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

*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

[#]Co-sharing first authors

Supplementary Information

Supplementary Tables

Table S1. Ribo-seq and proteogenomic analyses of 3,669 annotated genes. Column A provides the identifiers beginning with the GenBank accession, followed by the nt coordinates and a "+" symbol to indicate location on the forward strand and a "-" symbol for the reverse strand (example: NC_000911.1:7229-8311:+ for *psbA2*, *slr1311* gene). Proteogenomically detected proteins are labeled "Yes" in column M. Genes were considered translated if the RIBO-WT-avg_TE was ≥0.288 in column P.

-- See separate Excel file –

Table S2. **List of 2,711 genes classified as novel or as ORF corrections.** The type of novelty or correction is indicated in column A as "Unannotated", "Internal_OutofFrame" (IOF), "Internal_Inframe" (IIF), "N-terminal_extension" and "Truncated" (evidence that ORF should be truncated). Entries in red (n=808) were manually curated and commented. The manual inspection of this subset (column F) yielded 18 % of entries that were considered real, including 13 % encoding proteins of ≤70 amino acids. The number of membrane helices predicted by the DeepTMHMM algorithm[1] is indicated in column C. Overlaps with the findings by Peng et al.[2] are indicated in column E. In columns G–K information and comments are given for the manually inspected instances. Coordinates and genetic information for all entries is in columns L–U. Options to filter the data: Columns V–X contain the translational efficiencies calculated from Ribo-seq and RNA-seq coverage in two replicates, Y–AA provide the coverage from TIS-Ribo-seq. The respective sequencing coverages given in Reads Per Kilobase of transcript per Million mapped reads (RPKM) are provided in columns AB–AI. Columns AJ–BU provide values for prepared disome and trisome fractions (if available) and for TTS-Ribo-seq.

-- See separate Excel file –

Table S3. **List of 165 genes encoding proteins ≤70 aa (KZS annotation).** Proteogenomically detected proteins are labeled "Yes" in column J. All entries were manually validated in the RIBOBASE database for the presence of 3' end peaks in TIS- and TTS-Ribo-seq (columns K and L), and for contiguous Ribo-seq coverage (column M). Proteins previously detected by mass spectrometry in the study by Baers et al.[3] are labeled "yes" in column N. The average and respective translational efficiencies calculated from Ribo-seq and RNA-seq coverage in two replicates (columns U–W), and from TIS-Ribo-seq (columns X–Z) are given, followed by the respective coverages given in Reads Per Kilobase of transcript per Million mapped reads (RPKM); columns AA–AI).

-- See separate Excel file –

Table S4. Deoxyribonucleotides and primers used in this manuscript. All primers are given in 5'→3' orientation. Capital letters indicate segments binding to *Synechocystis* 6803 sequences.

ID	Full Name	Sequence
VK1	infC-in fwd	tacaatcgccaagaagtATGAACACGACTATC
VK2	infC-in rev	tctcttttccgcggaCCGTTGCCTTGACTTT
VK3	Slr1397 fwd v2	tacaatcgccaagaagtATGGATAATCTAACCG
VK4	Slr1397 rev	tctcttttccgcggaTCTTCCATCGATTGG
VK5	Slr0601 fwd v2	tacaatcgccaagaagtATGTCCACGGAAAAC
VK6	Slr0601 rev	tctcttttccgcggagTGATTGCTGTGCTC
VK7	Sll0871 fwd v2	tacaatcgccaagaagtATGAACACTTTTTTGGC
VK8	Sll0871 rev	tctcttttccgcggagCCCCGGACCAACAG
VK9	crtR-in fwd	cgccaagaagtATGACCCTGAAAATG
VK10	crtR-in rev	tctcttttccgcggagTAGCTCATATTTGC
VK11	slr1923-in fwd v2	tacaatcgccaagaagtATGGCCGGTCGAGAG
VK12	slr1923-in rev	tctcttttccgcggagCATTTCACAACCAATG
VK13	Toop_PstI_puc19_short fwd3	ttcgctcggttgccgccgggcgtttttttattactgcaggagacgaaagggcctcgtg
VK14	puc19_short_SalI rev	gtcgacaaaggccaggaaccgtaaaaagg
VK15	PpetE_puc19_short fwd 2	ccttttacggttcctggcctttgtcgacCTGGGCCTACTGGGCTATTC
VK16	ncl1450_3'UTR rev 2	aggccctttcgtctcctgcagtaataaaaaacgcccggcggcaaccgagcgaaTAATTCCCAACGAAGGCAAGC
VK17	PpetE rev	ACTTCTTGGCGATTGTATCTATAGG
VK18	3 x FLAG_ 3'UTR_Toop fwd	gattataaagatcatgatgg
VK19	ncl1450 rev	tgagcgtcatACTTCTTGGCGATTGTATCTATAGG
VK20	Ncl1450 fwd	gccaagaagtATGACGCTCATCGACACC
VK21	TU1220#1 fwd	agatacaatcgccaagaagtATGAAACATTCACGGAGAAATTTTTTCG
VK22	TU1220#1 rev	ccatcatgatctttataatcATATGGTGCGAGTCCCCC
VK23	TU1220#2 fwd	agatacaatcgccaagaagtATGATGGGGAATATCAGTTATG
VK24	TU1220#2 rev	ccatcatgatctttataatcTTCACATTTACCACGGCAAG
VK25	TU1183 fwd	agatacaatcgccaagaagtATGAGCAAATCTGCCGTTC
VK26	TU1183 rev	ccatcatgatctttataatcGTTGTCCTTATTGCCACAG
VK27	Ncr1610 fwd	agatacaatcgccaagaagtATGAACACCAGAACCCAAAC
VK28	Ncr1610 rev	ccatcatgatctttataatcAGCGACCACAACGGGGGT
VK29	TU1447 fwd	agatacaatcgccaagaagtATGGTGGGGGCGGGCTGG
VK30	TU1447 rev	ccatcatgatctttataatcACCCATCTGGCCGTCTCGG
VK31	Ncr1470 fwd	agatacaatcgccaagaagtATGTGGAACGCGAACCAC
VK32	Ncr1470 rev	ccatcatgatctttataatcACACCTCCTTGATTGTATGAC
VK33	TU2407 fwd	agatacaatcgccaagaagtGTGGAGGTCAAAGGGCGATC
VK34	TU2407 rev	ccatcatgatctttataatcGCAGTATCTAGAACAACTAGAGCC
VK35	Slr2031_as fwd	agatacaatcgccaagaagtATGGGCGGCGTTGCTATAG
VK36	Slr2031_as rev	ccatcatgatctttataatcATTATGACCATGACTAGGGGG
VK37	Norf2 fwd	agatacaatcgccaagaagtATGTACGCAATAGAGTTTG
VK38	Norf2 rev	ccatcatgatctttataatcGATCCATACATCATCCTC
VK39	TU7087 fwd	agatacaatcgccaagaagtATGCAACTAAAACACTGGCAATCTC
VK40	TU7087 rev	ccatcatgatctttataatcTGGCTGGTTTCCAGCCCA
VK41	nsiR7_ORF fwd	agatacaatcgccaagaagtATGAAACCGACTCATTTC
VK42	nsiR7_ORF rev	ccatcatgatctttataatcACTATCCGTAACATTTGAC
VK43	SPA-tag backbone fwd	tgacaagtagCGCCTCCATTCCCCAACG
VK44	SPA-tag fwd	ccgcggaaaagagaagatgg
VK45	SPA-tag rev	AATGGAGGCGctacttgtcatcgtcatcc
VK46	Ncr1610 backbone rev	ttttccgcggaAGCGACCACAACGGGGGTC
VK47	TU7087 backbone rev	ctcttttccgcggaTGGCTGGTTTCCAGCCCA
VK48	Slr2031-as backbone rev	ctcttttccgcggaATTATGACCATGACTAGGGGG
VK49	TU1220#1 backbone rev	ctcttttccgcggaATATGGTGCGAGTCC
VK50	TU1220#2 backbone rev	ctcttttccgcggaTTCACATTTACCACG
VK51	TU1183 backbone rev	ctcttttccgcggaGTTGTCCTTATTGC
VK52	TU1447 backbone rev	ctcttttccgcggaACCCATCTGGCCGT
VK52	Ncr1470 backbone rev	ctcttttccgcggaACACCTCCTTGATT

VK53	TU2407 backbone rev	ctcttttccgcggaGCAGTATCTAGAACA
VK54	Ncr1420 fwd	tacaatcgccaagaaGTGTGATATCTGTGAA
VK55	Ncr1420 rev	tctcttttccgcggagCCATTGTCCTGGTC
VK56	ssr0758_fwd	cgccaagaagtATGAAGAAAGAGTA
VK57	Ssr0758 rev	tctcttttccgcggaAAATGAAAGTTCTCG
VK58	Ncl1350 fwd	tacaatcgccaagaagtATGTTCTGGAAGC
VK59	Ncl1350 rev	tctcttttccgcggagAGGGAAGTTTCCTTG
VK60	pVZ322_seq fwd	tggttaattggttgtaacactggcag
VK61	pVZ322_seq rev	gtaataccatgaaaaataccatgctcag
VK62	pUC19_shorter_seq fwd	gaaatgttgaatactcatactcttcc
VK63	pUC19_shorter_seq rev	atagtcctgtcgggtttcgcc
P11-45	pUC19_fwd_shorterbackbone2	agctcactcaaaggcggtaa
P11-46	pUC19_rev_shorterbackbone2	tcaccgtcatcaccgaaacg
G3-64	pUC19s-PpetE_Fwd	cgtttcggtgatgacggtgaCTGGGCCTACTGGGCTATTC
G3-65	oop-pUC19s_Rev	ttaccgcctttgagtgagctaataaaaaacgcccggcg
VR277	PpetE_fwd2	ggattacagatcctctagagCTGGGCCTACTGGGCTATTC
VR278	PpetE_rev1	cgttgttcatACTTCTTGGCGATTGTATCTATAGG
VR279	as_psbC_fwd	gccaagaagtATGAACAACGTGGGTTCCG
VR280	as_psbC_rev	tttccgcggaCCTCTGGCATGCTGGTCG
VR281	Spa-3UTR-oop_fwd	atgccagaggtccgcggaaaagagaagatg
VR282	Spa-3UTR-oop_rev	tatgctcttctgctcctgcaaataaaaaacgcccggcg
VR285	PpetE_rev2	cgactttcatACTTCTTGGCGATTGTATCTATAGG
VR286	as_Csx18_fwd	gccaagaagtATGAAAGTCGAACAGGCC
VR287	as_Csx18_rev	tttccgcggaTACTATTTTTGTGGTTGGGG
VR288	SPA_UTR_oop_fwd	aaaaatagtatccgcggaaaagagaagatg
101	Slr1079-Leadless_fwd	ctttttagactggtcgtaatgaaATGGCAAGTTTTCTGGCTTTAC
102	Slr1079+Flag_rev_new	ccatcatgatcatgatctttataatccatTTCGCCGTATTCCTGGAG
105	NesSlr1079+SynRBS_fwd	aatatacaaaggaggtagaaATGGCAACATCGACACCACCCCCATCCCC
107	Nes1079-Flag_rev	gaatttggtaccgagctgcagTTAGGGTTGCTCTGGCTTCTGGGCTAGGG
109	Slr0489-Leadless_fwd	ctttttagactggtcgtaatgaaATGGCAACCTTCTTGGCCC
113	NesSlr0489+SynRBS_fwd	gtttataatatacaaaggaggtagaaATGGCAACATCGACACCACCCCCTTCCAACTC
115	Nes0489-Flag_rev	ggaatttggtaccgagctgcagTCAGGGCTGATCTGGTATCTGGGTTAGGGG
117	Slr0489_5'Fl_fwd	cacgaggccctttcgtctATGGGTAAATTGCGGCTGAG
118	Slr0489_5'Fl_rev	gcgttgacatcactctgtacGTAACAATATTGATCTGTGCTGGAAC
119	Slr0489_StrepR_fwd	GCACAGATCAATATTGTTACgtacagagtgatgtcaacgcc
120	Slr0489_StrepR_rev	CTAAAAAACCTAACTCTTTCATtattatcgtagttgctctcagagttg
121	Slr0489_5'UTR_fwd	ctgagagcaactacgataataATGAAAGAGTTAGGTTTTTTAGATGTTCC
122	Slr0489_5'UTR_rev	caataaattagggctcgccatGGCATCATTCTAGCTACTTCAAGC
123	Slr0489_3'Fl_fwd	gaagtagctagaatgatgccATGGCGAGCCCTAATTTATTGC
124	Slr0489_3'Fl_rev	ctttttacggttcctggcctttAACTGCTTTAAACCGTCCACTG
133	133-SegSlr0489_fwd	cgtgttgcaatggttggctc
134	134-SegSlr0489_rev	ggaagcagacaaaaactattaatttggcc
RB2/110	Slr0489+Flag_rev_new	ccatcatgatctttataatccatGGGCTCGCCATACTCTTGG
RB3/104	Slr1079+SynRBS_fwd	aatatacaaaggaggtagaaATGGCAAGTTTTCTGGCTTTAC
RB4	Slr1079+Flag_rev_new	ccatcatgatctttataatccatTTCGCCGTATTCCTGGAG
RB5/116	NesSlr0489+NatUTR_fwd	ctttttagactggtcgtaatgaaTTTCTGTATGCTGTAGCGGCAT
RB6/114	NesSlr0489+Flag_rev_new	cgccatcatgatctttataatccatGGGCTGATCTGGTATCTGGGTTAGGGG
RB7/108	NesSlr1079+NatUTR_fwd	ctttttagactggtcgtaatgaaTTCCTTTATGCCGTTGCGGCCC
RB8/106	NesSlr1079+Flag_rev_new	catgatctttataatccatGGGTTGCTCTGGCTTCTGGGCTAGGG
RB9	pUC19_Rha_3xFLAG_fwd	atggattataaagatcatgatgg
RB10	pUC19_Rha_3xFLAG_rev	ttctacctcctttgtatattataaac
RB11	pUC19_Rha_fwd	ctgcagctcggtaccaaattc
RB12	pUC19_Rha_noRBS_rev	ttcattacgaccagtctaaaaag
RB13	slr0489_fw	ttaataaggagatataccATGGCAACCTTCTTGGCC

RB14	pACYDuet_slr0489-6H-rev	gccccaaggggttatgctagttagtggtgatgatggtgatgGGGCTCGCCATACTCTTGG
RB15	His_tag_pACYCDuet_fw	catcaccatcatcaccactaac
RB16	pACYCDuet_rev	ggtatatctccttattaaagttaaac
RB17	slr0489short_mut_fw	ATTGAGGTTTTAcGGCAACATCGACACC
RB18	slr0489short_mut_rev	TAAAACCTCAATAATGCCGCTAC
RB19/ 111	Slr0489-Flag_rev	ggaatttggtaccgagctgcagTTAGGGCTCGCCATACTCTTGG

Table S5a. **A standard iPtgxDB search database with 145,955 protein entries (plus contaminants) was created.** Excluded were 3,030 possible N-terminal extensions shorter than 6 aa (not identifiable in a mass spectrometer), 70 entries annotated as pseudogenes by PGAP and 372 entries whose internal start site would not be distinguishable from a shorter proteoform of the longer PGAP annotation (if both start with a methionine).

