The structure of a highly conserved picocyanobacterial protein reveals a Tudor domain with a novel tRNA binding function

Katherine M Bauer[1], Rose Dicovitsky[2], Maria Pellegrini[2], Olga Zhaxybayeva[3,4] and Michael J Ragusa[1,2*]

[1]Department of Biochemistry & Cell Biology, Geisel School of Medicine at Dartmouth, Hanover NH 03755; [2]Department of Chemistry, Dartmouth College, Hanover NH 03755; [3]Department of Biological Sciences, Dartmouth College, Hanover NH 03755; [4]Department of Computer Science, Dartmouth College, Hanover NH 03755

[*]To whom correspondence should be addressed: Michael J. Ragusa: Department of Chemistry, Dartmouth College, Hanover NH 03755; michael.j.ragusa@dartmouth.edu

Keywords: NMR spectroscopy, *Prochlorococcus*, *Synechococcus*, nucleic acid binding, Tudor domains

---

ABSTRACT

Cyanobacteria of the *Prochlorococcus* and marine *Synechococcus* genera are the most abundant photosynthetic microbes in the ocean. Intriguingly, the genomes of these bacteria are very divergent even within each genus, both in gene content and at amino acid level of the encoded proteins. One striking exception to this is a 62 amino acid protein, termed *P*rochlorococcus/*S*ynechococcus Hyper Conserved Protein (PSHCP). PSHCP is not only found in all sequenced *Prochlorococcus* and marine *Synechococcus* genomes but it is also nearly 100% identical in its amino acid sequence across all sampled genomes. Such universal distribution and sequence conservation suggests an essential cellular role of the encoded protein in these bacteria. However, the function of PSHCP is unknown. We used Nuclear Magnetic Resonance (NMR) spectroscopy to determine its structure. We found that 52 of the 62 amino acids in PSHCP form a Tudor domain, while the remainder of the protein is disordered. NMR titration experiments revealed that PSHCP has only a weak affinity for DNA, but an 18.5 fold higher affinity for tRNA, hinting at an involvement of PSHCP in translation. Computational docking and mutagenesis studies identified a positively charged patch surrounding residue K30 that serves as the primary docking site for tRNA on PSHCP. These results provide the first insight into the structure and function of PSHCP and suggest a new function for Tudor domains in recognizing tRNA.

With the mean annual average of 3.6×10^{27} cells, cyanobacteria from genera *Prochlorococcus* (1) and *Synechococcus* (2) numerically dominate microbial communities of the global ocean (3). Because of such abundance, these photosynthesizing bacteria are extremely important players in the global carbon cycle (3). The small genomes and limited gene content of individual *Prochlorococcus* and *Synechococcus* cells (1.6-2.9 Mb, encoding ~1700-3100 genes; (4)) are counterbalanced by a diverse collective gene pool, which in the case of *Prochlorococcus* is estimated to contain 80,000 genes (5,6). The genomes of *Prochlorococcus* and *Synechococcus* are also surprisingly divergent: in pairwise comparisons, the genome-wide average amino acid identity (AAI; (7)) of the encoded proteins is often below 60% even within each genus (8). Maintenance of such remarkable sequence and gene content divergence remains poorly understood (6). However, one protein-coding gene is a curious exception, with its product showing almost 100% amino acid sequence conservation across all currently available *Prochlorococcus* and marine *Synechococcus* genomes ((9,10) and Supplementary Datasets 1

and 2). This gene encodes a 62 amino acid protein of unknown function, dubbed PSHCP for *Prochlorococcus*/*Synechococcus* Hyper Conserved Protein, (9,10). Although at the time of its discovery the gene encoding PSHCP was limited to the cyanobacterial clade consisting of *Prochlorococcus* and marine *Synechococcus*, sequence similarity searches now identify the PSHCP gene in genomes of freshwater *Synechococcus*; several *Cyanobium* spp. from both freshwater and marine environments; in cyanobacterial sponge symbiont "Candidatus *Synechococcus spongarium*" (11); in three species of *Paulinella*, which is a photosynthetic protist with a chromatophore thought to be derived from marine *Synechococcus* spp. (12); and in metagenomically-assembled cyanobacteria from globally distributed marine, brackish and freshwater environments. Within cyanobacteria, all characterized PSHCP-containing organisms form a clade (13) within the order *Synechococcales* (14). Hence, the PSHCP protein remains extremely conserved in a specific subgroup of cyanobacteria (Supplementary Datasets 1 and 2), but is undetectable outside of this subgroup.

