/10.1073/pnas.1015154108.

66. Kopf, M., Klähn, S., Pade, N., Weingärtner, C., Hagemann, M., Voß, B., and Hess, W.R. (2014). Comparative genome analysis of the closely related *Synechocystis* strains PCC 6714 and PCC 6803. DNA Res. *21*, 255–266. https://doi.org/10.1093/dnares/dst055.

67. Scholz, I., Lange, S.J., Hein, S., Hess, W.R., and Backofen, R. (2013). CRISPR-Cas systems in the cyanobacterium *Synechocystis* sp. PCC6803 exhibit distinct

processing pathways involving at least two Cas6 and a Cmr2 protein. PLoS ONE *8*, e56470. https://doi.org/10.1371/journal.pone.0056470.

68. Shah, S.A., Alkhnbashi, O.S., Behler, J., Han, W., She, Q., Hess, W.R., Garrett, R.A., and Backofen, R. (2019). Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. RNA Biol. *16*, 530–542.

69. Susanto, F.A., Jones, A.D., Donnelly, S.L., and Lundquist, P.K. (2025). Cyanoglobule lipid droplets are a stress-responsive metabolic compartment of cyanobacteria and the progenitor of plant plastoglobules. Plant Cell *37*, koaf127. https://doi.org/10.1093/plcell/koaf127.

70. Kopf, M., and Hess, W.R. (2015). Regulatory RNAs in photosynthetic cyanobacteria. FEMS Microbiol. Rev. *39*, 301–315. https://doi.org/10.1093/femsre/fuv017.

71. Huckauf, J., Nomura, C., Forchhammer, K., and Hagemann, M. (2000). Stress responses of *Synechocystis* sp. strain PCC 6803 mutants impaired in genes encoding putative alternative sigma factors. Microbiology *146*, 2877–2889. https://doi.org/10.1099/00221287-146-11-2877.

72. Trautmann, D., Voß, B., Wilde, A., Al-Babili, S., and Hess, W.R. (2012). Microevolution in cyanobacteria: re-sequencing a motile substrain of *Synechocystis* sp. PCC 6803. DNA Res. *19*, 435–448. https://doi.org/10.1093/dnares/dss024.

73. Katoh, A., Sonoda, M., and Ogawa, T. (1995). A possible role of 154-base pair nucleotides located upstream of ORF440 on CO2 transport of *Synechocystis* sp. PCC 6803. In Photosynthesis: From Light to Biosphere (Kluwer Academic Publishers), pp. 481–484.

74. Kamei, A., Ogawa, T., and Ikeuchi, M. (1998). Identification of a novel gene (*slr2031*) involved in high-light resistance in the cyanobacterium *Synechocystis* sp. PCC 6803. In Photosynthesis: Mechanisms and Effects: Volume I–V: Proceedings of the XIth International Congress on Photosynthesis, Budapest, Hungary, August 17–22, 1998, G. Garab, ed. (Springer Netherlands), pp. 2901–2904. https://doi.org/10.1007/978-94-011-3953-3_680.

75. Bouillet, S., Arabet, D., Jourlin-Castelli, C., Méjean, V., and Iobbi-Nivol, C. (2018). Regulation of σ factors by conserved partner switches controlled by divergent signalling systems. Environ. Microbiol. Rep. *10*, 127–139. https://doi.org/10.1111/1758-2229.12620.

76. Behler, J., Sharma, K., Reimann, V., Wilde, A., Urlaub, H., and Hess, W.R. (2018). The host-encoded RNase E endonuclease as the crRNA maturation enzyme in a CRISPR–Cas subtype III-Bv system. Nat. Microbiol. *3*, 367–377. https://doi.org/10.1038/s41564-017-0103-5.

77. Oeser, S., Wallner, T., Schuergers, N., Bučinská, L., Sivabalasarma, S., Bähre, H., Albers, S.-V., and Wilde, A. (2021). Minor pilins are involved in motility and

natural competence in the cyanobacterium Synechocystis sp. PCC 6803. Mol. Microbiol. *116*, 743–765. https://doi.org/10.1111/mmi.14768.

78. Kaushal, P., and Lee, C. (2021). N-terminomics - its past and recent advancements. J. Proteomics *233*, 104089. https://doi.org/10.1016/j.jprot.2020.104089.

79. Diesh, C., Stevens, G.J., Xie, P., De Jesus Martinez, T., Hershberg, E.A., Leung, A., Guo, E., Dider, S., Zhang, J., Bridge, C., et al. (2023). JBrowse 2: a modular genome browser with views of synteny and structural variation. Genome Biol. *24*, 74. https://doi.org/10.1186/s13059-023-02914-z.

80. Hemm, M.R., Weaver, J., and Storz, G. (2020). *Escherichia coli* small proteome. EcoSal Plus *9*. https://doi.org/10.1128/ecosalplus.ESP-0031-2019.

81. Peng, Z., Xiao, Q., and Wan, C. (2025). Identification and validation of SmORF-encoded peptides by genomics and proteomics in five cyanobacteria. J. Proteome Res. https://doi.org/10.1021/acs.jproteome.5c00685.

82. Jurėnas, D., Fraikin, N., Goormaghtigh, F., and Van Melderen, L. (2022). Biology and evolution of bacterial toxin–antitoxin systems. Nat. Rev. Microbiol. *20*, 335–350. https://doi.org/10.1038/s41579-021-00661-1.

83. Sailer, A.-L., Brendel, J., Chernev, A., König, S., Bischler, T., Gräfenhan, T., Urlaub, H., Gophna, U., and Marchfelder, A. (2025). Internal in-frame translation generates Cas11b, which is important for effective interference in an archaeal CRISPR-Cas system. Front. Microbiol. *16*, 1543464. https://doi.org/10.3389/fmicb.2025.1543464.

84. McBride, T.M., Schwartz, E.A., Kumar, A., Taylor, D.W., Fineran, P.C., and Fagerlund, R.D. (2020). Diverse CRISPR-Cas complexes require independent translation of small and large subunits from a single gene. Mol. Cell *80*, 971–979.

85. Kongari, R., Rajaure, M., Cahill, J., Rasche, E., Mijalis, E., Berry, J., and Young, R. (2018). Phage spanins: diversity, topological dynamics and gene convergence. BMC Bioinformatics *19*, 326. https://doi.org/10.1186/s12859-018-2342-8.