Name	Annotati-ons	Clus-ters	New clus-ters	New trun-cations	New exten-sions	Total clus-ters	Total ids
RefSeq	3,692	3,692	3,692	0	0	3,692	3,692
GenBank	3,564	3,564	89	146	425	3,781	4,352
Kazusa	3,275	3,274	38	148	16	3,819	4,554
Prodigal	3,724	3,724	101	205	65	3,920	4,925
ChemGenome	5,071	5,071	1,802	63	997	5,722	7,787
In-silico ORFs	153,404	107,718	102,033	61	39,546	107,755	149,427

Table S5b. **A custom iPtgxDB search database with 6,928 protein entries (plus contaminants) was created.** Excluded were 109 possible N-terminal extensions shorter than 6 aa (not identifiable in a mass spectrometer), 70 entries annotated as pseudogenes by PGAP, and 523 entries whose internal start site would not be distinguishable from a shorter proteoform of the longer PGAP annotation (if both start with a methionine).

Name	Annotati-ons	Clus-ters	New clus-ters	New trunca-tions	New extensi-ons	Total clus-ters	Total ids
RefSeq	3,692	3,692	3,692	0	0	3,692	3,692
GenBank	3,564	3,564	89	146	425	3,781	4,352
Kazusa	3,275	3,274	38	148	16	3,819	4,554
Prodigal	3,724	3,724	101	205	65	3,920	4,925
Ribo-Seq	2,711	2,627	2,220	322	163	6,140	7,630

Table S6. List of proteins enriched in co-IP experiments with Slr0489L-3xFLAG and control, ranked by statistical significance (Student‘s t-test). The tagged proteins used as bait in the experiment (Slr0489L-3xFLAG) and in the control (sfGFP-3xFLAG) are highlighted in boldface letters. This table relates to results shown in **Figure 6D**. For the histograms of the label-free quantification (LFQ) value distribution, see **Figure S16**.

-- See separate Excel file –

Table S7. List of proteins enriched in co-IP experiments with Slr0489S-3xFLAG and control, ranked by statistical significance (Student‘s t-test). The tagged proteins used as bait in the experiment (Slr0489S-3xFLAG) and in the control (sfGFP-3xFLAG) are highlighted in boldface letters. This table relates to results shown in **Figure 6E**. For the histograms of the LFQ data distribution, see **Figure S17**.

-- See separate Excel file --

Table S8. List of proteins enriched in co-IP experiments with Slr1079L-3xFLAG and control, ranked by statistical significance (Student‘s t-test). The tagged proteins used as bait in the experiment (Slr1079L-3xFLAG) and in the control (sfGFP-3xFLAG) are highlighted in boldface letters. This table relates to results shown in **Figure S8B**. For the histograms of the LFQ data distribution, see **Figure S18**.

-- See separate Excel file --

Table S9. List of proteins enriched in co-IP experiments with Slr1079S-3xFLAG and control, ranked by statistical significance (Student‘s t-test). The tagged proteins used as bait in the experiment (Slr0489S-3xFLAG) and in the control (sfGFP-3xFLAG) are highlighted in boldface letters. This table relates to results shown in **Figure S8C**. For the histograms of the LFQ data distribution, see **Figure S19**.

-- See separate Excel file --

Supplementary Figures

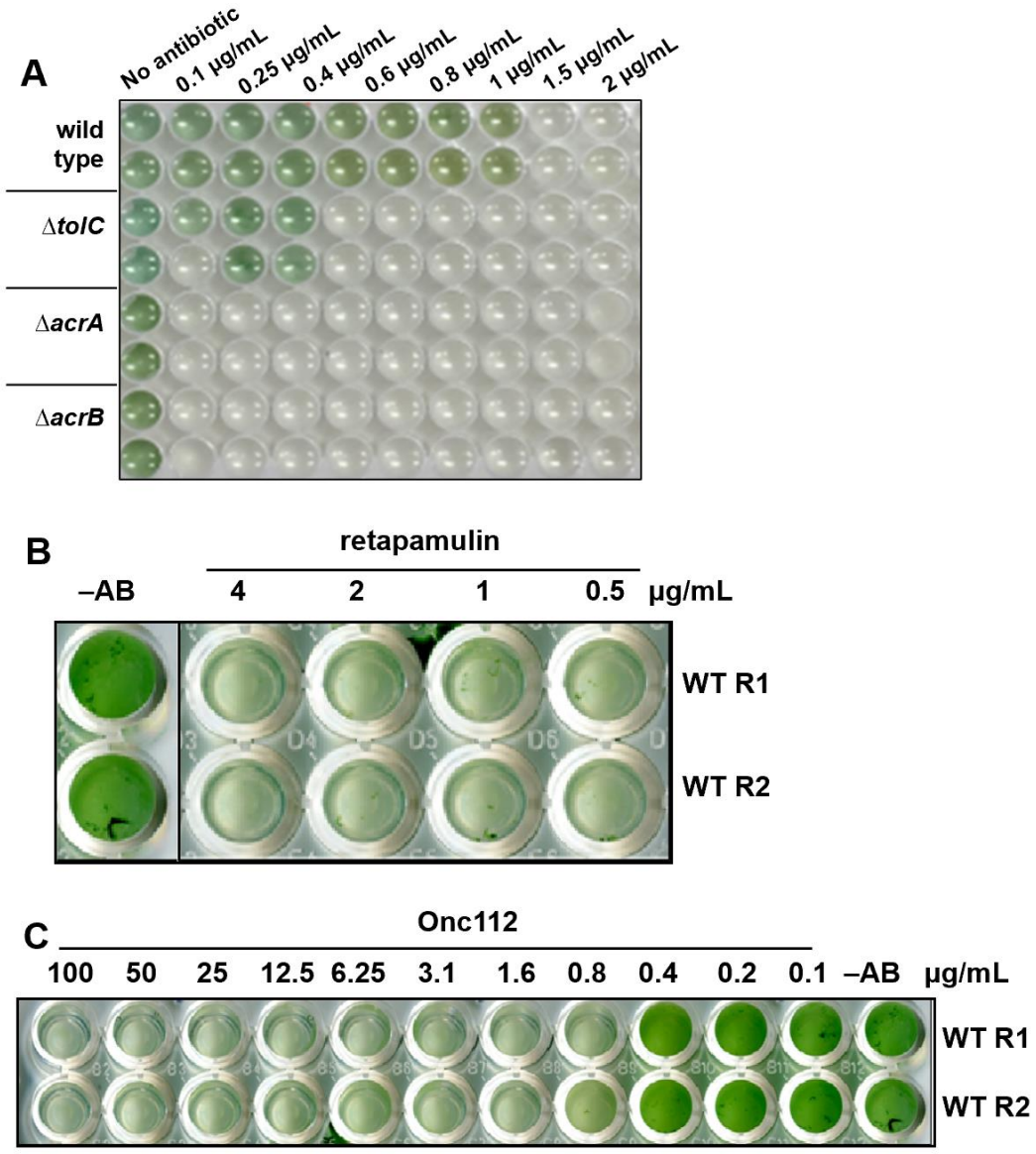


Figure S1. Assays to define the minimal inhibitory concentration for retapamulin and Onc112 in *Synechocystis* 6803. A. The wild type and different *tolC*-like transporter mutants[4] were grown in BG11 liquid culture volumes of 200 µL in the presence of retapamulin and documented after 7 days in two replicates. **B.** Wild type sensitivity test toward higher retapamulin concentrations in BG11 liquid cultures after 7 d in two replicates. **C.** Test of wild type sensitivity toward Onc112 after 7 d in two replicates. The respective concentrations are indicated in all panels (-AB, no antibiotic was added).

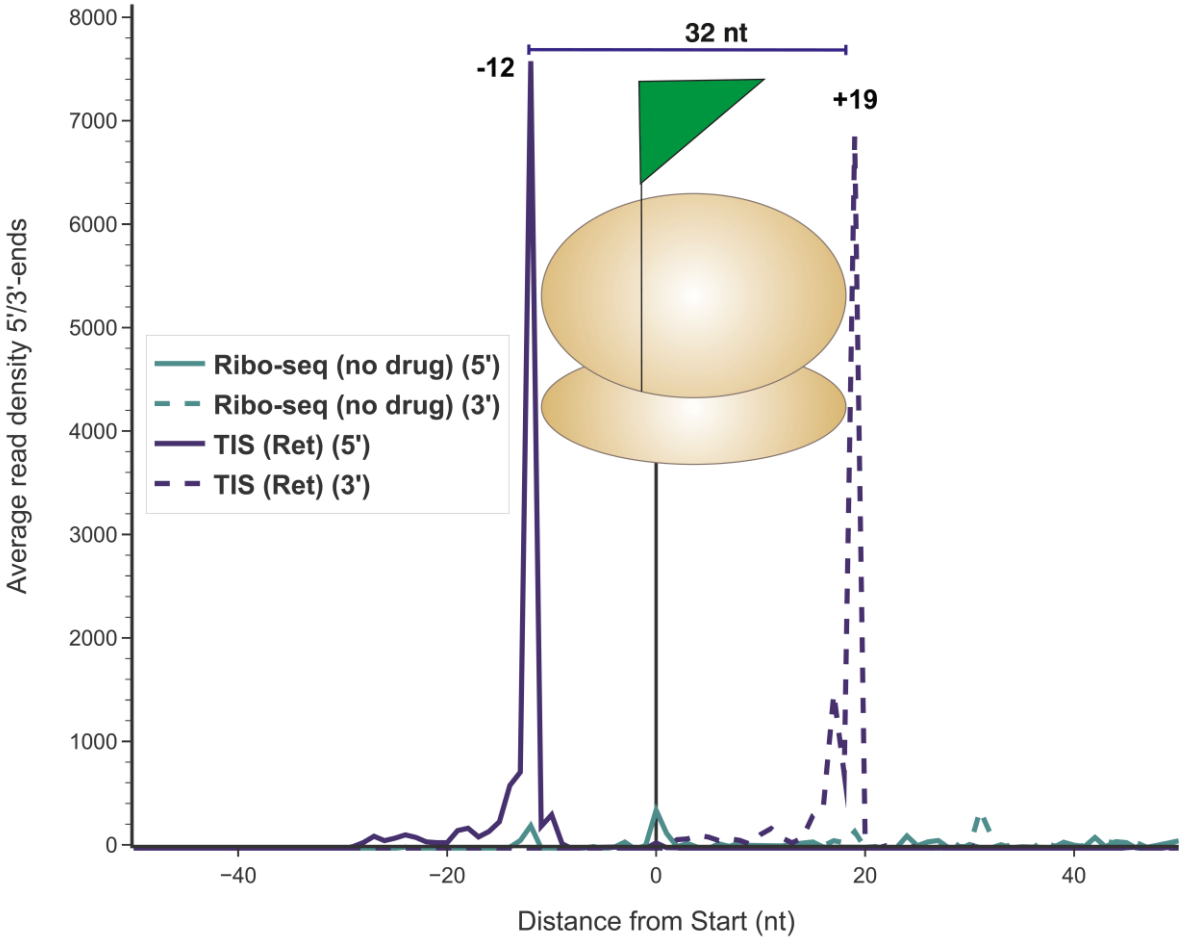


Figure S2. Meta-analysis of footprints enriched in TIS-Ribo-seq for very short ORFs. A sharp 5' end was observed for on average 6,567 reads 12 nt upstream and 19 nt downstream of the start codons of 39 small ORFs (≤50 aa). This demonstrates that sORFs exhibit translational initiation patterns comparable to those observed for other ORFs in **Figure 1E**.