The remarkable conservation of the PSHCP protein and retention of this gene even within extremely reduced chromatophore genomes (12) suggest that this protein may encode an important housekeeping function. In three examined strains of *Prochlorococcus* and marine *Synechococcus*, the gene is expressed, and its protein product is abundant (10). A large proportion of positively charged amino acids (16%) (9), a predicted isoelectric point of 11.3 (9), and the protein's association with the ribosomal protein L2 in pull-down assays (10) prompted hypotheses of PSHCP involvement in binding of either DNA or RNA, and its possible association with a ribosome. Yet, the amino acid sequence of PSHCP shows no significant sequence similarity to any protein domain with a known function.

To gain more insight into the possible function of PSHCP, we determined the structure of the PSHCP protein using Nuclear Magnetic Resonance (NMR) spectroscopy and investigated the protein's interactions with nucleic acids *in vitro*.

RESULTS

The structure of PSHCP

We began by screening different constructs of PSHCP for their behavior for structure determination. An initial round of construct screening revealed that residues 57-62 could be removed without altering the overall stability of the protein. This is in good agreement with disorder prediction using the IUPred server (15), which suggests that residues 57-62 are disordered. $PSHCP_{1-56}$ could be easily concentrated above 1 mM making it ideal for structure determination. Secondary structure probabilities and predicted order parameters S2, as calculated using TALOS-N (16) from ^{1}H,^{13}C,^{15}N chemical shifts, show that the first three residues of $PSHCP_{1-56}$ are disordered and that the remainder of the protein contains five β-strands (Fig. S1A and S1B). To further investigate the dynamics of $PSHCP_{1-56}$ we performed a ^{1}H-^{15}N heteronuclear Nuclear Overhauser Effect (NOE) experiment. In the ^{1}H-^{15}N heteronuclear NOE experiment a peak intensity ratio closer to one indicates little motion of the N-H bond on the picosecond to nanosecond timescale while a smaller peak intensity ratio corresponds to more motion on this time scale and thus residues which are disordered. The ^{1}H-^{15}N heteronuclear NOE data recorded on $PSHCP_{1-56}$ is in good agreement with the secondary structure probabilities and predicted order parameter data, confirming that the structured core of the protein corresponds to residues 4-56 (Fig. S1C).

For the structure calculation of $PSHCP_{1-56}$, 1195 NOE derived distance constraints and 44 dihedral angle constraints were used (Table 1). Excluding the first three amino acids, this resulted in 22.1 distance constraints per amino acid. The 20 lowest energy structures align with an average pairwise backbone RMSD of 0.72 ± 0.15 Å for residues 4-29 and 35-56 (Fig. 1A). Residues 30-34 form a flexible loop between β-strands two and three, and thus were excluded from the RMSD calculation. The overall fold of PSHCP contains five β-strands, which form a single antiparallel β-sheet with an overall barrel-like architecture (Fig. 1B and 1C).

To determine if PSHCP is part of a conserved domain family, we used the Dali server (17) to search the Protein Data Bank (PDB) for

proteins with structural similarity to PSHCP. PSHCP aligns well to a number of Tudor domains. The highest structural similarity is to the Tudor domain from the human TDRD3 protein: the 52 amino acids of PSHCP (residues 4-56) align to it with a backbone RMSD of 1.5 Å (Fig. 2A-C). Tudor domains are small, approximately 60 amino acid long, domains containing four or five β-strands in a barrel-like fold. Most known Tudor domains are found in eukaryotes (98.97% of the InterPro records for Tudor domain), where they are often present in 2-3 adjacent copies within large, multidomain proteins (18). In contrast, full length PSHCP consists entirely of a single Tudor domain (residues 4-56).