86. Cahill, J., Rajaure, M., O'Leary, C., Sloan, J., Marrufo, A., Holt, A., Kulkarni, A., Hernandez, O., and Young, R. (2017). Genetic analysis of the lambda spanins Rz and Rz1: identification of functional domains. G3 GenesGenomesGenetics *7*, 741–753. https://doi.org/10.1534/g3.116.037192.

87. Doherty, E.E., Adler, B.A., Yoon, P.H., Hsieh, K., Loi, K., Armbruster, E.G., Lahiri, A., Bolling, C.S., Wilcox, X.E., Akkati, A., et al. (2025). A miniature CRISPR–Cas10 enzyme confers immunity by inhibitory signalling. Nature, 1–8. https://doi.org/10.1038/s41586-025-09569-9.

88. Sullivan, A.E., Nabhani, A., Izrailevsky, D.S., Schinkel, K., Hoffman, C.R.K., Robbins, L.K., Nagy, T.A., Duncan, M.L., Ledvina, H.E., Erbse, A.H., et al. (2025). The Panoptes system uses decoy cyclic nucleotides to defend against phage. Nature, 1–9. https://doi.org/10.1038/s41586-025-09557-z.

89. Thomas, J.-C., Ughy, B., Lagoutte, B., and Ajlani, G. (2006). A second isoform of the ferredoxin:NADP oxidoreductase generated by an in-frame initiation of translation. Proc. Natl. Acad. Sci. U. S. A. *103*, 18368–18373. https://doi.org/10.1073/pnas.0607718103.

90. Perez-Riverol, Y., Bai, J., Bandla, C., García-Seisdedos, D., Hewapathirana, S., Kamatchinathan, S., Kundu, D.J., Prakash, A., Frericks-Zipper, A., Eisenacher, M., et al. (2022). The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. *50*, D543–D552. https://doi.org/10.1093/nar/gkab1038.

91. Lorenz, R., Bernhart, S.H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F., and Hofacker, I.L. (2011). ViennaRNA Package 2.0. Algorithms Mol. Biol. *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

92. Darty, K., Denise, A., and Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. Bioinformatics *25*, 1974–1975.

TABLES

Table 1. SEPs validated in this study by tagging and western blotting. The SEP ID is given (ID), followed by the coordinates and orientation (O). The SEPs are sorted by length (L), and the sequences are given. If proteogenomic support was found, the number of identified peptides is given in column P. All sORFs are located on the chromosome (acc. NC_000911), or, if indicated, on plasmid pSYSA (AP004311).

ID	Location	O	L	Sequence	Class	P	Figure
infC-IOF,	343,330-343,407	(+)	25	MNTTIKCASIRPNASSKLGIKSRQR	IOF		Fig. 5A, C
asPsbC	1,347,490 - 1,347,576	(+)	28	MNNVGSVSRSIPFSNPAAAARARPACQR	as		Fig. S6
Ncl1450	2,854,576 - 2,854,686	(-)	36	MTLIDTNSLHNAAFVLGTVLFVGILVWSVWFAFHMD	Ig/as		Fig. 4
Ncr1470	2,854,583 - 2,854,705	(+)	40	MWNANHTDQTRIPTKRTVPNTKAALWRLLVSMSVIQSRRC	Ig/as		Fig. 4,
crtR-IOF	982,131-982,259	(-)	42	MTLKMTQTILFPPAVPSSSLPPGSSTMRSFSLNGGYGANMSY	IOF		Fig. 5A
Sml0012	903,627 - 903,764	(-)	45	MDYLEKLLDKLRELAQKLIEGLLGPQGEPEPELIPIPVNDRRSRR	uORF	1	Fig. 5A, Fig. S8A
Ssr70935 (TU7087)	93,076 - 93,216 pSYSA	(+)	46	MQLKHWQSQKQEQKKSPGAPKKVRRQKARLQALRKKLHSSWAGNQP	ig		Fig. 5A, Fig. S8C
slr1923-IOF	2,211,455 – 2,211,595	(+)	46	MAGRENWVKKMRFILVFTKKCSPPRKKNLSPGPSGQALSAPLVVKC	IOF		Fig. 5A
NsiR7	450,353 - 450,505	(+)	49	MKPTHFIGHCPLLEILSVVQYRIADNANTKTICACISPNCGPMVSNVTDS	ig		Fig. 5A, Fig. S10B
Ncr1420	2,741,413 - 2,741,571	(+)	52	VISVKISSQPLRPSPYKAYRDPDTGQWRIAKQNREEQNNNSPNATDQDQDNG	ig	6	Fig. 5A, Fig. S10A
Ncl1350	2,706,192 - 2,706,365	(-)	57	MFWKLRQVSRFWFCYLNQPIFSRESKSTWRLSHFWSMYQISLLESCWQKSPIKETSL	ig		Fig. 5A, Fig. S9
rsbU-as	782,439 - 782,627	(-)	62	MGGVAIGKYARFGVIFRIKQGDKAMGIFQVGIHYPVEVVKNLGWTMAVQHFCPQHPPSHGHN	as	1	Fig. 5A, B
Norf2	365,077 - 365,283	(+)	68	MYAIEFEADLQDGVLTIPDHYKSLKNQHVRVVIMLDGHGNELELRALSDHSATLIDEWHDPSEDDVWI	ig	4	Fig. 5A, Fig. S8B
asCsx18	88,276 - 88,488 pSYSA	(-)	70	MKVEQATNRVLIITAAKPKGAISKITKVIRPLAMAKTRLRAKRSREGEYMLRATVRKVAEILTPPTTKIV	as		Fig. S11
Ncr1610 sORF1	3,274,013 - 3,274,231	(+)	72	MNTRTQTKIIHEGDYMAEVQVELTYTDHDWSPYLSLTEAQKLDQLRLALRRNDVKTASSLARVYHLTPVVVA	ig	3	Fig. 4F, 5A, Fig. S7B