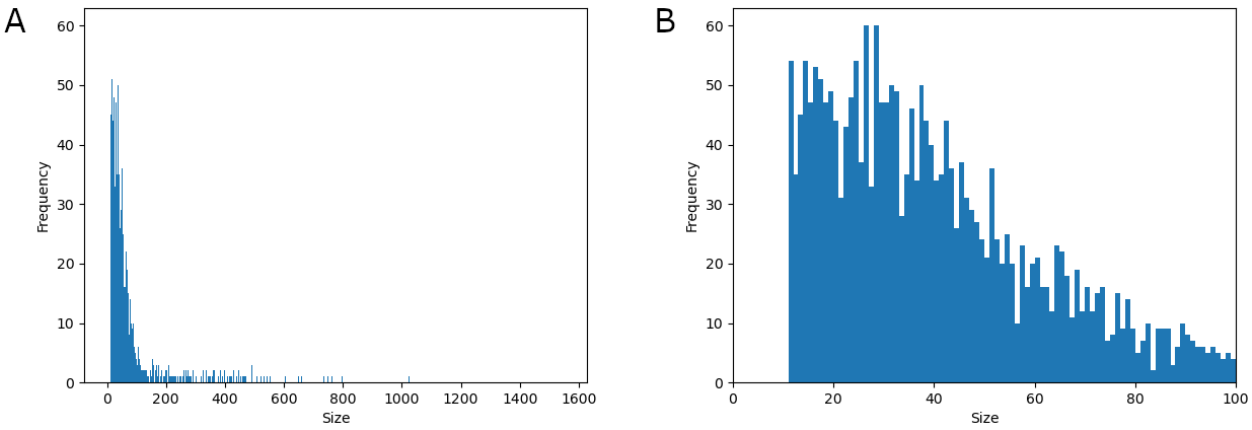


Figure S3. Histogram of the lengths of 2,711 Ribo-seq candidates. **A.** Size histogram (in amino acids (aa)) showing the lengths of all top candidates, which are enriched in candidates below 150 aa. **B.** Zoom in to the subset ≤100 aa, here with a bin size of 1.

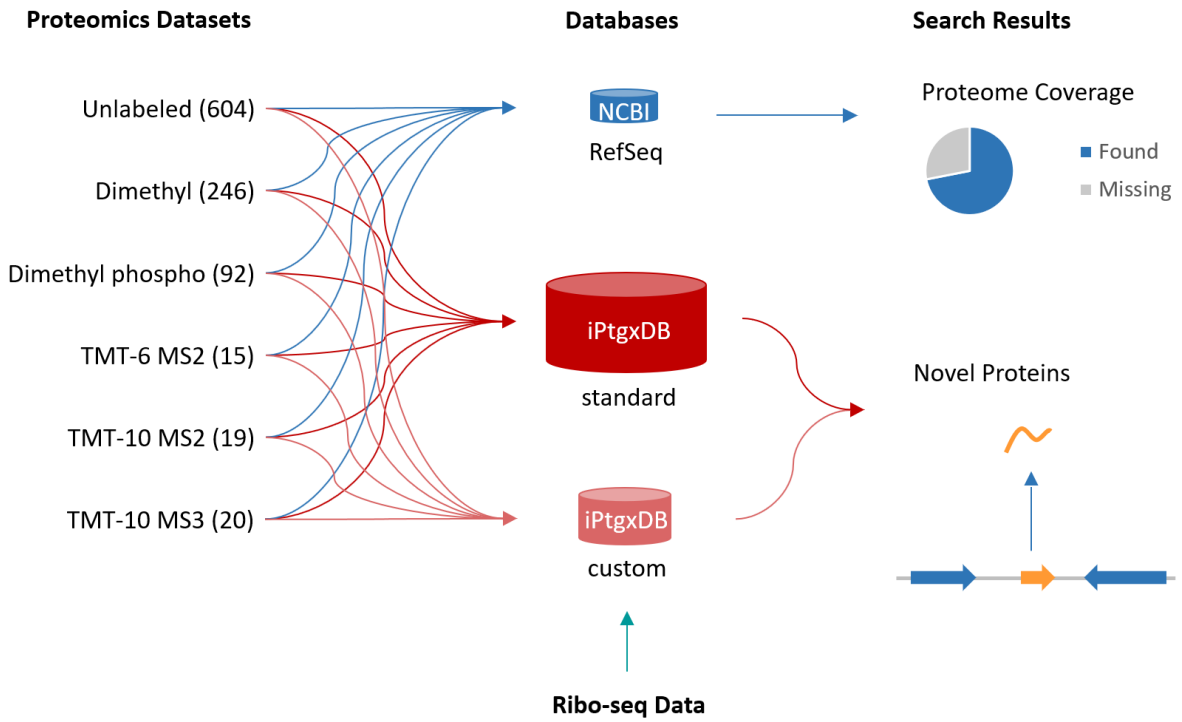


Figure S4. Proteogenomic analyses. Searches against three different databases (DBs) were carried out in order to first establish a rough estimate of the overall proteome coverage (using NCBI RefSeq as search DB) and to then provide protein expression evidence for so far unannotated protein-coding genes of *Synechocystis* 6803.

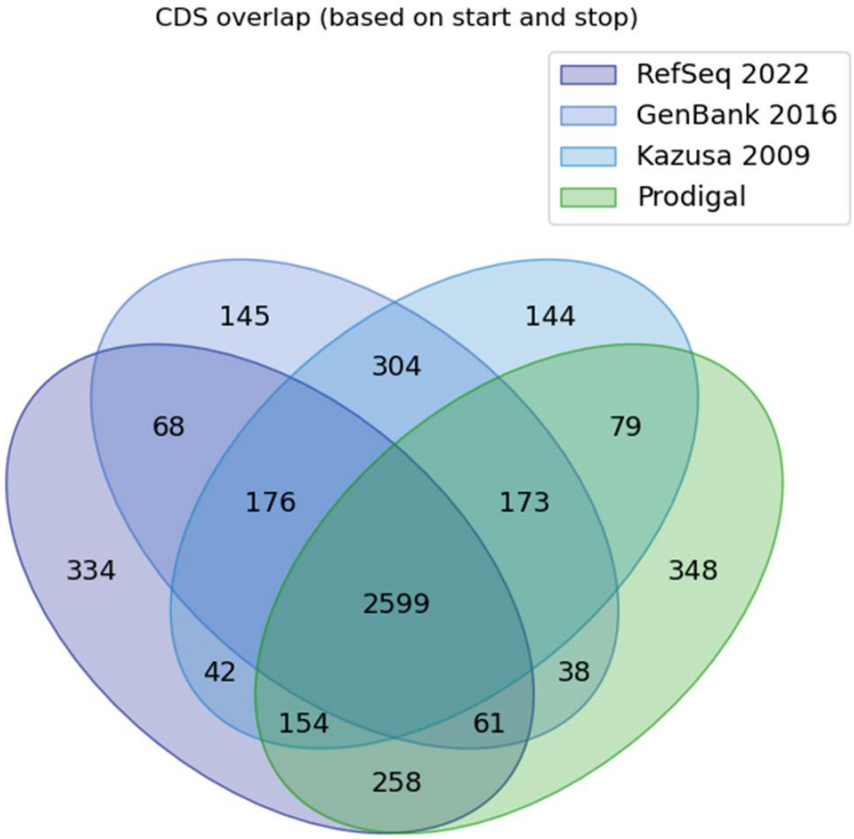


Figure S5. Overlap between different annotations of the *Synechocystis* 6803 genome. Shown are the precise matches in the number of genes sharing identical start and stop codon assignments in the NCBI RefSeq 2022 (ID: GCF_000009725.1) (2022), the GenBank (2016) and a version of the Kazusa (KZS) annotation from 2009, modified in-house, plus an *ab initio* gene prediction using Prodigal 2.6.3[5]. Only 2,599 genes are precisely shared between the four datasets.

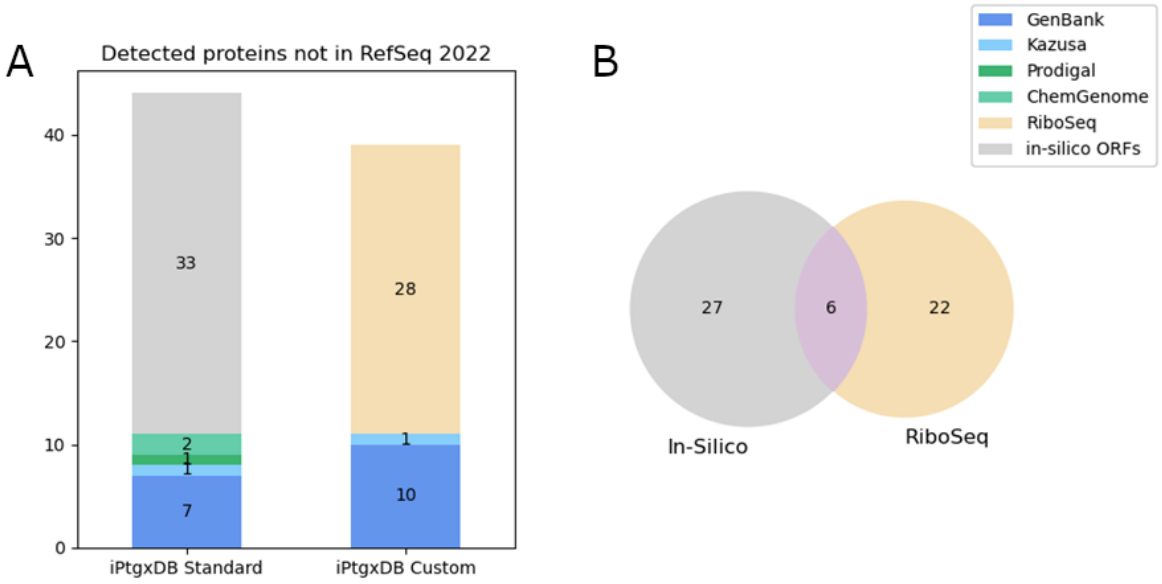


Figure S6. **Novel proteins compared to the RefSeq 2022 annotation identified by searches against a standard and custom iPtgxDB. A.** The GenBank and Kazusa identifications from the standard iPtgxDB are also among the novel proteins identified in the custom iPtgxDB. **B.** Of the 28 novel proteins identified in the custom iPtgxDB that were solely annotated based on RiboSeq data, 6 were also predicted as *in-silico* ORFs and identified in the standard iPtgxDB.

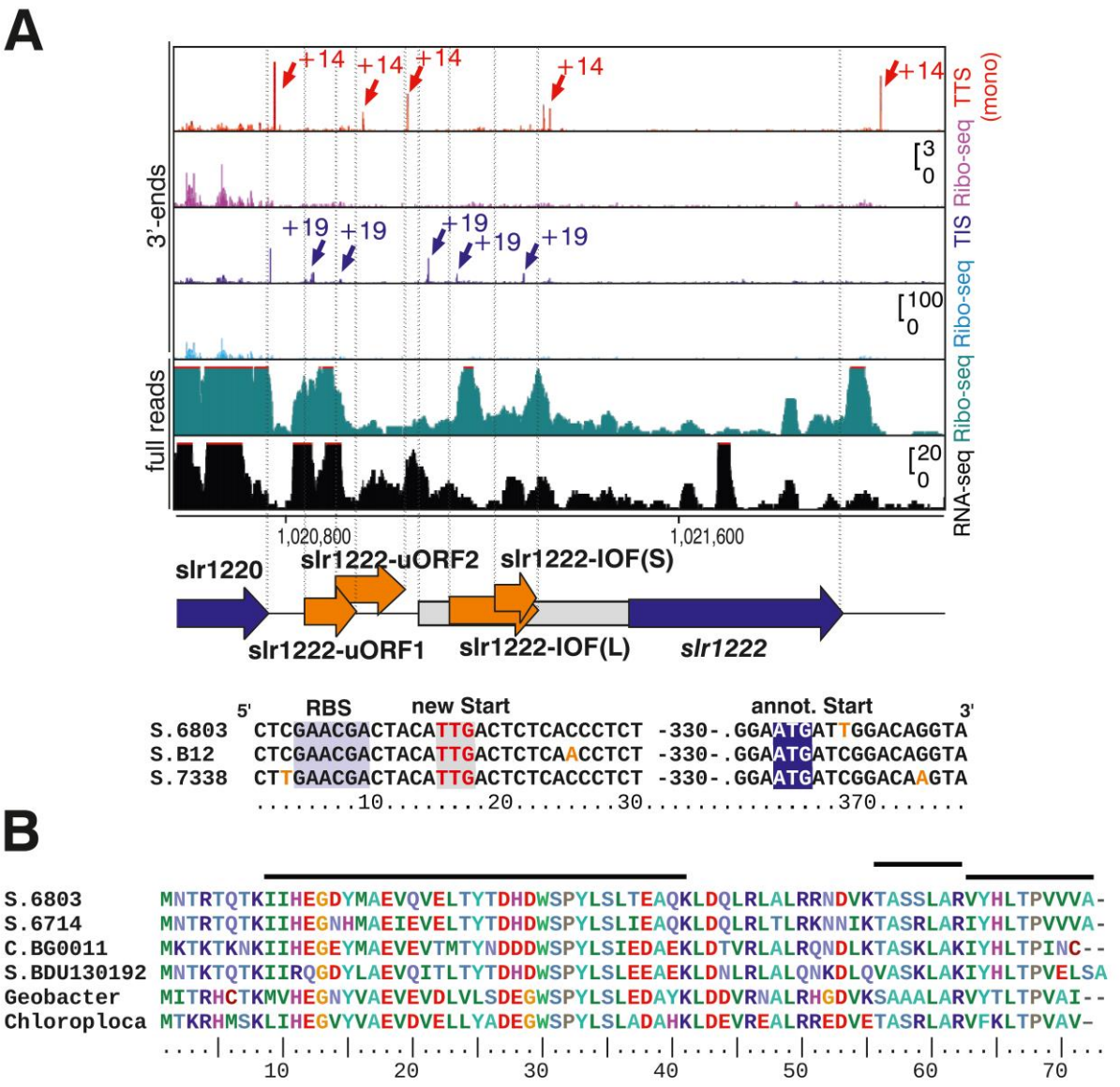


Figure S7. Ribosome profiling aids the re-annotation of genes and the discovery of novel SEPs.