PSHCP is unlikely to be involved in recognition of methylated proteins

The best-characterized functions of Tudor domains are to bind methylated lysine and arginine residues (18,19). This frequently results in the targeting of Tudor domain containing proteins to histones to regulate DNA damage responses and chromatin remodeling. The interaction of Tudor domains with methylated residues occurs through a conserved aromatic cage (Fig. 2D) (18,20). This aromatic cage is composed of four aromatic residues, which are individually located in β-strands one through four of Tudor domains. However, PSHCP contains only one aromatic residue in any of the positions corresponding to the aromatic cage residues and is therefore unlikely to bind methylated amino acids (Fig. 2E).

PSHCP has a weak affinity for DNA

Tudor domains have also been shown to regulate DNA function by directly binding DNA (21,22). In PSHCP, ten of 62 amino acids are positively charged (Fig. 3A). The majority of these residues cluster in one of two locations on the surface of PSHCP, generating two distinct positively charged patches that may support DNA binding (Fig. 3B). To determine if PSHCP is able to bind to DNA, we incubated ^{15}N-labeled $PSHCP_{1-56}$ with different concentrations of a double stranded GC-rich DNA sequence (dsGC DNA) that has been previously shown to bind the double interdigitated Tudor domain of retinoblastoma-binding protein 1 (RBBP1) (21). When titrating dsGC DNA into PSHCP seven peaks shifted in the $PSHCP_{1-56}$ ^{1}H-^{15}N HSQC spectrum, demonstrating that $PSHCP_{1-56}$ can bind to DNA (Fig. 4A). As only seven peaks shifted upon DNA binding, this suggests that $PSHCP_{1-56}$ does not undergo large structural changes upon DNA binding. The peaks for L14, E15, S16, G20 and V22 all shifted upon DNA binding. These five residues all sit in the loop between β-strand 1 and 2 suggesting that this region is important for DNA binding.

To determine the dissociation constant (K_d) for $PSHCP_{1-56}$-DNA binding, we monitored chemical shift perturbations (CSPs) at the four residues which undergo the largest CSPs in PSHCP (E15, G20, V22, T46) over a concentration range of 50 µM to 500 µM dsGC DNA (Fig. 4B and Table 2). The calculated K_d values ranged from 121.3 to 173.3 µM with an average K_d of 140.6 µM. This data demonstrates that PSHCP has a weak affinity for double stranded DNA. In addition, this K_d value is similar to the previously measured 84 µM K_d of the double interdigitated Tudor domain of RBBP1 for the same DNA sequence (21). It is interesting to note that despite the extended DNA binding surface in RBBP1 this resulted in only a 1.7 fold higher affinity for DNA than the single Tudor domain of PSHCP.

PSHCP binds tRNA with a low micromolar affinity

While most Tudor domains that have been studied are involved in the regulation of DNA, some Tudor domains are located in RNA-binding proteins and are involved in the regulation of RNA metabolism (19,23,24). In addition, the knotted Tudor domain from *S. cerevisiae* Esa1 has been shown to bind oligo-RNAs (25), and more recently the Tudor domain of *Escherichia coli* ProQ was shown to interact with the small non-coding RNA SraB (26). All of this supports a broader role for Tudor domains in the regulation of nucleic acids, including through direct binding of RNA.

Intriguingly, the PSHCP gene in completely sequenced genomes of marine *Synechococcus* and *Prochlorococcus marinus* spp. is flanked on the 5' end by a gene encoding

a tryptophan tRNA and on the 3' end by genes encoding an aspartic acid tRNA and the glutamyl-tRNA synthetase (Fig. S2 and refs. (9,10)). Additionally, in *P. marinus* str. MED4 and MIT9313 and in *Synechococcus* sp. WH8102, the PSHCP gene is sometimes co-transcribed with the Trp-tRNA gene (10). Since genes in bacteria are often clustered and co-transcribed based on their joint functionality in a process or a pathway, we hypothesized that PSHCP may interact with tRNA. To test this, we titrated a mixture of tRNA isolated from *E. coli* into ^{15}N labeled PSHCP and monitored CSPs in a series of ^{1}H-^{15}N HSQC spectra (Fig. 5A). At a 1.6:1 mixture of tRNA and PSHCP$_{1-56}$ almost every peak in the ^{1}H-^{15}N HSQC had broadened to the point where it was no longer visible (Fig. 5B). This suggests that PSHCP and the tRNA are forming a stable complex and that the peaks in the ^{1}H-^{15}N HSQC broadened due to the increased molecular weight from the approximately 27 kDa tRNA. Only five peaks did not completely broaden during the tRNA titration and these peaks corresponded to residues at the termini of PSHCP (residues 1-4 and 56), which therefore remain flexible after binding to tRNA. Due to the substantial broadening of the peaks in the ^{1}H-^{15}N HSQC we were only able to monitor CSPs for residues 2-4 which still had visible peaks after completion of the titration. Calculated K_d values for PSHCP$_{1-56}$-tRNA binding ranged from 2.9 to 10.4 µM with an average K_d of 7.6 µM (Fig. 5C and Table 3). As such, PSHCP displays an 18.5-fold higher affinity for tRNA than dsDNA, suggesting a potential role for PSHCP in protein translation.