FIGURES

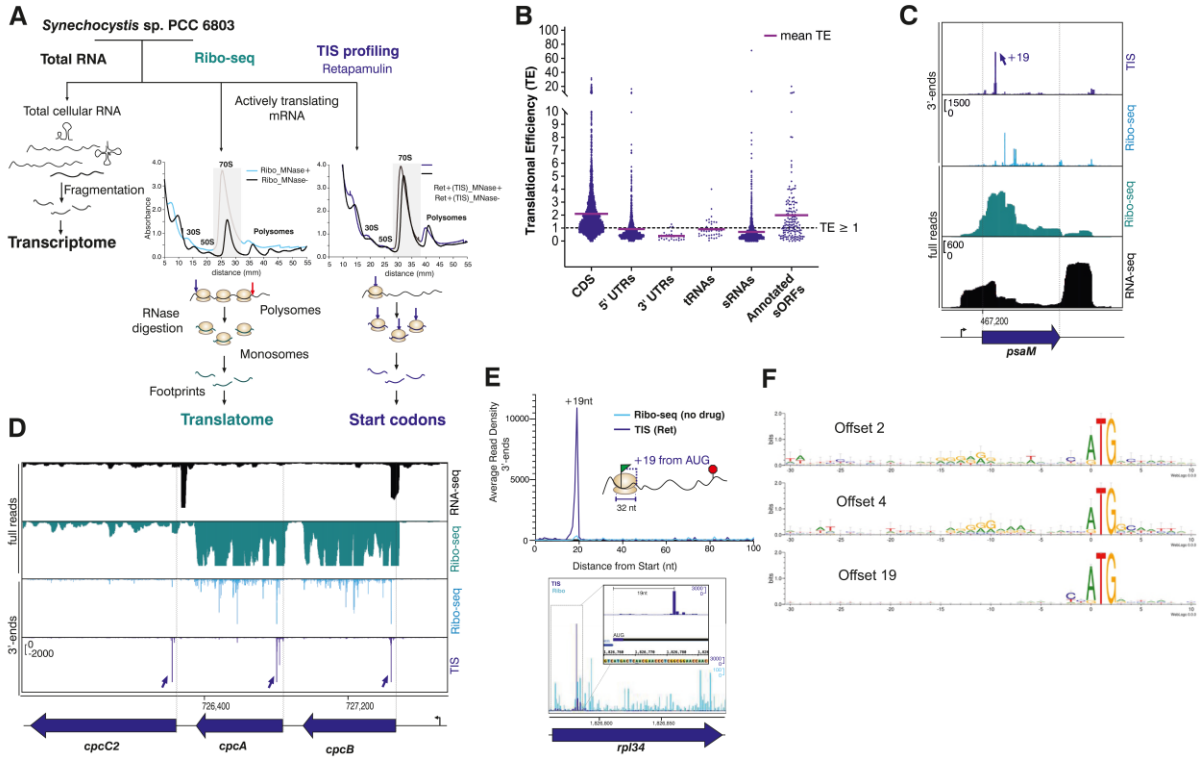

Figure 1. Establishment of TIS-Ribo-seq for the cyanobacterium *Synechocystis* 6803. **A.** Experimental set-up. Ribosomes were captured in the presence or absence of retapamulin on the translated mRNAs. In parallel, total RNA was prepared from aliquots of the same cultures. After sucrose gradient ultracentrifugation, absorption profiles were taken at 260 nm comparing samples from cultures treated with retapamulin (10 µg/mL) for 15 min and digested with MNase (Ret+(TIS)_MNase+; dark blue) or not (Ret+(TIS)_MNase-; black). The middle panel shows samples without retapamulin treatment, either after MNase digestion (Ribo_MNase+, light blue), or without (Ribo_MNase-, black). The MNase treatment depleted polysomes between 40 and 60 mm, enriching the 70S monosome peak at 30 mm. Retapamulin treatment stalled ribosomes at translation initiation sites while elongating ribosomes finished translation, resulting in an enrichment of 70S monosomes. Approximately 30 nt long footprints protected by and co-purified with 70S ribosomes were then subjected to cDNA library preparation and deep sequencing to identify the translatome and start codons, respectively. **B.** Translational efficiency (TE) for different RNA categories. The scatter plots show global TEs (TE = Ribo-seq/RNA-seq) computed from Ribo-seq replicates for all annotated coding sequences (CDS), 5′- and 3′-UTRs, tRNAs,

annotated small RNAs (sRNAs), and annotated sORFs encoding proteins of ≤70 aa. The purple lines indicate the mean TE of the respective transcript category. **C.** Upper 2 panels, read coverage from mapped 3′ ends from TIS-Ribo-seq (dark blue) and the Ribo-seq control (light blue), and from Ribo-seq and RNA-seq libraries (lower 2 panels) for the gene *psaM* encoding the photosystem I subunit XII. A sharp 3′ end was observed 19 nt downstream the annotated start codon. The Ribo-seq read coverage is enriched for the coding part, in contrast to the RNA-seq coverage, which starts at the position of the transcription start site (bent arrow). **D.** Read coverage for the *cpcBAC2* genes encoding the phycobilisome subunits CpcB, A and the rod linker polypeptide C2. **E.** Meta-analysis of footprints enriched in TIS-Ribo-seq. A sharp 3′ end was observed in 12,979 reads 19 nt downstream the annotated start codons of 582 genes (upper panel), exemplified in the lower panel for gene *rpl34* encoding the large ribosomal subunit protein L34. The 3′ ends of cDNA reads are depicted for ribosome-associated fractions after retapamulin treatment (dark blue) or without the drug (light blue lines). **F.** Enrichment of nucleotides associated with TIS-Ribo-seq 3′ end signals with an offset of 2 nt (29 genes), 4 nt (28 genes), or 19 nt (407 genes) from the annotated start codons after filtering for ≥10 reads. Sequence logos of the 30 nt preceding and ten nt following the start codons are shown. The preference for ATG over GTG codons is visible, as well as a slight preference for preceding pyrimidine residues. An enrichment for purine residues was only seen for the TIS-Ribo-seq signals with an offset of 2 nt in the region from -10 to -15 nt, or with an offset of 4 nt for the region from -7 to -14.