A. TIS and TTS profiling data re-annotate the start codon of *slr1222* and suggest the existence of four potential novel sORFs in the associated region. TIS profiling (purple) in combination with Ribo-seq coverage (turquoise, full reads) supports an N-terminal extension by 116 aa for Slr1222, yielding 321 aa in total. DNA sequence alignments of the translation initiation regions from different *Synechocystis* strains indicated the annotated and new start codons (lower panel). In combination with Ribo-seq & TTS profiling coverage (red track on top), TIS profiling supports the detection of two internal out-of-frame, overlapping sORFs, IOF(L) and (S) (60 aa and 15 aa) within the N-terminal extension of Slr1222, and two un-annotated upstream sORFs (slr1222-uORF1 and 2, 35 aa and 44 aa, respectively) in the intergenic region between *slr1220*

and *slr1222*. Arrows in purple indicate the respective TIS signals, arrows in red TTS signals. Note, that a transmembrane domain was predicted in Slr1222-IOF(L).

B. Homologs of Ncr1610-sORF1 exist in several cyanobacteria and many other species. The multiple sequence alignment shows in addition to Ncr1610-sORF1 putative homologs from the cyanobacteria *Synechocystis* 6714 (AIE73341), *Cyanothece* sp. BG0011 (WP_107667291), *Synechococcus* sp. BDU 130192 (WP_099239310), the Pseudomonadota *Geobacter* sp. DSM 9736 (WP_088535776) and the Chloroflexota ‘Candidatus *Chloroploca mongolica*‘ (WP_135477305). Horizontal lines above the alignment indicate peptides found for Ncr1610-sORF1. This Figure extends **Figure 4F**.

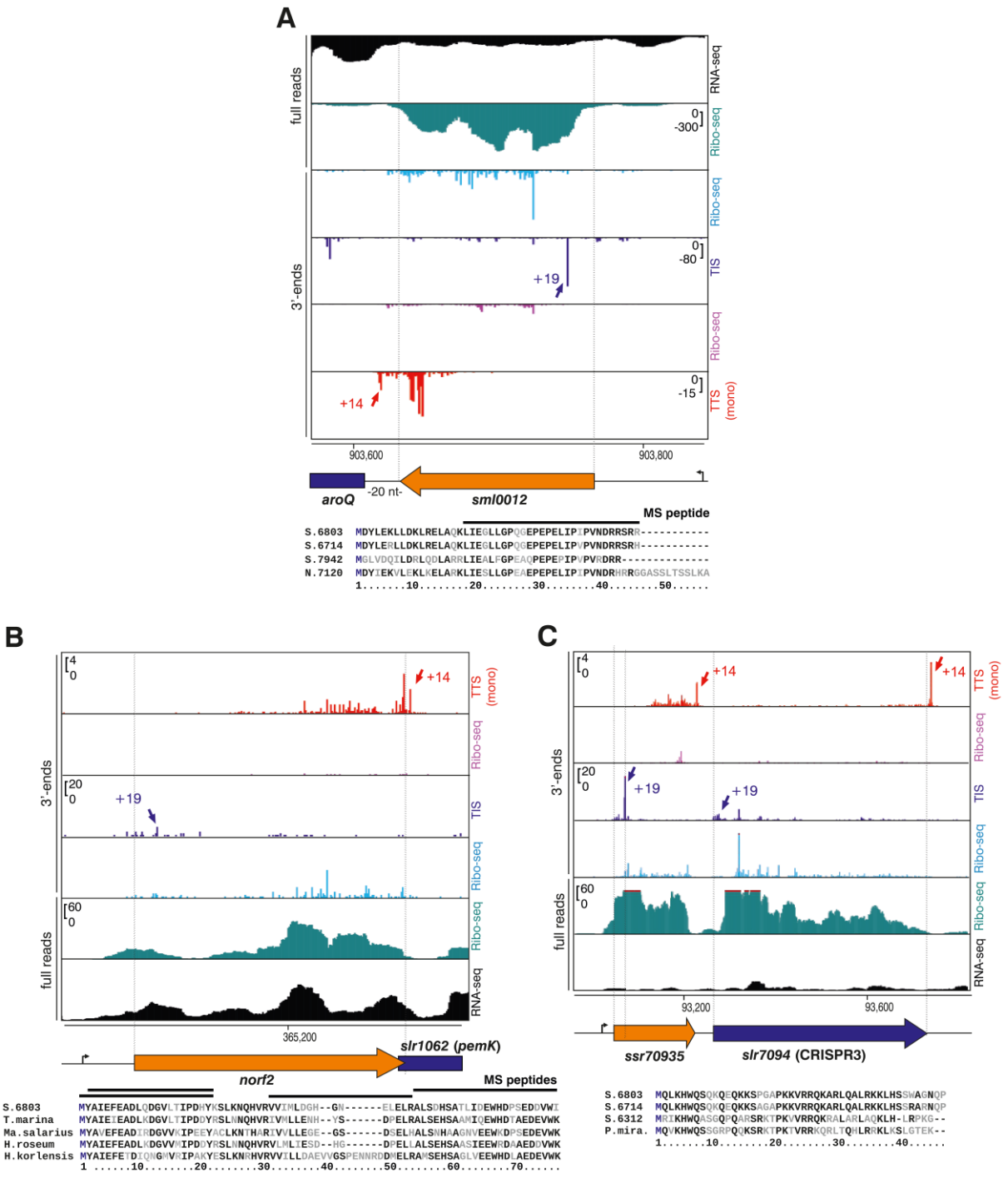


Figure S8. Small proteins encoded in sRNAs and unannotated genome regions.

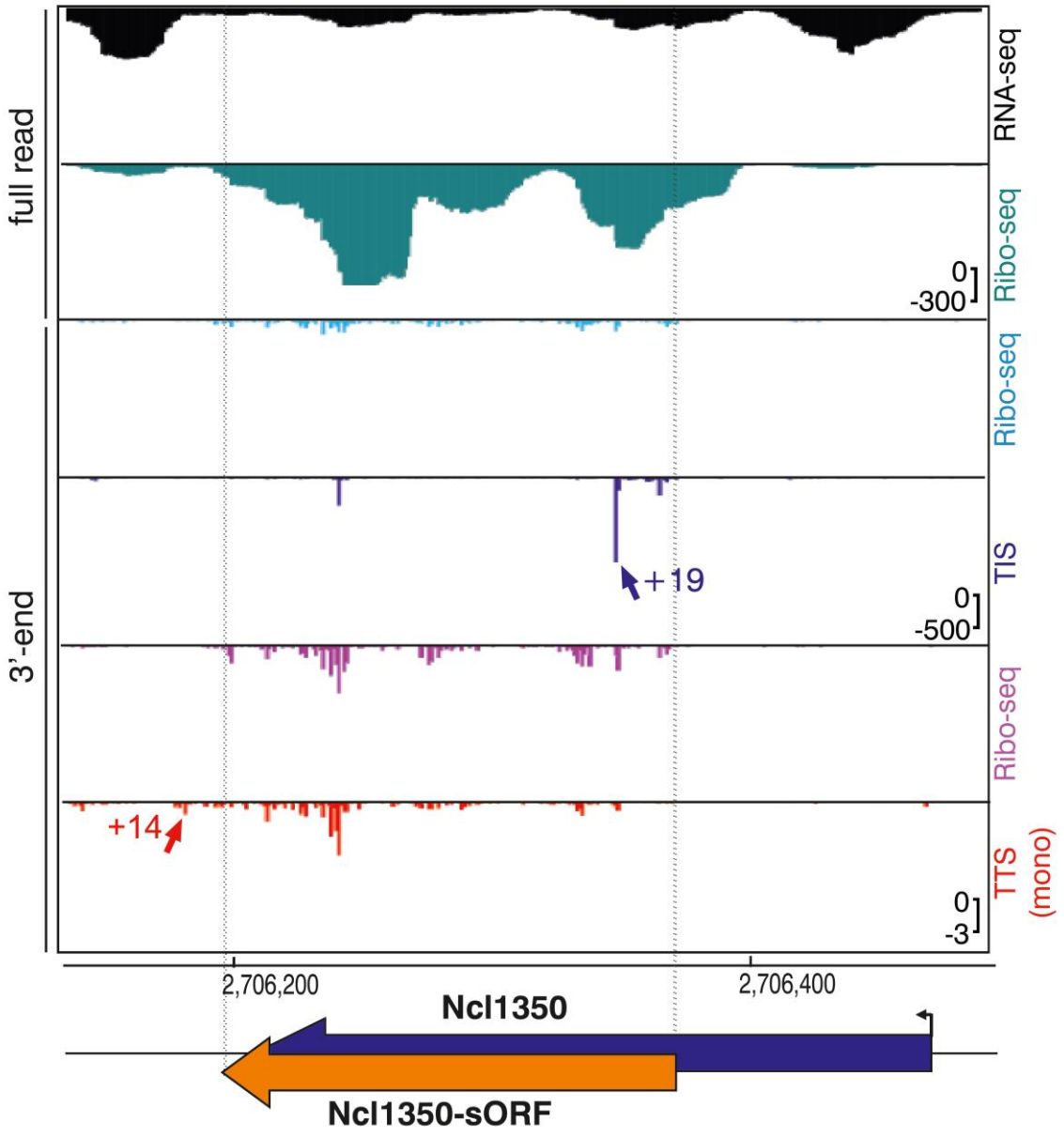
A. Sml0012 is a 45 aa protein with homologs in many other cyanobacteria. Read coverage from RNA-seq (top, black), Ribo-seq (second panel from top, turquoise), TIS- (purple) and TTS-Ribo-seq (red) indicates translation of the *sml0012* ORF. The TIS-Ribo-seq 3' ends map 19 nt downstream the start codon and the TTS-Ribo-seq 3' ends 14 downstream the stop codon (blue and red arrows). The Ribo-seq read coverage is enriched for the coding part, in contrast to the RNA-seq coverage, which starts at the transcription start site (bent arrow) 199 nt upstream the start codon. Transcription is

contiguous with the downstream located gene *aroQ* encoding the enzyme 3-dehydroquinate dehydratase that leads via the shikimate pathway to the synthesis of aromatic amino acids Y, F and Q. The multiple sequence alignment shows the comparison to Sml0012 homologs from *Synechocystis* 6714 (gene D082_00160), *Synechococcus elongatus* sp. PCC 7942 (accession WP_011242938), and *Nostoc* sp. PCC 7120 (gene *asr2781*). The location 17 to 25 nt upstream of *aroQ* is conserved in these strains. The horizontal line above the alignment indicates a peptide found for Sml0012. The protein was moreover validated after FLAG tagging by Western blotting (**Figure 5A**).

B. Norf2 was previously hypothesized as protein-coding gene *norf2* ("novel ORF 2") based on its transcription and the conserved reading frame [6]. Here, its translation as a 68 aa protein is supported by the Ribo-seq, TIS- and TTS-Ribo-seq data, by the identification of 4 peptides in the standard iPtgxDB that cover 86% of it (horizontal lines), and by Western blotting after adding a C-terminal FLAG tag (**Figure 5A**). We found no homologs for Norf2 in any other cyanobacteria, but some closely related proteins exist in gamma-proteobacteria. The alignment shows the comparison to homologs from *Thiocapsa marina* (WP_007191346), *Marinobacter salarius* (WP_269400048), *Halochromatium roseum* (WP_201215208), and *Halomonas korlensis* (WP_089797407). These proteins are annotated as hypothetical proteins; however, the gene location in the *Synechocystis* 6803 genome has been characterized as a genomic island [7], and a role in a toxin-antitoxin system with the overlapping gene *slr1062* encoding a PemK-type toxin[8] as an antitoxin is likely.

C. The here defined gene *ssr70935* is located directly downstream the CRISPR3 array, between *slr7093* and *slr7094*, on plasmid pSYSA[9], upstream of *slr7094* and *slr7095* with which it is cotranscribed as part of TU7087[10], hence forming a tricistronic operon. Its translation is supported by strong Ribo-seq, TIS- and TTS-Ribo-seq coverage and by Western blotting after FLAG tagging (**Figure 5A**). In the lower part, Ssr70935 is compared to homologs from *Synechocystis* 6714 (gene D082_40650), *Synechocystis* sp. PCC 6312 (AFY60879), *Synechococcus* sp. PCC 6312 (gene Syn6312_1729), and *Petrachloros mirabilis* (WP_275072620).