The structured core of PSHCP coordinates tRNA binding

To identify regions of PSHCP that may be important for tRNA binding, we monitored the peak intensity ratio between the mixture of PSHCP$_{1-56}$ with 7.4 µM tRNA and the unbound PSHCP$_{1-56}$ (Fig. 5D). As PSHCP binds to tRNA, the peaks undergo broadening and, as a result, the intensity of the peak decreases. The majority of the peaks had bound/unbound intensity ratios around 0.4, including many of the residues that cluster into the two positively charged surfaces on PSHCP. This suggests that these surfaces may play a role in tRNA binding (Fig. 3B).

To determine if the C-terminal tail of PSHCP (residues 57-62) plays a role in tRNA binding we repeated the HSQC titration experiment with ^{15}N labelled PSHCP$_{1-62}$. The addition of tRNA also led to significant line broadening of the PSHCP$_{1-62}$ ^{1}H-^{15}N HSQC spectra (Fig. 6A). While peak broadening was similar between PSHCP$_{1-56}$ and PSHCP$_{1-62}$, clear peak shifts at the N-terminus were no longer observed. As such, we were unable to determine a dissociation constant from these spectra. To determine a dissociation constant for both PSHCP$_{1-56}$ and PSHCP$_{1-62}$ we instead utilized isothermal titration calorimetry (ITC) (Fig. 6B, C). Dissociation constants for PSHCP$_{1-56}$ and PSHCP$_{1-62}$ binding to tRNA were 0.76 µM and 1.17 µM, respectively, suggesting that residues 57-62 do not play an important role in tRNA binding.

Residues involved in tRNA binding

To test the role of each of the positively charged patches on the surface of PSHCP in tRNA binding, one positively charged amino acid in each patch (R23 and K30) was chosen to be mutated to alanine. In addition, the loop in between β-strands 1 and 2 had large CSPs during DNA binding and this region may also play a role in tRNA binding. As such, two residues in this loop (E15 and S16) were also chosen to be mutated to alanine (Fig. 7A). We generated three double mutants in PSHCP$_{1-62}$, E15R23AA, E15K30AA and S16K30AA and performed tRNA titration experiments using these mutants. We monitored the bound/unbound intensity ratio at 5 different amino acids (L5, L14, V22, G33 and R43) that are distributed across the entire length of PSHCP.

The bound/unbound intensity ratios for all five residues in the WT protein were between 0.22 and 0.38 (Fig. 7B). For PSHCP E15R23AA the bound/unbound intensity ratios were nearly identical to the WT protein, ranging from 0.22 to 0.36. This demonstrates that this mutation has little effect on tRNA binding. The bound/unbound intensity ratios for PSHCP E15K30AA (0.56 to 0.81) and PSHCP S16K30AA (0.78 to 1.0) were significantly larger than the bound/unbound intensity ratios of the WT protein. This demonstrates that the E15K30AA and S16K30AA mutations led to

decreased peak broadening in the presence of tRNA, and therefore a reduction in tRNA binding.

To confirm these results, we collected ^{1}H-^{15}N HSQC spectra of 50 µM WT PSHCP and PSHCP mutant proteins in the presence of 90.7 µM tRNA (Fig. 7C). In the WT and E15R23AA ^{1}H-^{15}N HSQC spectra only 8 peaks are clearly visible. In contrast, in the ^{1}H-^{15}N HSQC for the E15K30AA and the S16K30AA mutant proteins there is much less broadening and as a result the majority of the peaks in the HSQC are visible confirming that PSHCP E15K30AA and S16K30AA have reduced binding for tRNA. Taken together, our mutagenesis data suggests that the positively charged patch, which surrounds K30, is the main binding site for tRNA. Since the S16K30AA showed increased unbound/bound peak intensity ratios when compared to E15K30AA, the loop between β-strands 1 and 2 may also play a secondary role in tRNA binding.