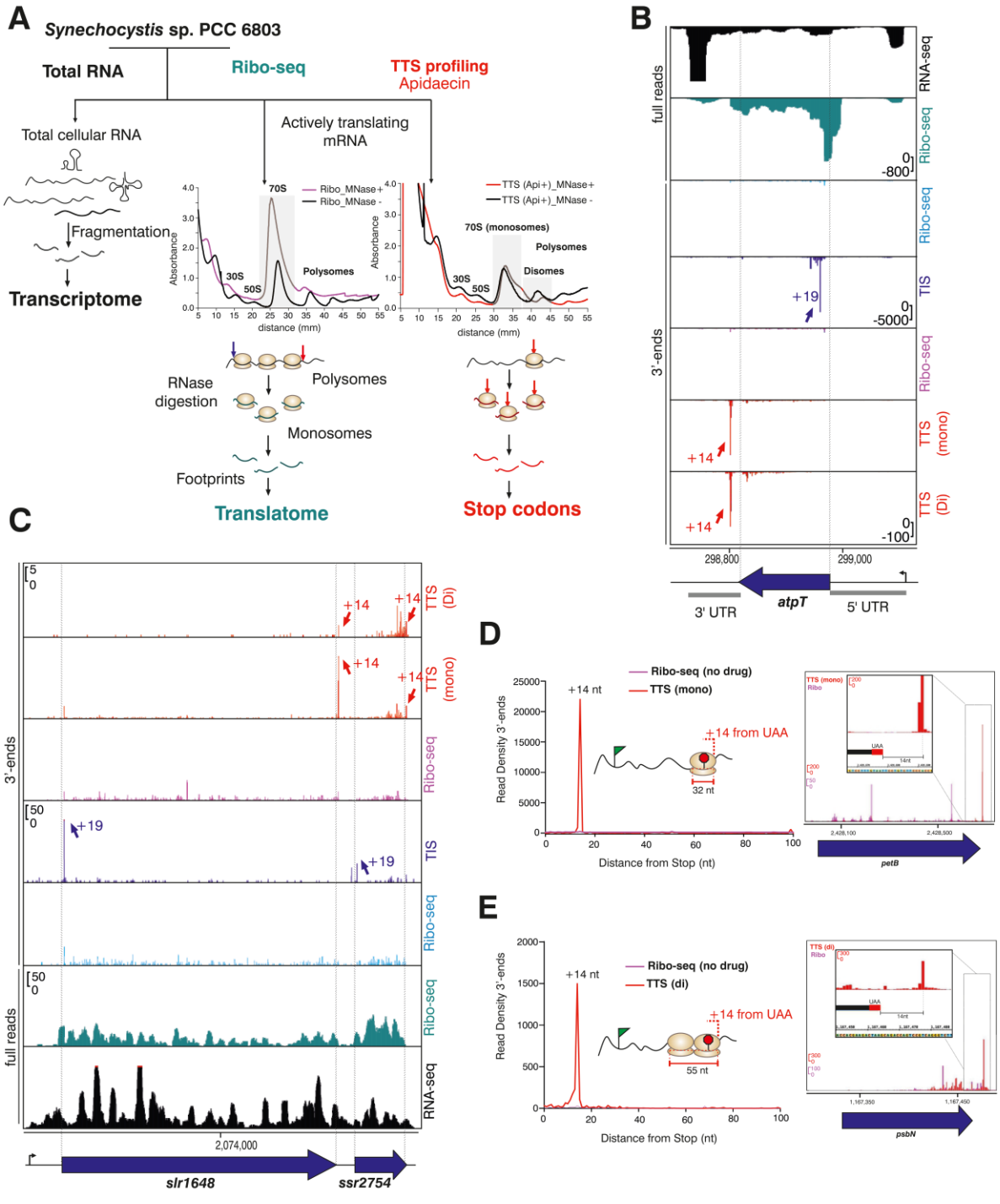

Figure 2. Establishment of TTS-Ribo-seq for *Synechocystis* 6803. A. Experimental set-up. Gradient profiles comparing samples from cultures treated with 10 µM Api137 for 15 min (right panel) digested with MNase (TTS (Api+)_MNase +; red) or without (TTS (Api+)_MNase -; black) compared to samples without Api137 treatment (panel in the middle), after MNase digestion (Ribo_MNase +, pink), or

without (Ribo_MNase -, black). After RNase treatment, approximately 30 nt long footprints protected by and co-purified with 70S ribosomes were obtained, subjected to cDNA library preparation and deep sequencing to identify the translatome and stop codons, respectively. In parallel, transcriptome samples were prepared and analyzed. **B.** TTS and TIS ribosome profiling for the gene *atpT* encoding the ATP synthase inhibitor AtpΘ[17]. Read coverage from RNA-seq and Ribo-seq libraries (upper 2 panels), mapped 3′ ends from Ribo-seq as control and TIS-Ribo-seq (middle panels) and from Ribo-seq control and TTS-Ribo-seq prepared from monosomes and disomes (lower 3 panels). The Ribo-seq read coverage is enriched for the coding part, in contrast to the RNA-seq coverage, which starts at the position of the transcription start site (bent arrow). The TIS-Ribo-seq 3′ ends map 19 nt downstream the start codon and the TTS-Riboseq 3′ ends 14 downstream the stop codons. **C.** TTS and TIS ribosome profiling for the genes *slr1648* and *ssr2754*. Ribosomes jam into blocked termination sites resulting in visible polysome fractions in Api-treated samples prepared from disomes as well as monosomes (TTS-seq, upper 2 panels). The gene *slr1648* encodes the carotenoid oxygenase SynDiox2 involved in the biosynthesis of retinal[49], *ssr2754* encodes a YefM-type antitoxin that forms an antitoxin-toxin gene pair with *ssr2755*[50]. **D.** Meta-analysis of footprints enriched in TTS-Ribo-seq from prepared monosomes (replicate 2, 32 nt reads, counts per million (CPM)-normalized). A sharp 3′ end was observed 14 nt downstream of the annotated stop codons across 1,055 genes (left panel), exemplified for gene *petB* (right panel). The 3′ ends of cDNA reads are depicted for ribosome-associated fractions after Api137 treatment (red) or without the drug (pink lines). **E.** Metagene analysis of footprints enriched in TTS-Ribo-seq from prepared disomes for 358 genes (replicate 2, 56 nt reads, CPM-normalized) (left panel), exemplified for the *psbN* gene (right panel).