.




```
S.6803     MFWKLRQVSRFWFCYLNQPIFSRESKSTWRLSHFWSMYQISLLESCWQKSP-IKETSL-
S.6714     MFWKLRQGFRFWFCYLNQPLFSRESKATWRLTHFWSMYQISLLESCWQKSP-IKETSL-
Lept.7376  MFTRVRQVVDPLFQYLNQPIGKSEQCAVWNPNRFWYLYKINLLEKCWLKDPAVKGQKAY
Glt.7822   MMWKIRHFCFAFWLYLNQPLEGSESQSVWHMSRFWYLYKIQFLETCLEKDINSESHYTQ
           1 ........10........20........30........40........50.......
```

Figure S9. Ncl1350 was originally classified as an sRNA[10], but actually contains a translated sORF of 57 aa as indicated by the coverage from Ribo-seq, TIS- and TTS-Ribo-seq that was after SPA tagging validated by Western blotting (**Figure 5A**). The TSS was mapped to position 2,706,438 yielding a 5'UTR of 71 nt. Homologues of similar size exist in many different cyanobacterial species, here shown in the alignment with putative proteins from *Synechocystis* 6714, *Leptolyngbya* sp. PCC 7376 and *Gloeothece verrucosa* PCC 7822. These proteins are annotated (accession numbers AIE75908, AFY36466 and ADN14821.

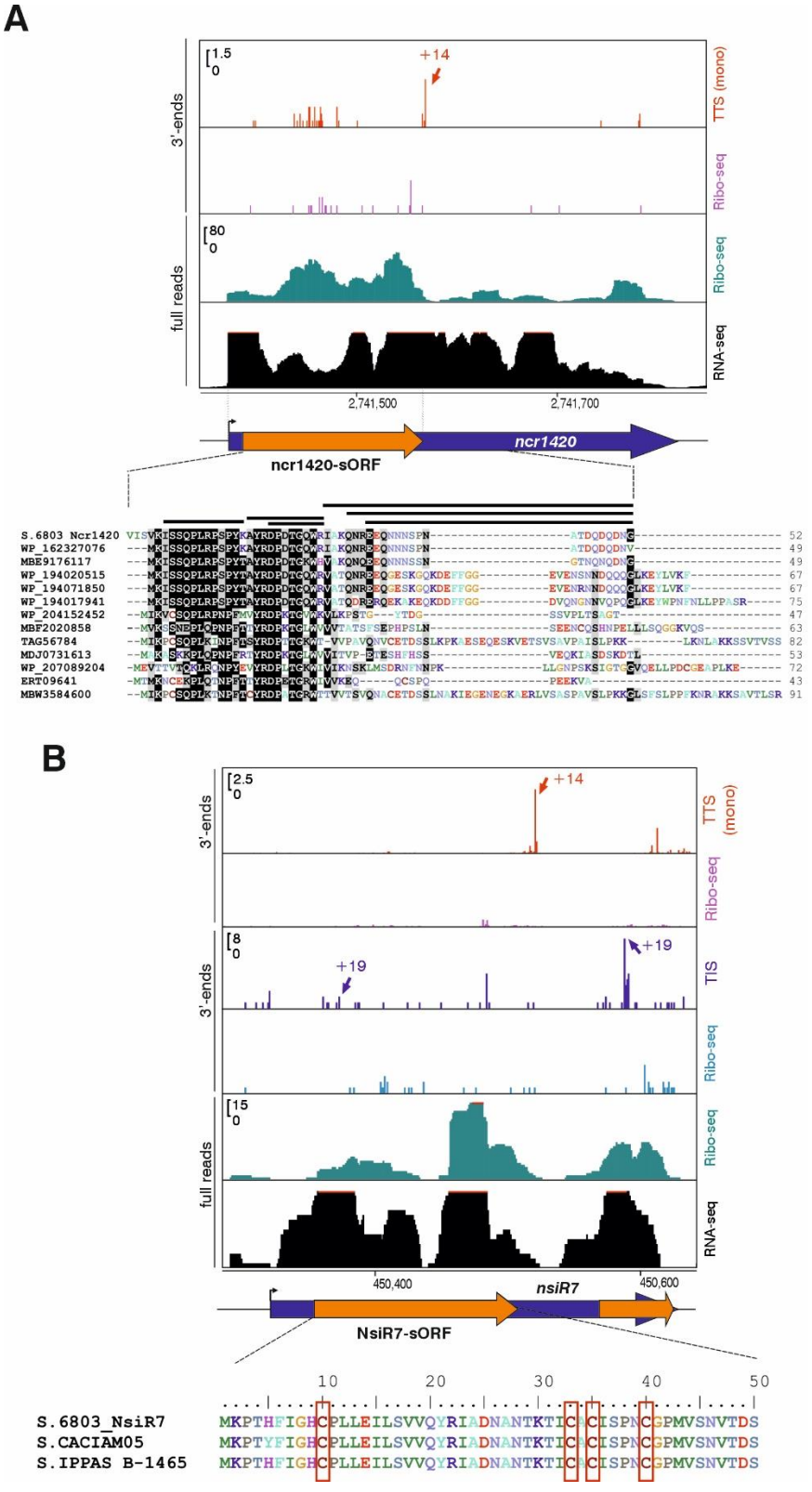


Figure S10. Homologs of small proteins encoded on previously defined sRNAs in *Synechocystis* 6803, as predicted on the basis of TIS- and TTS-Ribo-seq data. **A.** The sRNA Ncr1420 (synonym SyR21) was originally classified as a non-coding RNA within TU2899[10], but actually is translated into a 52 aa SEP, supported by TTS data

with a 3’ end offset of 14 nt (red arrow) and confirmed after SPA tagging by Western blotting **(Figure 5A**). Moreover, six different peptides were identified in the iPtgxDB. The gene was annotated in RefSeq 2022 as SGL_RS19480. Homologues of similar size exist in several different cyanobacteria, the alignment shows the comparison to homologs from *Synechocystis* sp. CACIAM 05 (WP_162327076), *Synechocystis salina* LEGE 06155 (MBE9176117), *Synechocystis salina_1* (WP_194020515), *Synechocystis* sp. LEGE 06083 (WP_194071850), *Synechocystis salina_2* (WP_194017941), *Leptolyngbya* sp. CCY15150 (WP_204152452), *Hydrococcus* sp. (C42_A2020_068), *Hydrococcus* sp. C42_A2020_068 (MBF2020858), *Oscillatoriales cyanobacterium* (TAG56784), *Crocosphaera* sp. (MDJ0731613), *Phormidium pseudopriestleyi* (WP_207089204), *Lyngbya aestuarii* BL J (ERT09641), and *Cyanobacteria bacterium* 0813 (MBW3584600). Several more putative homologs can be identified by TBlastN.

B. The sRNA NsiR7 was originally classified as non-coding RNA Ncr0210 in TU3138[10]. It belongs to the 33 most abundant ncRNAs in *Synechocystis* 6803[11] and is strongly induced under nitrogen starvation and was therefore renamed to nitrogen starvation induced RNA 7 (NsiR7), controlled by the transcription factor NtcA[12]. It contains a translated sORF encoding a 50 aa protein, confirmed by Western blotting after 3xFLAG tagging **(Figure 5A**). Homologs of similar size exist in several different *Synechocystis*, an identical protein is encoded in *Synechocystis* sp. IPPAS B-1465. The alignment shows the comparison to homologs in *Synechocystis* sp. CACIAM 05 (CA05) and *Synechocystis* S. IPPAS B-1465. A downstream encoded very short sORF potentially encodes a protein of 17 aa (NsiR7-dORF), for which homologs can be predicted in *Synechocystis* sp. IPPAS B-1465 and *Synechocystis* sp. CACIAM 05, but not in *Synechocystis* 6714. Four conserved cysteine residues are boxed. Both sORFs are supported by TIS data with a 3’end offset of 19 nt and a clear TTS signal with a 3’ end offset of 14 nt in case of NsiR7 (red arrow).

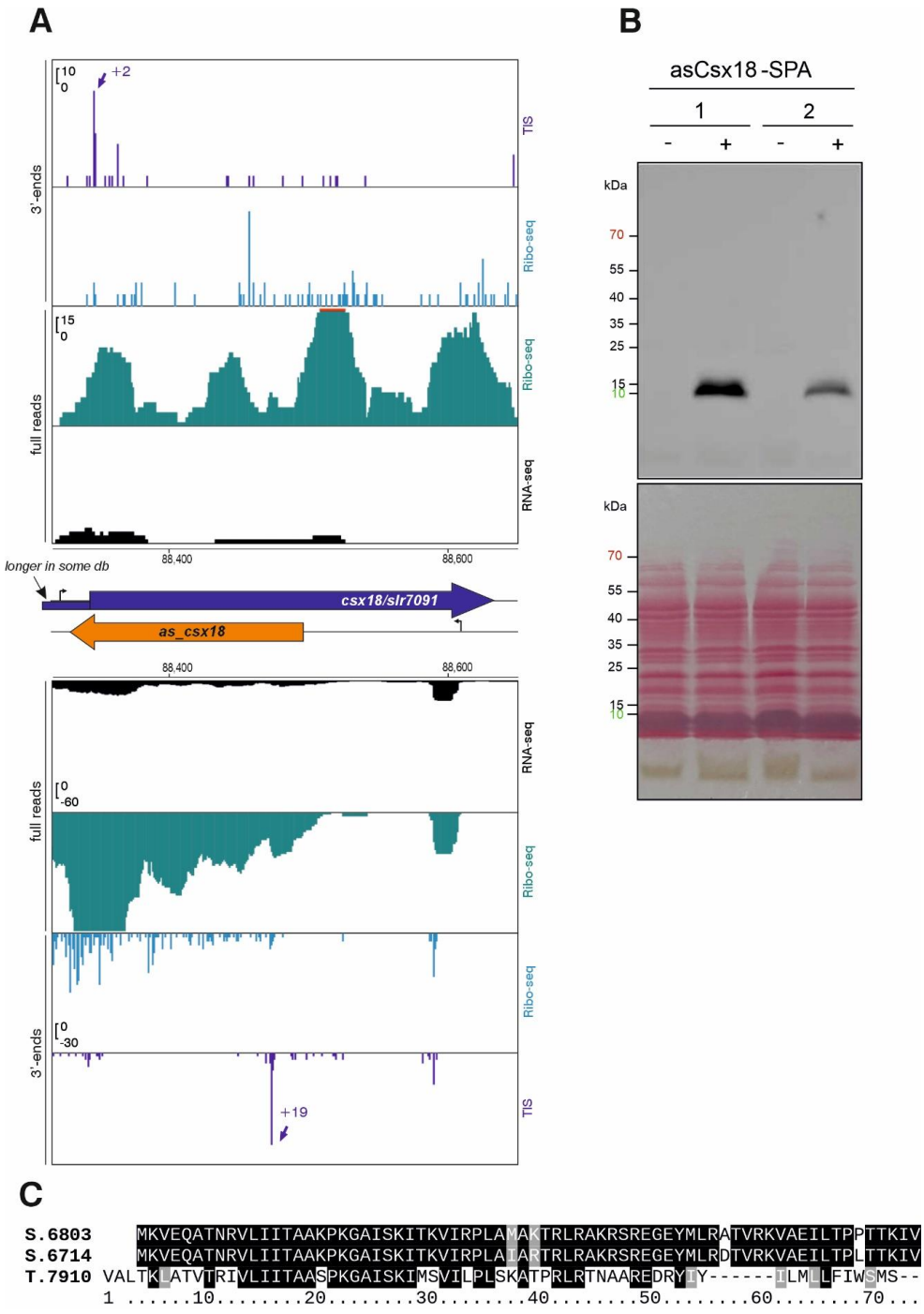


Figure S11. Coding potential in the asRNA as_csx18 to gene *slr7091* encoding a CRISPR-associated Csx18 protein. A. Both DNA strands are expressed, supported by RNA-seq (black), Ribo-Seq (turquoise), and TIS-Ribo-seq (purple) datasets. The characteristic TIS-Ribo-seq signals 3' of the start codons were mapped at position +19 for the sORF in *as_csx18* and at position +2 (which was one of the offsets found for the TIS-Ribo-seq 3′ end signals, cf. **Figure 1E**) for *slr7091*. Note, that Slr7091 is annotated in the KZS, RefSeq and GBK databases as a 122 aa protein, but according to our data it is translated from a shorter ORF encoding a 96 aa protein, consistent

with the Prodigal prediction. This example illustrates how experimental data overlaid on top of integrated annotations from different genome centers can be used to consolidate the annotation differences. **B.** Western Blot verification of the dual SPA-tagged asCsx18 protein expressed from plasmid pVZ322 under control of the P_{petE} promoter induced by copper addition to 1.25 µM (+). Samples without induction served as negative controls (-). In this gel, ten µg protein was loaded per lane, separated on a 12% SDS-PAA gel, run for 22 h. The antiserum was M2 anti-FLAG HRP (Sigma-Aldrich A8592). The lower panel shows the ponceau red-stained membrane. **C.** The sequence alignment shows possible homologs of the 70 aa asCsx18 protein encoded by the *as_csx18* in *Synechocystis* 6714 and *Tolypothrix* sp. PCC 7910 identified by TBlastN in antisense orientation to their respective *csx18* genes. The respective ORFs are shown in full lengths. The gene in *Synechocystis* 6714 is located on plasmid pSYLA[13] and according to dRNA-seq data it is also associated with an asRNA[14]. All three *csx18* loci are located close to a CRISPR array (distance of three genes or less).