Structural model of tRNA-PSHCP interactions

To further evaluate possible tRNA binding interfaces on PSHCP, we performed computational docking using the HDOCK program (27), which is optimized for protein and nucleic acid docking. In these docking experiments *E. coli* fMet initiator tRNA (PDB: 5L4O) was docked onto $PSHCP_{4-56}$ which was kept rigid. The top five models from the docking converged, and in all five models the tRNA molecule packs against K30 (Fig. 8A), further supporting a role for this residue in tRNA binding. The tRNA molecules also pack against the loop between β-strands 1 and 2 in support of this loop as a secondary docking site for tRNA. In our docking results the tRNA is rotated into different orientations confirming that no exact orientation of the tRNA was favored. In contrast, the top five models docking tRNA with the PSHCP K30A mutant no longer converge to surround the K30A (Fig. 8B). Instead, the tRNA molecules were found to be docked at several different locations along the surface of PSHCP, suggesting that there is no longer a preferred docking site on PSHCP. These computational docking models are in good agreement with the mutagenesis experiments described above, suggesting that the positively charged surface surrounding K30 is the primary binding site on PSHCP for tRNA.

DISCUSSION

We used NMR spectroscopy to determine the structure of the highly conserved 62 amino acid cyanobacterial protein PSHCP. PSHCP residues 4-56 contain a Tudor domain, with five β-strands arranged in a barrel-like architecture. The remainder of the protein is disordered. Two positively charged patches on the surface of PSHCP hinted that it may bind directly to nucleic acids. We found that PSHCP binds to *E. coli* tRNA with low micromolar affinity and an 18.5-fold higher affinity than for double stranded DNA. This suggests that PSHCP may bind specifically to tRNA. Additionally, the sequences flanking PSHCP on both sides of cyanobacterial genomes are tRNA, further supporting a role for PSHCP in tRNA function and translation. Residues 57-62 did not enhance the binding affinity of PSHCP for tRNA and it remains to be seen what role this disordered tail may have in PSHCP function.

The majority of the Tudor domains that have been studied are eukaryotic Tudor domains that regulate histone function by interacting with methylated lysine residues (18). However, there are several examples of Tudor domains that can bind directly to nucleic acids instead of methylated proteins. These nucleic acid binding Tudor domains are broadly distributed from bacteria to eukaryotes, suggesting that the ancestral function of Tudor domains may be nucleic acid binding rather than protein binding. Structures of such nucleic acid binding Tudor domains have been determined for the methyltransferase Esa1 from *S. cerevisiae* (25), the DNA damage response protein 53BP1 from *Homo sapiens* (28), the histone binding protein RBBP1 from *H. sapiens* (21), and RNA binding protein ProQ from *E. coli* (26). The Tudor domain of PSHCP shares structural similarity with the core region of these four nucleic acid binding Tudor domains, but there is a surprising divergence in the mechanism of nucleic acid binding between all of these different Tudor domains, as detailed below.

Both 53BP1 and RBBP1 contain two adjacent Tudor domains (21,28). 53BP1 contains two sequential Tudor domains that are separated

in primary sequence and do not share any secondary structure. In contrast, RBBP1 forms an interdigitated Tudor domain where some of the secondary structure of the two Tudor domains alternate in the primary sequence. In addition, the two Tudor domains of RBBP1 share 2 β-strands generating a larger rigid scaffold for DNA binding. DNA binding to either 53BP1 and RBBP1 relies on residues in both Tudor domains (28). RBBP1 contains a positively charged pocket in between the two Tudor domains that allows for DNA binding (21). This is a distinct mode of nucleic acid binding formed by the overall interdigitated structure of the two Tudor domains. PSHCP aligns well to either individual Tudor domain from these structures but does not contain an extended surface like that of 53BP1 or RBBP1 for nucleic acid binding.