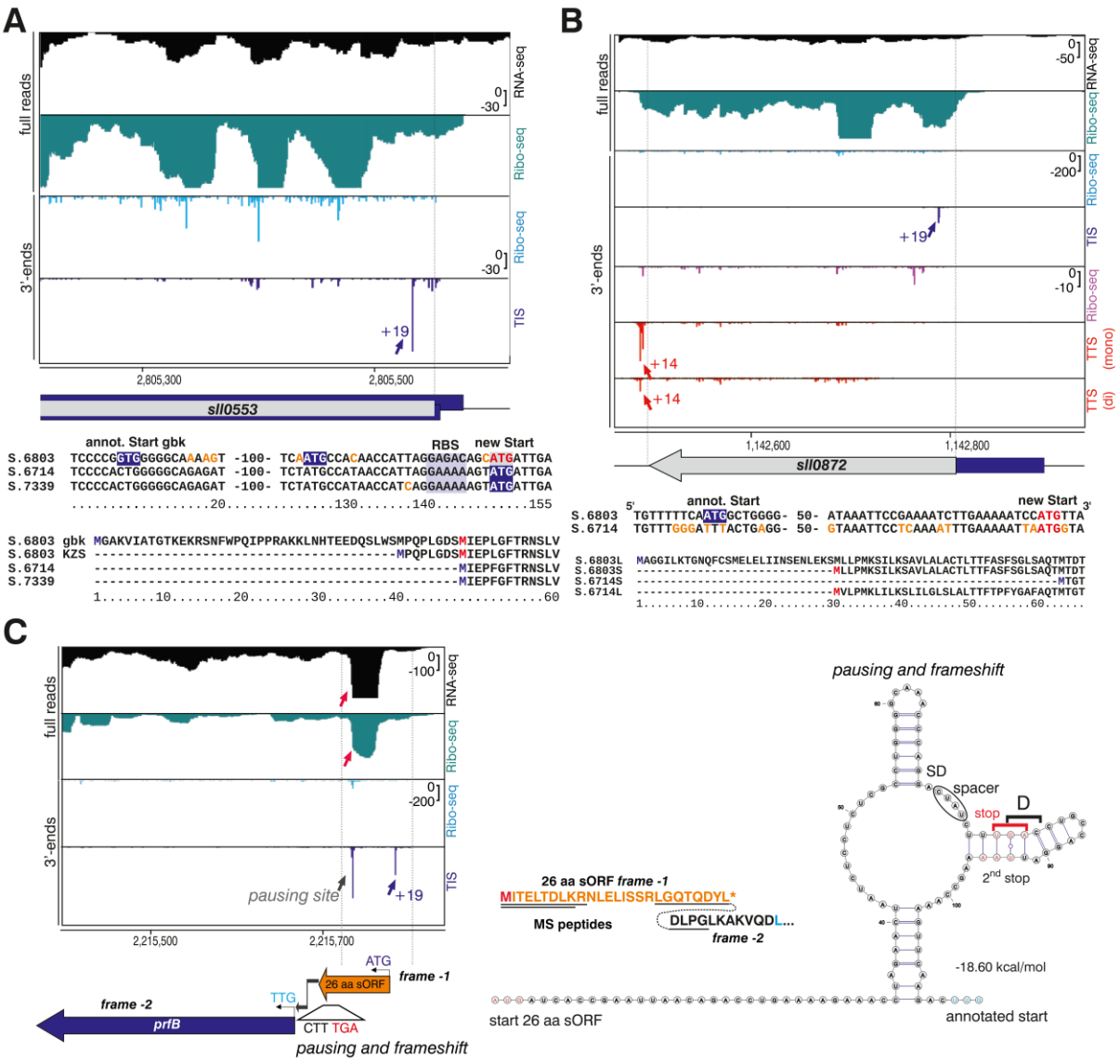

Figure 3. Start codon re-annotation and detection of ribosomal frameshifting. A. Re-annotation of the start codon of *sll0553* based on TIS profiling data. The annotated longest gene model in GBK is 1038 nt, encoding a 345 aa homolog with a GTG start codon (WP_010873707.1), while the KZS annotation suggests a 306 aa protein. However, the Ribo-seq coverage and TIS-Ribo-seq mapping indicate a reading frame encoding a 298 aa protein. Blue: current annotations (345 and 306 aa). Grey: Novel annotation based on TIS profiling data (298 aa). The enriched TIS peak associated with the correct start codon is indicated with a black arrow in the TIS library and indicates an N-terminal truncation for the *sll0553*-encoded protein. The lower panel shows DNA and protein alignments of the translation initiation regions from different *Synechocystis* strains. Thus, the TIS data supports a shorter form of *sll0553* in *Synechocystis* 6803, consistent with the annotations in *Synechocystis* strains PCC

6714 and 7339. The annotated and new start codons are indicated. The corrected protein length of 298 aa is supported by 19 different peptides from the proteogenomics analysis, while no peptides were found for the elongated N termini based on the GBK and KZS annotations. **B.** Correction of Sll0872. It is annotated as a 393 nt ORF in *Synechocystis* 6803 (GBK and KZS annotations) yielding a 130 aa protein (S.6803L; BAL28778.1), while the reading frame of the homolog D082_22070 in *Synechocystis* 6714 with 69 codons is much shorter (S.6714S). The Ribo-seq data suggest that both variants are incorrect. The predicted new start is depicted in red in both strains. **C.** Ribosome frameshifting. The *prfB* gene (*sll1865*) encodes peptide chain release factor 2. RNA-seq coverage (top, black) indicates contiguous transcription from a TSS approximately 140 nt upstream the annotated reading frame beginning with a TTG start codon. Ribo-seq coverage in the 2nd panel from top indicates translation in the upstream region, supported by a TIS-Ribo-seq signal 19 nt downstream of a suitable start codon (blue line +19 in the 4th panel from top), pointing at a 26-codon reading frame finishing with a TGA stop codon. However, both datasets show an unusual, abundant read accumulation in the same subregion (red arrows) and another TIS-Ribo-seq signal, which is not associated with a start codon. This pattern is consistent with ribosomal pausing and +1 ribosomal frameshifting, which is strongly supported by mapped peptides (black horizontal lines under the aa sequence). Two (MITELTDLK and MITELTDLKR) map to the upstream 26 aa frame, including the start amino acid, while a third (LGQTQDYLDLPG) directly spans the frameshifting site (lines connected by a dashed line). Elements involved in the frameshifting are annotated alongside the RNA secondary structure with a minimum free energy of -18.60 kcal/mol. The shift site is preceded by a short spacer and an internal Shine-Dalgarno (SD) sequence. A second UAA stop site downstream is indicated that is a conserved feature of these elements to suppress translation from re-initiation at the UUG. The RNA secondary structure was predicted by RNAfold[91] with default parameters and visualized with VARNA version 3.93[92].