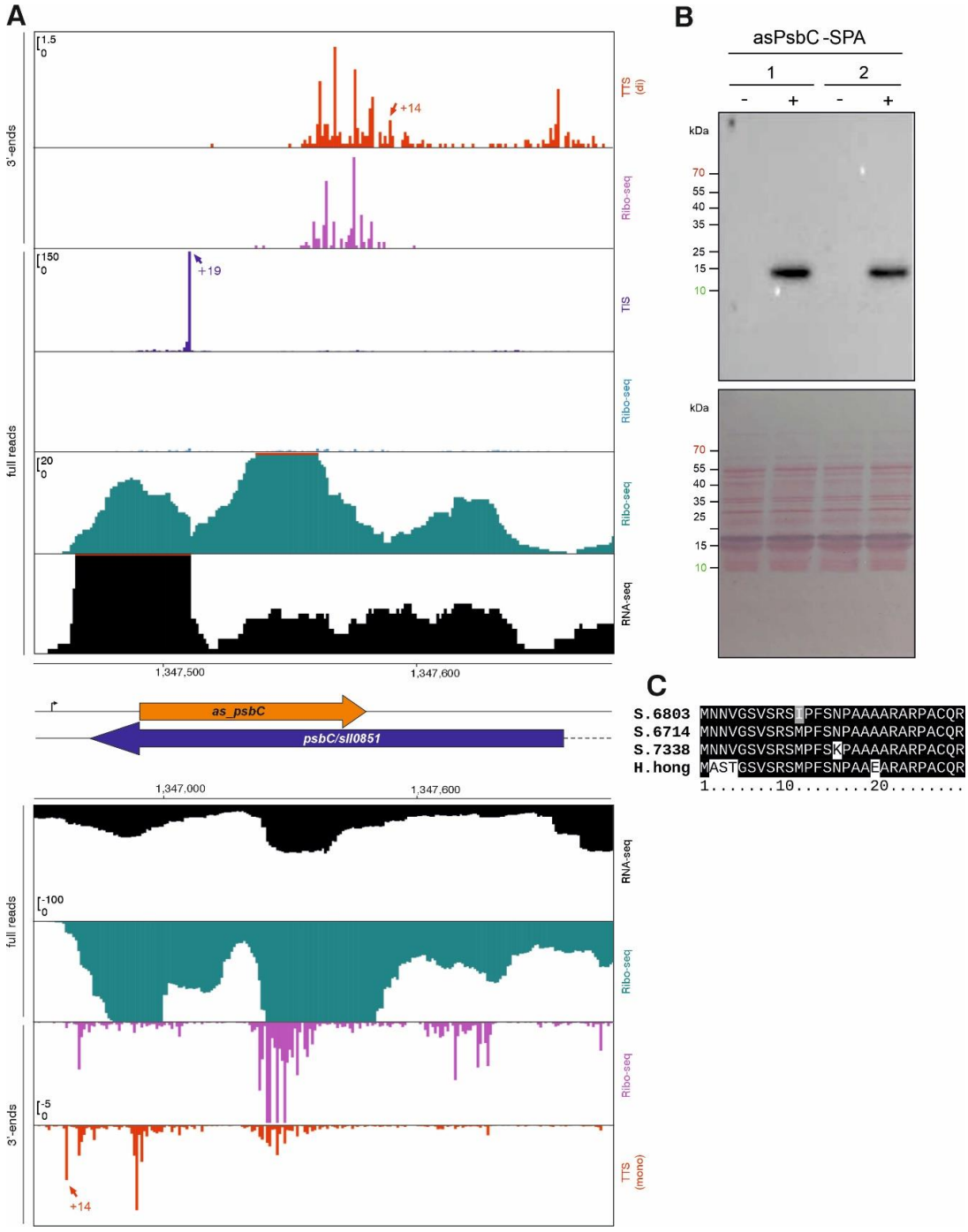


Figure S12. Coding potential for a 28 aa protein in the asRNA *as_psbC* to gene *sll0851* encoding the photosystem II protein PsbC/CP43. Both DNA strands are expressed, supported by RNA-seq (black), Ribo-Seq (turquoise), TIS-Ribo-seq (purple) and TTS-Ribo-seq (red) datasets. The characteristic TIS-Ribo-seq signal located 19 nt 3' of the start codon in the *as_psbC* transcript were mapped at position +19 (out of the shown region for *psbC*). TTS-Ribo-seq signals located 14 nt 3' of the stop codons in the as_psbC transcript and *psbC* mRNA are labeled by red arrows.

B. Western Blot verification of the SPA-tagged asPsbC protein expressed from plasmid pVZ322 under control of the P_{petE} promoter induced by copper addition to a final concentration of 1.25 µM (+). Samples without induction served as negative controls (-). In this gel, 3 µg protein was loaded per lane, separated on a 15% SDS-PAA gel, run for 22 h. The antiserum was M2 anti-FLAG HRP (Sigma-Aldrich A8592). The lower panel shows the ponceau red-stained membrane.

C. The sequence alignment shows possible homologs of the asPsbC protein in *Synechocystis* strains 6714 and 7338, and in *Halomicronema hongdechloris* C2206 identified by TBlastN in antisense orientation to their respective *psbC* genes. However, it is unknown if corresponding asRNAs exist in these strains and the respective ORFs are also longer than the displayed segments.

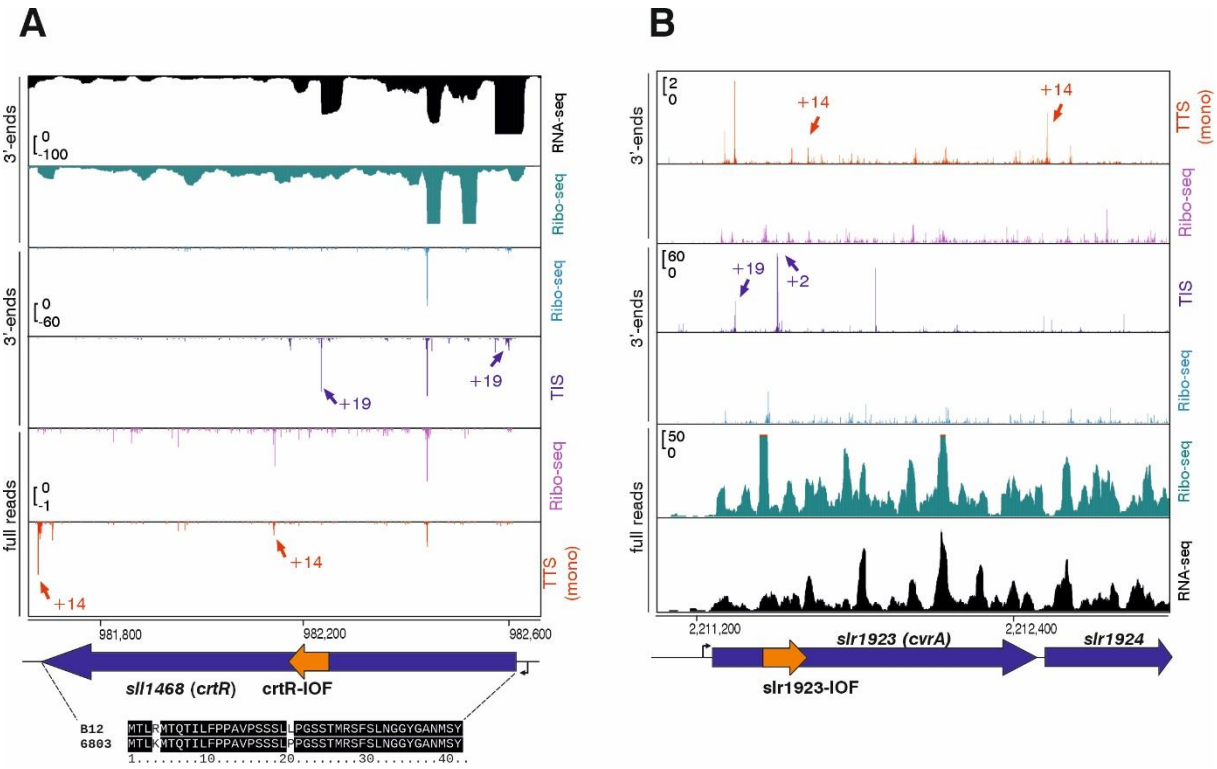
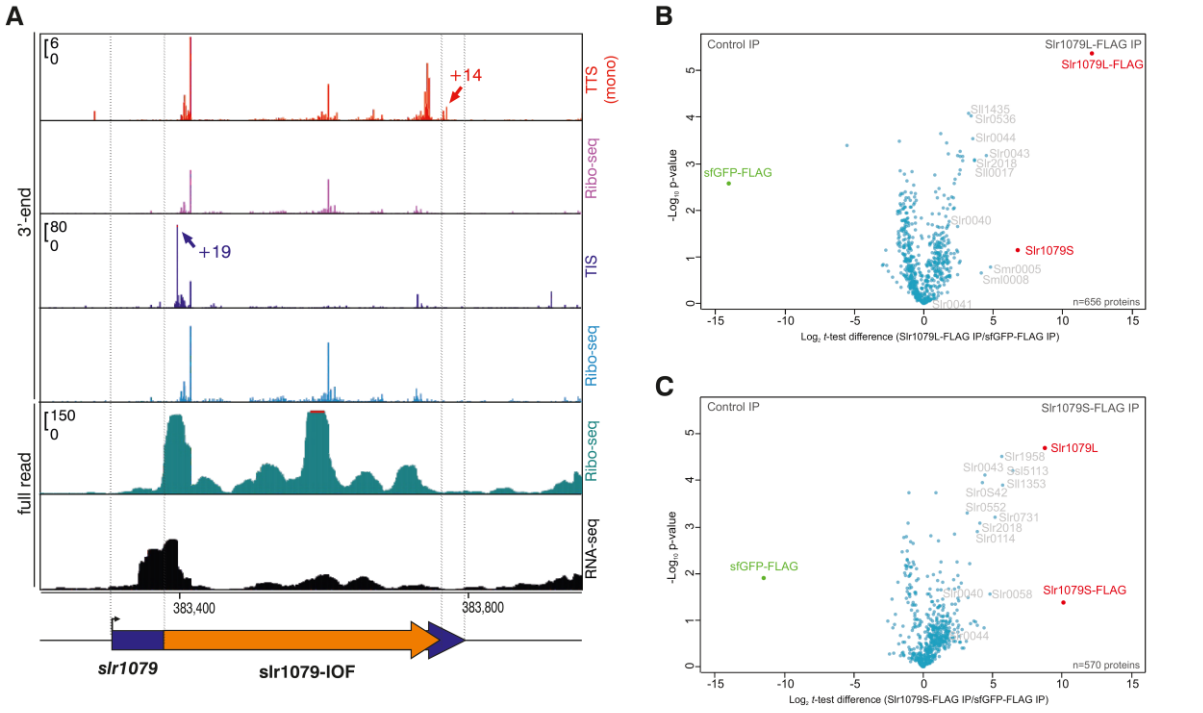


Figure S13. sORFs identified in mRNAs out-of-frame. Alternative, gene-internal out-of-frame (IOF) start sites were predicted for **A.** *crtR* and **B.** *cvrA* on the basis of TIS-RiboSeq (blue) and TTS-RiboSeq (red) data. TIS-RiboSeq reads 3'ends accumulated close to the potential internal start codon with a 3'end offset of 19 nt for the main genes and the IOF in *crtR/sll1468*, while a 3'end offset of 2 nt was detected for the IOF in *cvrA*/*slr1923*. Both main genes encode important enzymes of pigment biosynthesis. CrtR is the beta-carotene hydroxylase converting beta-carotene to zeaxanthin[15], CvrA is the vinyl reductase essential for the conversion from divinylchlorophyll(ide) to normal chlorophyll(ide)[16,17]. The sORF in *crtR* can also be identified in *Synechocystis* sp. B12, but generally these IOF sORFs are not conserved. In all panels, the Ribo-seq read coverage (turquoise) is enriched for the coding segments, in contrast to the RNA-seq coverage (black), which starts at the position of the transcription start site (bent arrow). The here identified sORFs are indicated by orange arrows.



D

>Slr1079L
MASFLALLPRSLTTFLYAVAALLRFYGNIDTTPIPRIPLTIFGWSFLAFTLGTAALLVNLGLEWNTGNRSRNREIETRERETRRD
NLADEERNRASEEREKADRERDRADQERNRADQERKRAASRARIQNRGFVLQTRYQLAPSPEARATLIDFLSFLQEYGE

>Slr1079S
MATSTPPPSPESPSPSLDGASWRSRWAQRLSWLISALSGTLEIEAETERLKQENAKLDEIILQTKKETERLRKEKKPIEKEIE
QIKKETEQIRKENERLAVLESKIEDLFSKHGINLPLAQKPEQP

E

>Slr1082L
MASFLALLPRSLTTFLYAVAALLRFYGNIDTTPFQLIPLTILDWSFLAFTLGTAALLVNLGLEWNSGNRSRNRADQERDRTDQE
RDRADQERQRAARRARIQNRFFLRPTRYQLTPSPDTRAALIDFLSFLQEYGE

>Slr1082S
MATSTPPPSNSSPSPSSTGVSWRSRWAQRLSWLISALSGTLEIEAETEQIRKETEQIKKETEQIRKDNERLAVLESKIDSSFD
QHGINLPLAQTPEQP

Figure S14. A protein Slr1079S is encoded by a long IOF (orange arrow) within the *slr1079* reading frame (blue arrow). A. Read coverage from Ribo-seq and RNA-seq libraries is shown in the lower 2 panels. The Ribo-seq read coverage is clearly higher at the beginning of the IOF segment. Mapped signals from TIS-Ribo-seq at the characteristic position 19 nt 3' of the respective start codon (two central panels including control), and 14 nt 3' of the respective stop codon from TTS-Ribo-seq (two upper panels including control) are highlighted by the blue and red arrows.