Esa1 contains a knotted Tudor domain due to the presence of two additional β-strands (β0 and β6) which lie just before and just after the classical Tudor domain (25). PSHCP aligns well to the short version of the Esa1 Tudor domain, which lacks these additional β-strands. The short version of Esa1 is incapable of binding to RNA while the knotted Tudor domains binds to poly(U) RNA with a 21.6 μM K_d (25). This binding interaction has the highest affinity of any of the interactions tested. It is surprising then that PSHCP is able to bind tRNA with an approximately 1 μM affinity as measured by ITC, even though it is most similar in structure to the short version of Esa1 which is incapable of binding RNA.

The *E. coli* ProQ protein contains two domains, an N-terminal FinO like domain and a C-terminal Tudor domain (26). Both of the domains interact with small RNAs to facilitate binding with low nanomolar affinities (29). The primary interaction region on ProQ for RNA appears to be through the FinO domain (26). However, two regions on the ProQ Tudor domain were also shown to interact with RNA. One of these regions which spans β-strands 2 and 3 overlaps with the surface of PSHCP that coordinates tRNA binding. While these Tudor domains share a common RNA binding surface, the location of the positively charged amino acids in these regions are in different β-strands. In PSHCP, K30 is located on β-strand 2 while in ProQ R214 is located on β-strand 3. The binding affinity of the ProQ Tudor domain for RNA has not been experimentally determined and therefore it is unclear if the ProQ Tudor domain has a similar affinity for RNA as PSHCP.

All of the discussed 'nucleic acid binding' Tudor domains exist as part of multidomain proteins that range in size from 232 to 1257 amino acids. The interaction of these proteins with nucleic acids often requires more than one domain for efficient binding. Since PSHCP is only 62 amino acids, it is rather remarkable that it has such a strong binding affinity for tRNA in such a small number of amino acids. It will be interesting to determine exactly what role the PSHCP-tRNA interaction plays in a subgroup of *Synechococcales* and why the protein is ~100% conserved at an amino acid level when only one face of the protein appears to be required for tRNA binding. For example, a point mutation at R23 had no effect on tRNA binding, yet the residue is 100% conserved across all PSHCP-containing taxa (Supplementary Datasets 1 and 2). Perhaps, PSHCP is also involved in binding to other targets. In pull-down assays, PSHCP was observed to be associated with the ribosomal protein L2 (10). Taken together with tRNA binding ability reported here, it is tempting to speculate that PSHCP is part of the ribosome complex and is involved in protein translation. Such a function could explain the extraordinary amino acid conservation of PSHCP, as translational machinery proteins are among the most conserved. Further studies are needed to evaluate this hypothesis.

EXPERIMENTAL PROCEDURES

Protein Expression and Purification

$PSHCP_{1-56}$ and $PSHCP_{1-62}$ were cloned into the ligation independent cloning vector (1B) which was a gift from Scott Gradia (Addgene, 29653). Plasmids containing PSHCP were transformed into BL21 (DE3) star cells (Invitrogen, C601003). PSHCP mutants were generated using Q5 Site-Directed Mutagenesis Kit (NEB, E0554S). Cells were grown in Terrific Broth (Fisher, BP9728-2) to an optical density at 600 nm of approximately 3.0. For unlabeled protein, expression was induced with 1 mM isopropylthio-β-D-galactoside (IPTG; Amresco, 97061) and cultures were grown for 18 hours at 18ºC. For uniformly labelled ^{15}N or ^{15}N and ^{13}C