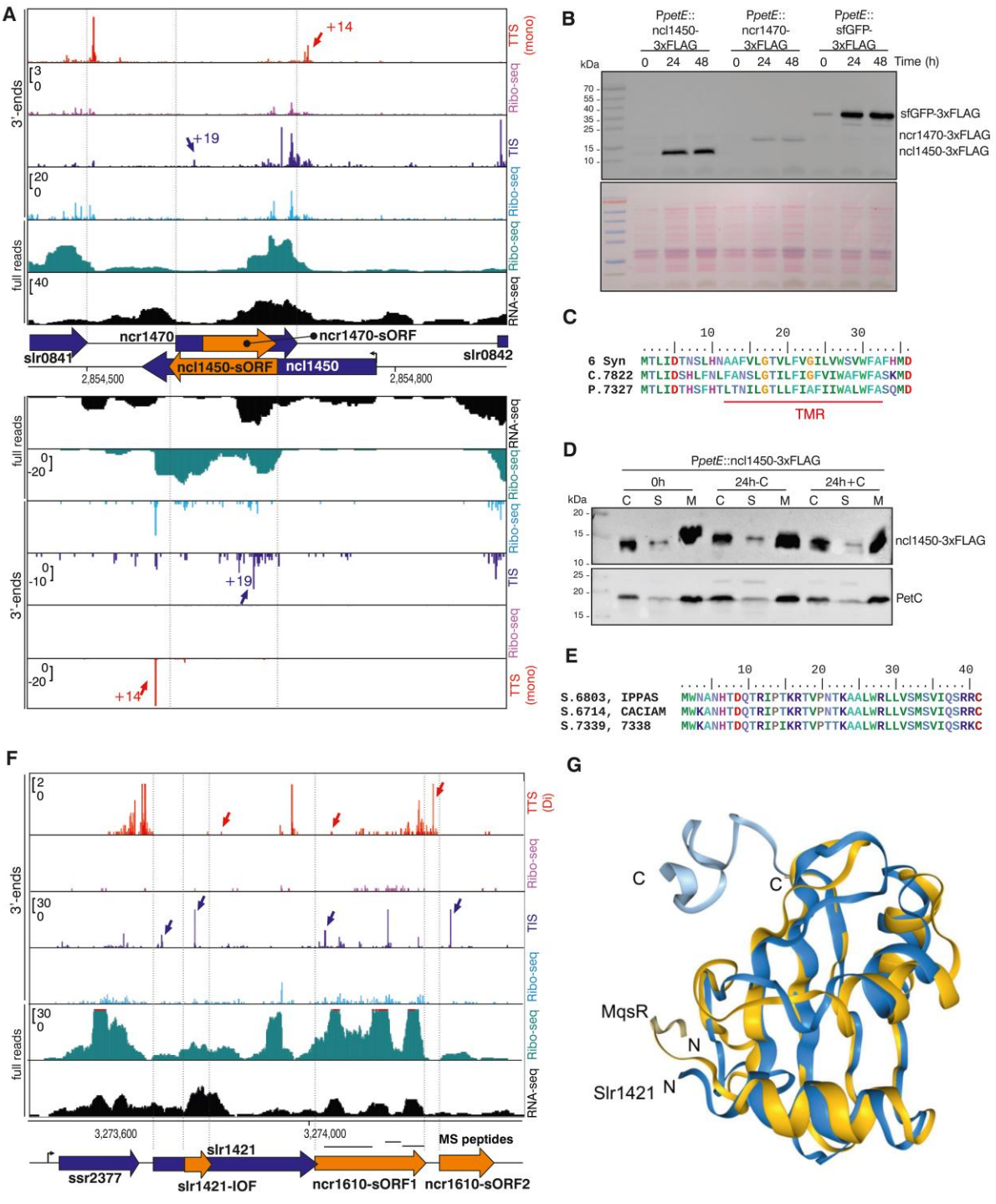

Figure 4. Identification of novel sORFs encoding toxin-antitoxin candidates. A. Ncr140 and Ncl1450 were annotated as a pair of overlapping sRNAs[51], the genomic arrangement is shown in the center of panel A. Ribo-seq, TIS- and TTS-Ribo-seq coverage indicate the translated segments in each transcript (orange arrows). **B.** Both coding sequences were fused to the copper-regulated P_{petE} promoter and a segment encoding a C-terminal 3xFLAG tag. A sequence encoding superfolder GFP (sfGFP) was engineered as control. The time course of protein induction after induction is

shown. The lower panel shows the corresponding ponceau red-stained membrane. **C.** Alignment of 6 identical Ncl1450 homologs in *Synechocystis* strains 6803, 6714, 7338, 7339, IPPAS B-1465 and CACIAM 05 (6 Syn), with homologs from *Cyanothece* sp. PCC 7822 and *Pleurocapsa* sp. PCC 7327. A predicted transmembrane domain is indicated (red line, TMR). **D.** Fractionation of total cellular protein (C) into soluble (S) and membrane (M) fractions yielded a similar localization for Ncl1450 in the membrane fraction as the known membrane protein PetC. Mouse monoclonal anti-FLAG serum coupled to horseradish peroxidase (A8592, Sigma) was used at a titer of 1:5000, anti-PetC antiserum (AS08 330, Agrisera) was used at a titer of 1:2,000 and developed with horseradish peroxidase-conjugated anti rabbit secondary antiserum (A8275, Sigma-Aldrich) at a titer of 1:10,000. **E.** Ncr1470 is only found in *Synechocystis* strains. The alignment shows all putative homologs that could be identified. Three sequence variants were found, with identical sequences each in *Synechocystis* 6803 and IPPAS B-1465 (top), in strains CACIAM 05 and PCC 6714 (center), and in strains PCC 7339 and PCC 7338 (bottom). None of the respective genes is annotated. **F.** Coding potential around the *slr1421* locus. TIS- and TTS-Ribo-seq data suggest two sORFs encoding small proteins of 71 and 34 aa within the previously defined sRNA *ncr1610*[51]. The coding potential of sORF1 was verified immunologically (**Figure 5A**) and by the detection of 3 peptides (horizontal tabs). Its start codon overlaps the stop codon of *slr1421*. Bent arrows indicate transcription start sites in panels A and F. **G.** Alphafold-predicted structure of Slr1421 (blue) and overlay with the best-matching structure identified by Foldseek (yellow), the protein MqsR (TM score: 0.6003).