B. Co-IP analysis of Slr1079L. The volcano plot shows the protein interaction partners of Slr1079L-3xFLAG identified by mass spectrometry compared to sfGFP-3xFLAG. Triplicate samples were analyzed (**Figure S18, S19, S20**). The two most enriched proteins, Slr1079L and Slr1079S are colored red, the locus IDs are given for co-

enriched proteins in grey. For the complete list of 656 detected proteins and further details, see **Table S8**.

C. Co-IP analysis of Slr1079S. Details as in panel B. For the complete list of 570 detected proteins, see **Table S9.** The results of an analogous co-IP experiment for *slr0489* encoding Slr0489L/S are shown in **Figure 6D**. **D.** Sequences of Slr1079L and Slr1079S. The black horizontal lines indicate proteogenomically detected peptides for both proteins directly proving their simultaneous presence in the cells. **E.** Sequences of Slr1082L and Slr1082S. Proteogenomically detected peptides were only found for Slr1082S (horizontal lines).

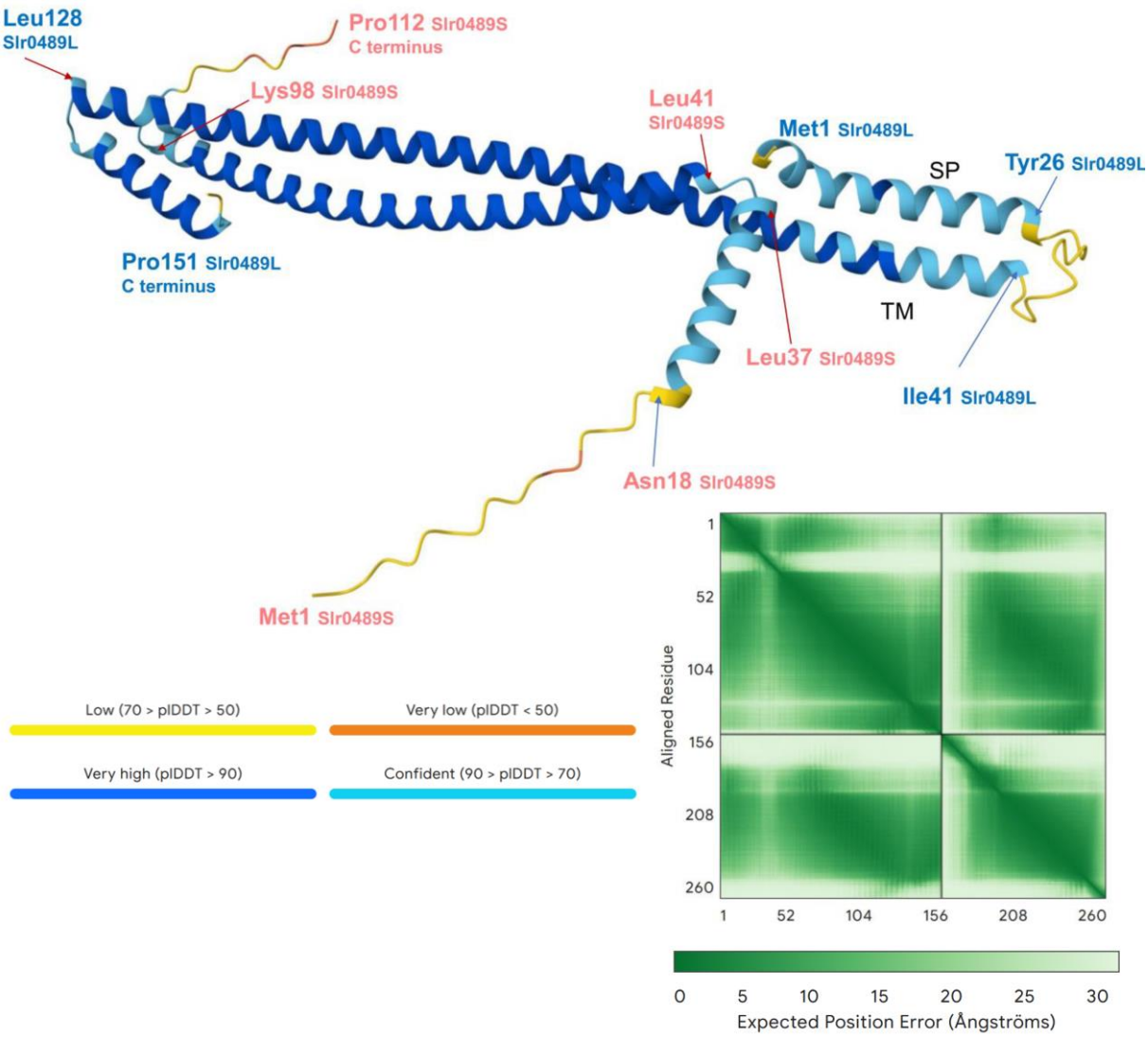


Figure S15. Alphafold prediction of the Slr0489L/Slr0489S heterodimer. Alphafold models long helical regions in both proteins which align over ~50% of their respective lengths. Several residues are given for orientation for Slr0489L (light blue) and Slr0489S (rose). The two predicted membrane-spanning regions in Slr0489L are indicated (TM1 and TM2). Support scores were ipTM = 0.58 and pTM = 0.59, note that support was with ipTM = 0.65 and pTM = 0.67 better for the heterooctamer (4 copies of Slr0489L and Slr0489S each). Here, the heterodimer is shown for better clarity.

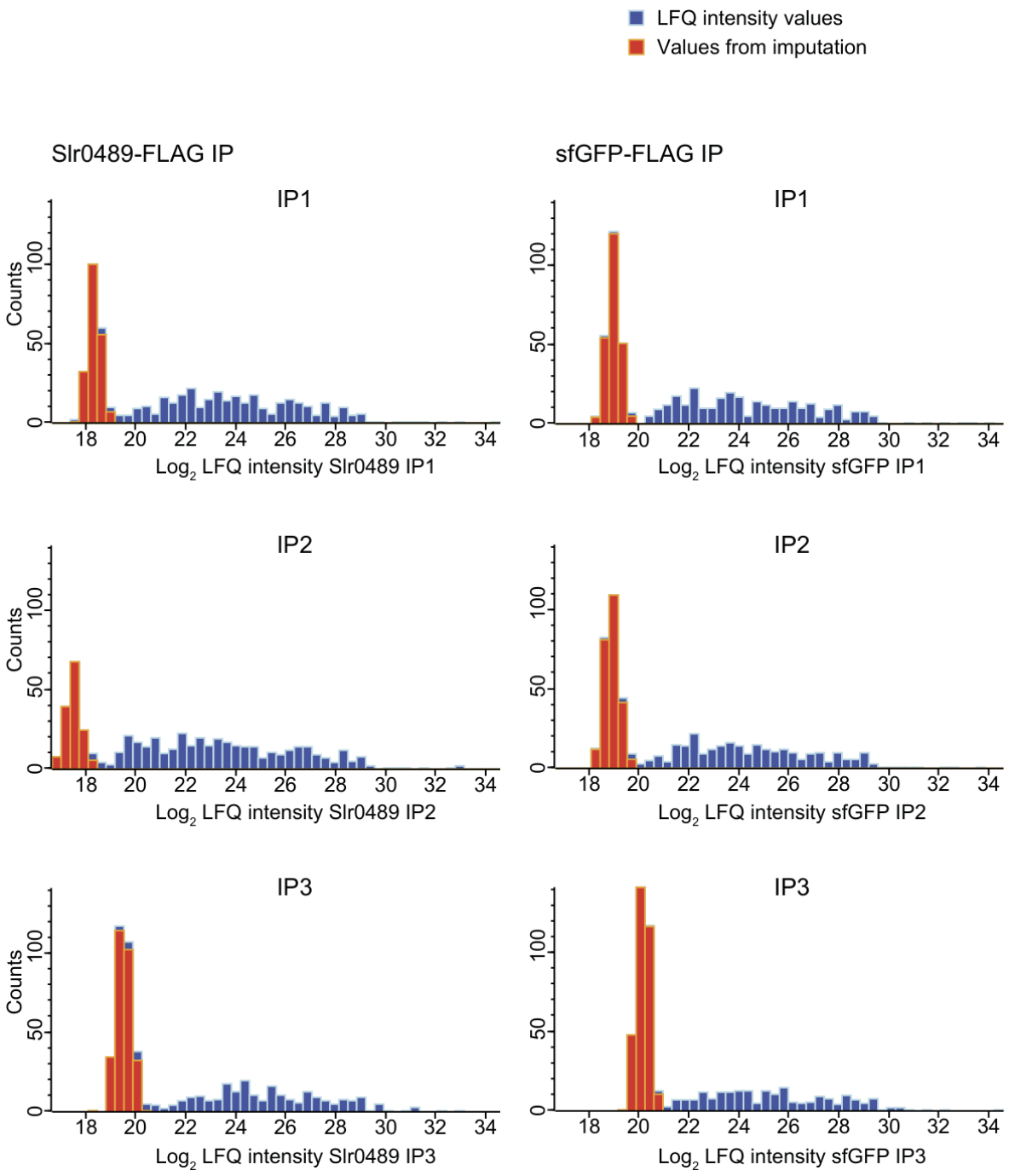


Figure S16. Histograms of the LFQ data distribution for the co-IP with Slr0489L-3xFLAG compared to sfGFP-3xFLAG. This figure extends **Figure 6D** and **Table S6**.

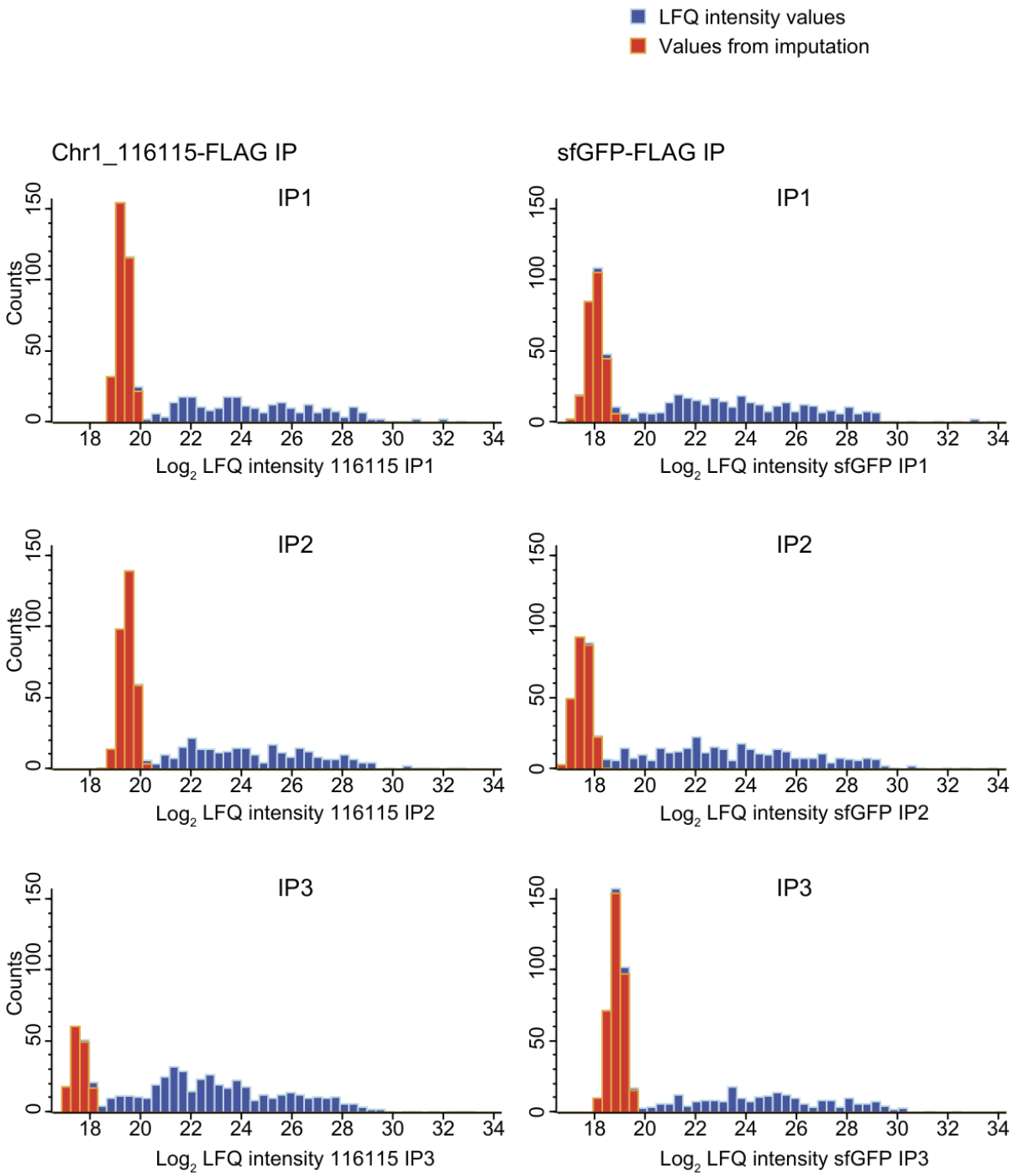


Figure S17. Histograms of the LFQ data distribution for the co-IP with Slr0489S-3xFLAG compared to sfGFP-3xFLAG. This figure extends **Figure 6E** and **Table S7**.