protein, cells were pelleted (3,000 x g, 20 min) and resuspended in M9 minimal media comprising 3 g/L ^{15}N ammonium chloride (Cambridge Isotope Laboratories, NLM-467-25) with 10 g/L glucose or 10 g/L ^{13}C glucose (Cambridge Isotope Laboratories, CLM-1396-25). Expression was induced using 1 mM isopropylthio-β-D-galactoside (IPTG; Amresco, 97061) and cultures were grown for 18 hours at 18ºC. Cells were harvested and stored at -80ºC. Cell pellets were thawed and resuspended in 50 mM Tris, pH 8.0, 500 mM NaCl, 5 mM $MgCl_2$, 0.1% (v:v) Triton X-100 buffer containing protease inhibitors (Roche, 11836170001). Cells were lysed using a French Press (Thermo Electron, FA-032) and lysates were cleared by centrifugation for 50 min at 40,000 xg at 4ºC. Cleared lysates were incubated with TALON resin (Clontech, 635504) which was preequilibrated with 50 mM Tris, pH 8.0, 500 mM NaCl. Protein was eluted in buffer containing 200 mM imidazole. TEV protease was used to cleave off the hexahistidine tag. Cleaved protein was flowed over a desalting column to remove imidazole and a second purification using TALON resin was used to remove the TEV protease and any uncleaved protein. A final purification step was carried out using a HiLoad Superdex 75 PG column (GE Healthcare, 28-9893-33) equilibrated in 20 mM sodium phosphate pH 6.5 and 50mM NaCl for nucleotide binding experiments or 20 mM sodium phosphate pH 6.5, 200 mM NaCl, and 0.2 mM tris(2-carboxyethyl)phosphine (Amresco, K831-26) for structural studies.

NMR Spectroscopy and Structure Determination

All NMR experiments for assignment and structure determination were performed at 298 K on a Bruker Avance 700. The sequence-specific backbone assignment was determined using 2D [^{1}H ^{15}N] HSQC, 3D HNCA, 3D HN(CO)CA, 3D HNCO, 3D HN(CA)CO, 3D CBCA(CO)NH, 3D HNCACB and 3D HBHA(CBCACO)NH experiments. Aliphatic side chain assignments were determined using 3D (H)CCH-TOCSY, 3D HC(C)H-COSY, 3D HC(C)H-TOCSY, 3D H(CCCO)NH and 3D (H)CC(CO)NH experiments. Amide side chain assignments were determined using a 3D ^{15}N-resolved NOESY. ^{1}H chemical shifts were externally referenced to 0 ppm methyl resonance of 2,2-dimethyl-2-silapentane-5-sulfonate (DSS), whereas ^{13}C and ^{15}N chemical shifts were indirectly referenced according to the IUPAC recommendations (30). All NMR spectra were processed using Topspin 3.5 (Bruker). Processed spectra were analyzed using CARA (http://cara.nmr.ch/). Backbone and side chain assignments were deposited in the BMRB under accession 30559.

For structure calculation a 2D ^{1}H-^{1}H NOESY, 3D ^{15}N-resolved NOESY and 3D ^{13}C-resolved NOESY were recorded and processed using Topspin 3.5 (Bruker). UNIO was used for iterative automated NOE peak picking and NOE assignment by ATNOS/CANDID (31,32) and structure calculation with CYANA v2.1 (33,34). The 20 lowest energy structures were used for water refinement in CNS v1.3 with the RECOORD scripts (35). The quality of the final ensemble was verified using NMR-Procheck and the RMSD of these structures was determined using MolMol (36,37). The final structure was deposited in the PDB under ID 6NNB. All structure figures were generated using the PyMOL Molecular Graphics System, Version 1.8 (Schrödinger, LLC). APBS was used to generate the electrostatic surface representation of PSHCP (38).

HSQC Titration Experiments

dsGC DNA (5′-CCG CGC GCG CGG-3′) was synthesized by IDT. DNA was resuspended in 20mM sodium phosphate pH 6.5, 50 mM NaCl, heated at 100ºC for 5 min and cooled at room temperature to allow for the DNA to anneal. tRNA from *E. coli* (Sigma R1753) was resuspended in 20 mM sodium phosphate pH 6.5, 50mM NaCl. The concentration of the *E. coli* tRNA was determined by running a range of concentrations of the tRNA on an agarose gel and comparing them to a standard concentration. ^{15}N labeled PSHCP was mixed with nucleic acid samples to a final protein concentration of 50 µM. The DNA and tRNA concentrations were varied based on the overall titration. ^{1}H-^{15}N HSQC spectra were recorded on a Bruker Avance 600 MHz spectrometer using a 1.7 mm cryoprobe. CSP values were determined using $\delta\Delta = \sqrt{((\delta^{1}H)^{2} + 0.14(\delta^{15}N)^{2})}$. CSP data were fit using

MATLAB (MathWorks) to $f(x) = CSP_{max}*((K_d+x+P_{tot})-sqrt((K_d+x+P_{tot})^2-4*(x*P_{tot})))/(2*(P_{tot}))$ where P_{tot} is the total concentration of protein (50 µM). To determine the intensity ratio of peaks from the 1H-15N HSQC spectra, peak intensity values were determined using CARA (http://cara.nmr.ch/).