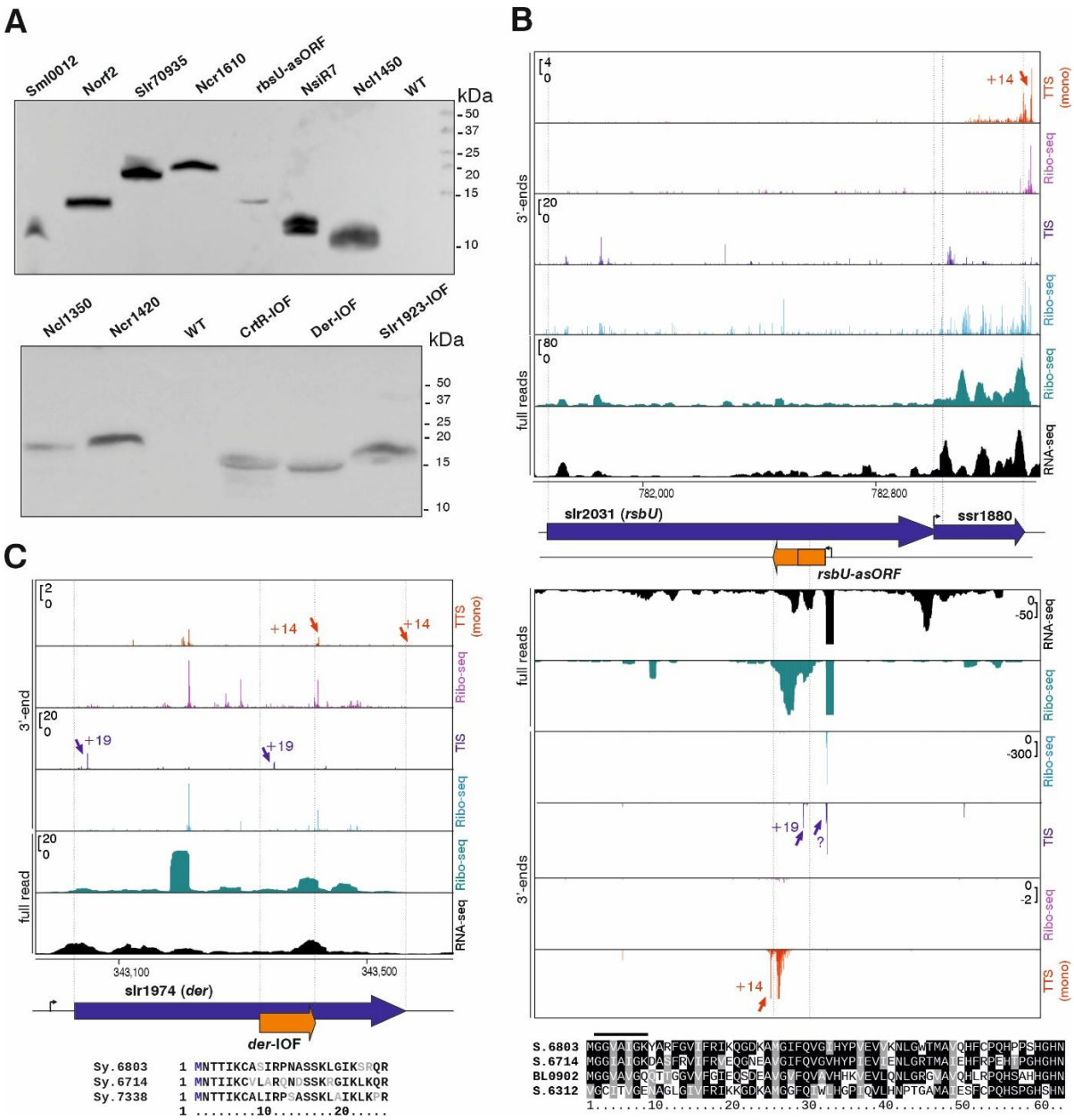

Figure 5. Validation of sORF candidates. A. Western Blot verification using anti-FLAG serum showing signals for 7 3xFLAG-tagged sORFs (upper panel) and 5 dual SPA-tagged proteins (lower panel) expressed from plasmid pVZ322 under control of the P_{petE} promoter induced by copper addition. *Synechocystis* wild type (WT) samples served as negative controls. **B.** Coding potential in an asRNA to *slr2031* (*rsbU*). Both DNA strands are expressed, supported by RNA-seq (black), Ribo-Seq (turquoise), TIS-Ribo-seq (dark blue) and TTS-Ribo-seq (red) datasets. The sequence alignment at the bottom shows possible homologs of the 62 aa protein encoded by the *rsbU*-asORF in *Synechocystis* strains 6714 and 6312, and in *Leptolyngbya* sp. BL0902 identified by TBlastN in antisense orientation to their respective *rsbU* genes. However,

it is unknown if corresponding asRNAs exist in these strains and the respective ORFs are longer than the shown sections. **C.** Evidence for an IOF encoding a 25 aa peptide within *slr0974* encoding the ribosome initiation factor IF-3. Bent arrows indicate transcription start sites.