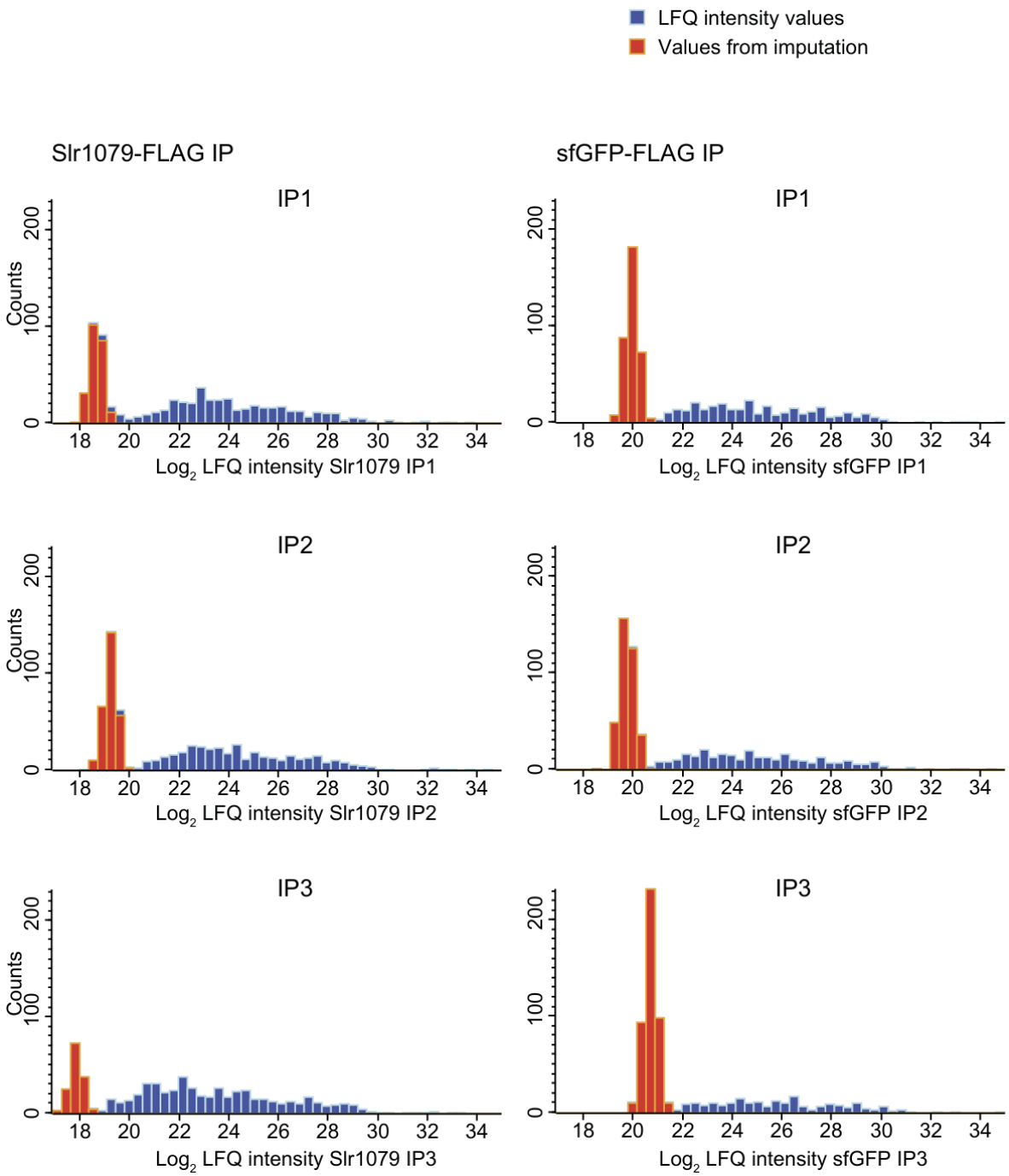


Figure S18. Histograms of the LFQ data distribution for the co-IP with Slr1079L-3xFLAG compared to sfGFP-3xFLAG. This figure extends **Figure S14B** and **Table S8**.

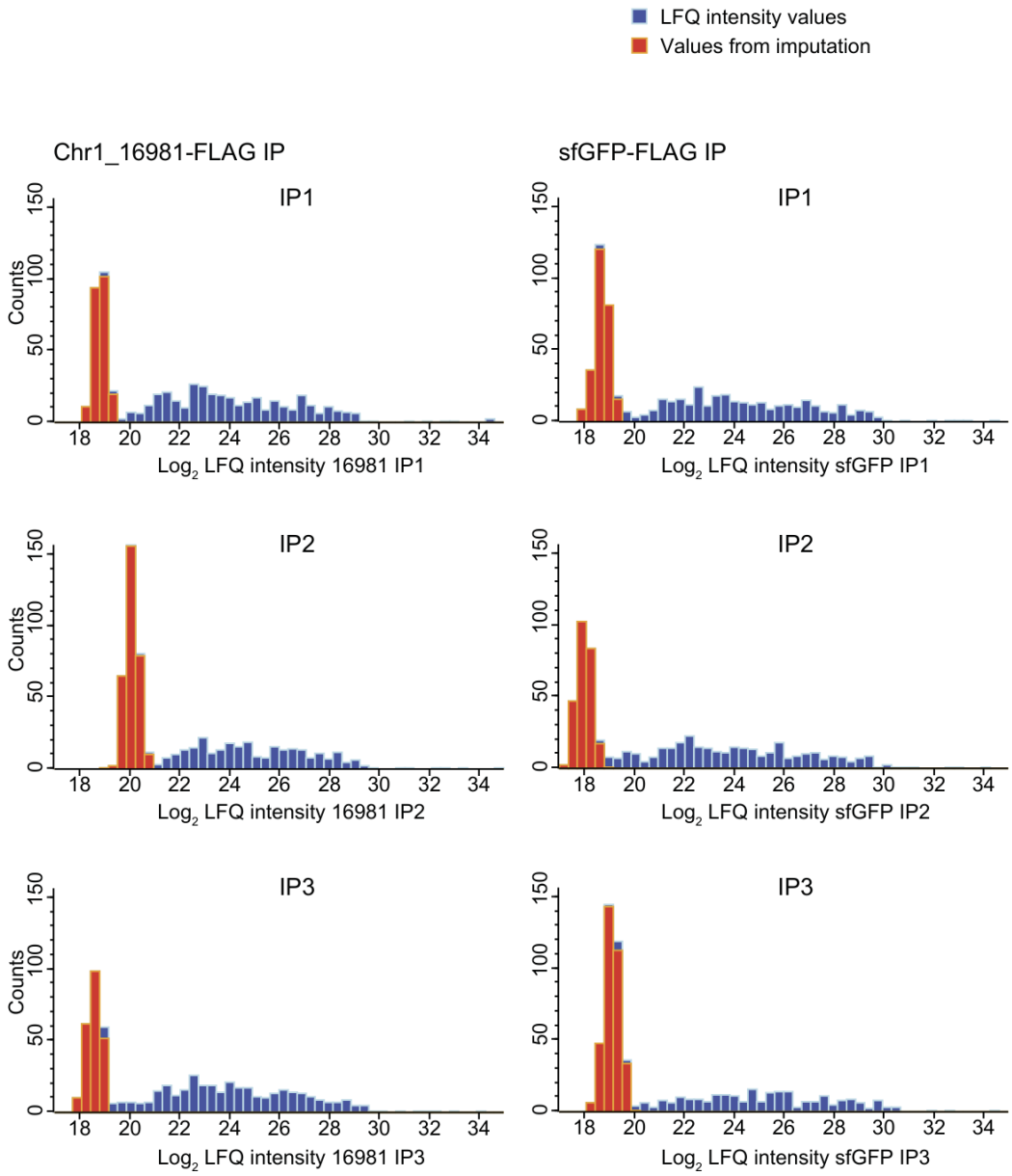


Figure S19. Histograms of the LFQ data distribution for the co-IP with Slr1079S-3xFLAG compared to sfGFP-3xFLAG. This figure extends **Figure S14C** and **Table S9**.

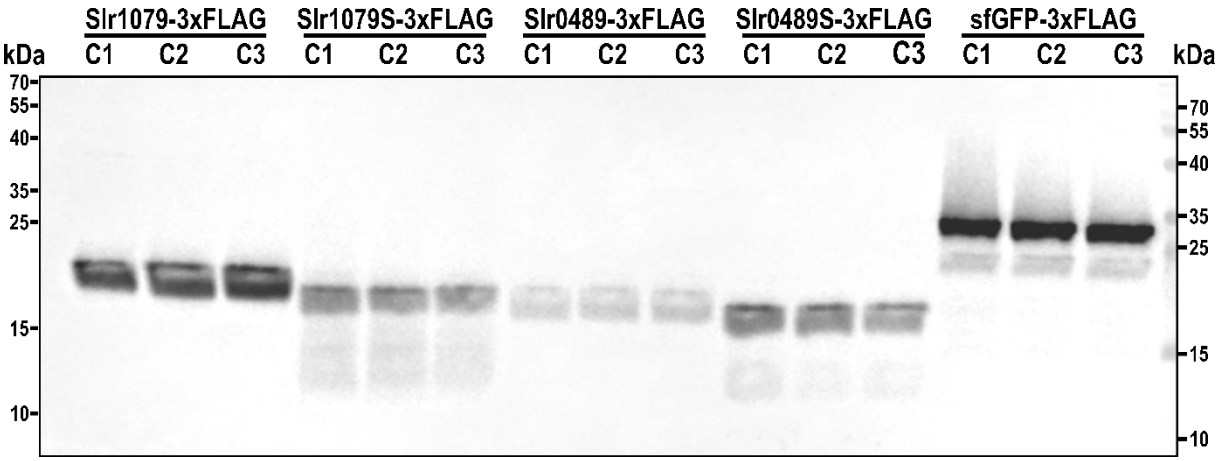


Figure S20. Western blot of overexpressed Slr0489L/S and Slr1079L/S. The genes *slr0489* and *slr1079* were amplified and cloned from the *Synechocystis* 6803 genome and inserted into expression vector pVZ322s under the control of the P_{rha} promoter and with a sequence encoding a C-terminal 3xFLAG tag. In parallel, the IOF for each gene was cloned with their native 5´UTRs into the same expression vector. The constructs were transformed into wild-type *Synechocystis* 6803. A construct encoding tagged sfGFP was used as control. The expression was induced upon addition of 0.6 mg/mL L-rhamnose for 24 h before harvesting. Co-immunoprecipitation assays were performed and bound proteins were eluted with 100 ng/µL 3xFLAG-peptide. For each elution, 20 µL were loaded on a 12% SDS-PAA gel. The 3xFLAG tagged proteins were revealed by Western Blotting with M2 anti-FLAG antiserum. Each strain was tested in triplicates (C1-3). This figure extends **Figure 6D/E** and **Figure S14B/C**.

Supplementary References

1. 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.

2. 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.

3. 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.

4. Oliveira, P., Martins, N.M., Santos, M., Pinto, F., Büttel, Z., Couto, N.A., Wright, P.C., and Tamagnini, P. (2016). The versatile TolC-like Slr1270 in the cyanobacterium *Synechocystis* sp. PCC 6803. Environ. Microbiol. *18*, 486–502. https://doi.org/10.1111/1462-2920.13172.

5. 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.

6. Mitschke, J., Georg, J., Scholz, I., Sharma, C.M., Dienst, D., Bantscheff, J., Voß, B., Steglich, C., Wilde, A., Vogel, J., et al. (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. Proc. Natl. Acad. Sci. USA *108*, 2124–2129. https://doi.org/10.1073/pnas.1015154108.

7. 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.

8. 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.

9. 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.

10. 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.

11. 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.

12. Giner-Lamia, J., Robles-Rengel, R., Hernández-Prieto, M.A., Muro-Pastor, M.I., Florencio, F.J., and Futschik, M.E. (2017). Identification of the direct regulon of NtcA during early acclimation to nitrogen starvation in the cyanobacterium *Synechocystis* sp. PCC 6803. Nucleic Acids Res. *45*, 11800–11820. https://doi.org/10.1093/nar/gkx860.

13. Kopf, M., Klähn, S., Voss, B., Stüber, K., Huettel, B., Reinhardt, R., and Hess, W.R. (2014). Finished genome sequence of the unicellular cyanobacterium *Synechocystis* sp. strain PCC 6714. Genome Announc. *2*. https://doi.org/10.1128/genomeA.00757-14.

14. Kopf, M., Klähn, S., Scholz, I., Hess, W.R., and Voß, B. (2015). Variations in the non-coding transcriptome as a driver of inter-strain divergence and physiological adaptation in bacteria. Sci. Rep. *5*, 9560. https://doi.org/10.1038/srep09560.

15. Masamoto, K., Misawa, N., Kaneko, T., Kikuno, R., and Toh, H. (1998). Beta-carotene hydroxylase gene from the cyanobacterium *Synechocystis* sp. PCC6803. Plant Cell Physiol. *39*, 560–564. https://doi.org/10.1093/oxfordjournals.pcp.a029405.

16. Islam, M.R., Aikawa, S., Midorikawa, T., Kashino, Y., Satoh, K., and Koike, H. (2008). *slr1923* of *Synechocystis* sp. PCC6803 is essential for conversion of 3,8-divinyl(proto)chlorophyll(ide) to 3-monovinyl(proto)chlorophyll(ide). Plant Physiol. *148*, 1068–1081. https://doi.org/10.1104/pp.108.123117.

17. Ito, H., Yokono, M., Tanaka, R., and Tanaka, A. (2008). Identification of a novel vinyl reductase gene essential for the biosynthesis of monovinyl chlorophyll in *Synechocystis* sp. PCC6803. J. Biol. Chem. *283*, 9002–9011. https://doi.org/10.1074/jbc.M708369200.