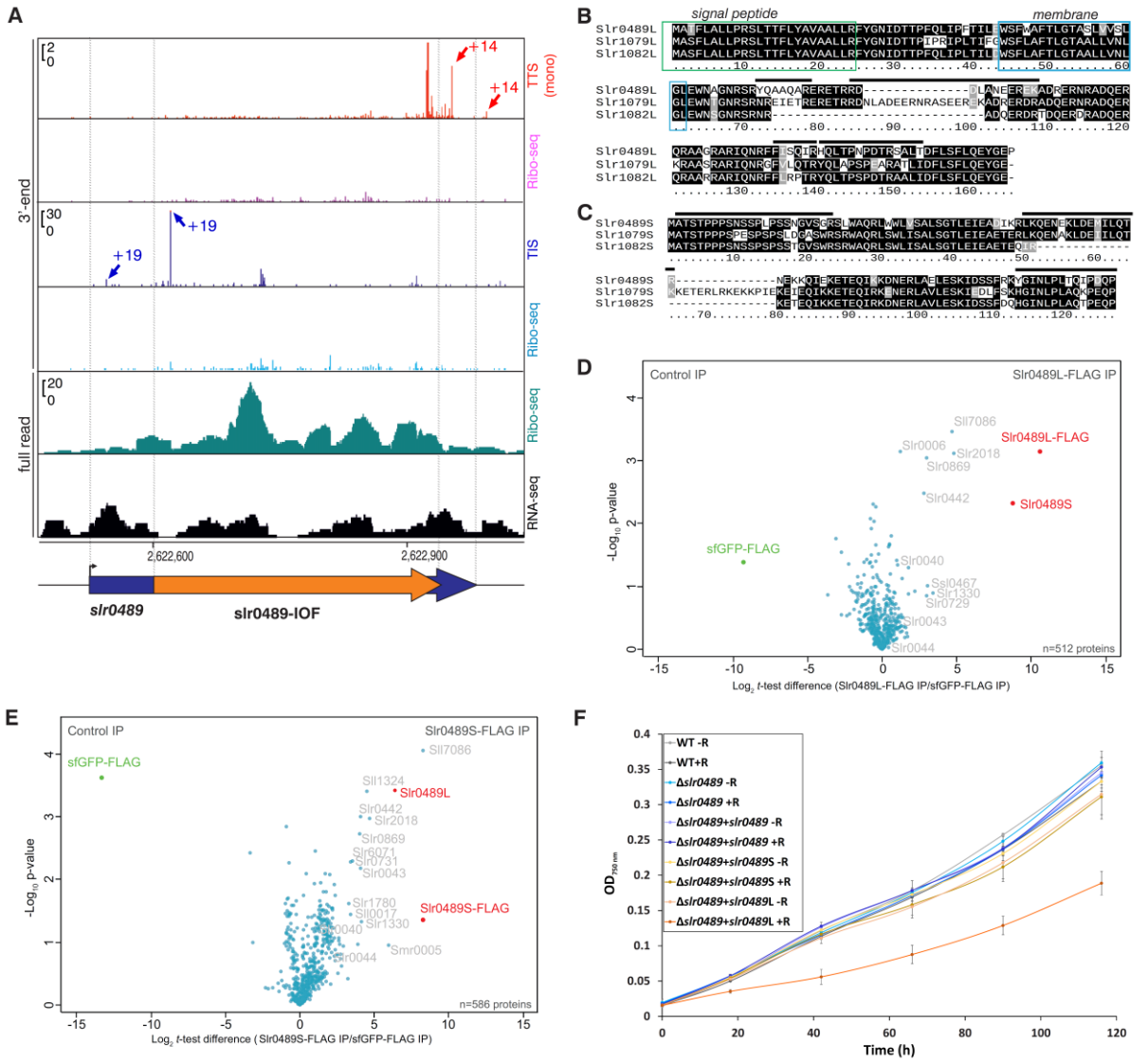

Figure 6. A family of genes encoding proteins in long internal out-of-frame reading frames. A. A protein Slr0489S is encoded within the *slr0489* reading frame (blue arrow) by a long IOF reading frame (orange arrow). Read coverage from Ribo-seq and RNA-seq libraries is shown in the lower 2 panels, the Ribo-seq read coverage is higher for the IOF segment. Mapped dual signals from TIS-Ribo-seq at the characteristic position 19 nt 3′ of the respective start codons (two central panels including control), and 14 nt 3′ of the respective stop codons from TTS-Ribo-seq (two upper panels including control) are shown. **B.** Sequence comparison of the proteins encoded by *slr0489* and two closely related loci, *slr1079* and *slr1082*. The proteins expressed from the annotated genes are labeled “L” for long. Signal peptides and transmembrane segments predicted by DeepTMHMM[60] are boxed. Black horizontal lines indicate some of the proteogenomically detected peptides for Slr0489L.

Altogether, 6 peptides were identified in the custom and 5 in the standard iPtgxDB. **C.** Sequence comparison of the proteins (labeled “S” for short) encoded IOF within *slr0489*, *slr1079* and *slr1082*. Black horizontal lines indicate some of the proteogenomically detected peptides for Slr0489S. **D.** Co-IP analysis of Slr0489L-3xFLAG in three replicates. The volcano plot shows the protein interaction partners of Slr0489L-3xFLAG identified by mass spectrometry compared to sfGFP-3xFLAG. The two most enriched proteins, Slr0489L and Slr0489S are colored red, the locus IDs are given for co-enriched proteins in grey. For the complete list of 512 detected proteins and further details, see **Table S6**. **E.** Co-IP analysis of Slr0489S-3xFLAG. Details as in panel D. For the complete list of 586 detected proteins, see **Table S7.** The results of an analogous co-IP experiment for *slr1079* encoding Slr1079L/S are shown in **Figure S14**. **F.** Effects of overexpressing *slr0489* (without FLAG tag) in a *Synechocystis* 6803 Δ*slr0489* deletion mutant, either with both intact reading frames (Δ*slr0489*+*slr0489* +R), of Slr0489S separately (Δ*slr0489*+*slr0489*S +R), or of Slr0489L from a variant in which a single point mutation prevented the expression of Slr0489S (Δ*slr0489*+*slr0489*L +R). The overexpression was triggered by the addition of rhamnose (+R), which was omitted from the respective controls (-R). The wild type (WT) and Δ*slr0489* deletion mutant served as further controls. The lines show the averaged increase in OD_{750} over 120 h. Cultures were measured in technical duplicates of two independent biological replicates.