Isothermal titration calorimetry

ITC was performed using a VP-ITC (Microcal). All proteins and tRNA were prepared in 20mM sodium phosphate pH 6.5, 50 mM NaCl for ITC experiments. Approximately 100 µM *E. coli* tRNA (Sigma R1753) was titrated into 10 µM PSHCP (either $PSHCP_{1-62}$ or $PSHCP_{1-56}$) at 24°C. tRNA concentration was determined by running a tRNA titration on a 1% agarose gel. tRNA samples were then compared to samples of known concentration. Data were analyzed using Origin (Microcal) to produce both ITC thermograms and binding isotherms.

tRNA Docking

tRNA docking onto PSHCP was performed using the HDOCK program (27). PSHCP was used as the receptor molecule which is held rigid in the docking. For the ligand molecule to dock onto PSHCP we used *E. coli* fMet initiator tRNA molecule (PDB: 5L4O), since this tRNA structure does not contain any ligands. Computational docking was performed without using any template structures for the complex allowing for the docking program to search the entire surface of PSHCP for potential areas to dock the tRNA.

Survey of PSHCP Taxonomic Distribution and Amino Acid Conservation

The amino acid sequence of the PSHCP gene from *Prochlorococcus marinus* str. MIT 9312 (RefSeq ID WP_002807701.1) was used as a query in a BLASTP search of the *nr* database via NCBI web site (accessed on September 12, 2018; BLAST v. 2.8.0+ (39); E-value cutoff 10^{-4}). Taxonomic distribution was examined via Taxonomy Report provided with the BLAST search results. Matches with sequences not identical to the query were retrieved and aligned in ClustalX v. 2.1 (40). Additionally, the same query sequence was used in a BLASTP search of the 703 genomes of *Prochlorococcus* and marine *Synechococcus* available via IMG/ProPortal (41). This database includes both isolates and single cell genomes, the latter ones with an average completion of 60% (41). The BLASTP 2.6.0+ search was carried out via IMG/MER web site at https://img.jgi.doe.gov/ on September 11, 2018, with E-value cutoff of 10^{-5}. The obtained 463 PSHCP homologs were retrieved and aligned in ClustalX v. 2.1 (40). Two poor quality sequences (defined as those containing at least one X) and one partial sequence at the beginning of a short contig in an incomplete genome were removed from the alignments. Both alignments are available as Supplemental Datasets 1 and 2.

Acknowledgments
This work was supported by Dartmouth start-up funds and the National Institutes of Health awards GM113132 and GM128663 to M.J.R, and by the Simons Foundation Investigator in Mathematical Modeling of Living Systems award 327936 to O.Z. We would like to thank Amanda Cockshutt and Kaitlyn Connelly for helpful discussions.

Conflict of Interest
The authors declare they have no conflicts of interests with the contents of this article.

Author Contributions
K.M.B. and R.D. purified all protein, analyzed NMR data, and collected ITC data. M.P. collected all of the NMR spectra. O.Z. carried out the survey of PSHCP taxonomic distribution. O.Z. and M.J.R. conceptualized the project. K.M.B., M.P., O.Z. and M.J.R. wrote and edited the manuscript.

References

1. Chisholm, S. W., Frankel, S. L., Goericke, R., Olson, R. J., Palenik, B., Waterbury, J. B., Westjohnsrud, L., and Zettler, E. R. (1992) Prochlorococcus-Marinus Nov Gen-Nov Sp - an Oxyphototrophic Marine Prokaryote Containing Divinyl Chlorophyll-a and Chlorophyll-B. *Arch Microbiol* **157**, 297-300
2. Waterbury, J. B., Watson, S. W., Valois, F. W., and Franks, D. G. (1986) Biological and ecological characterization of the marine unicellular cyanobacterium Synechococcus. in *Photosynthetic Picoplankton* (Platt, T., and Li, W. K. W. eds.). pp 71–120
3. Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincon, J., Zabala, L. L., Jiao, N., Karl, D. M., Li, W. K., Lomas, M. W., Veneziano, D., Vera, C. S., Vrugt, J. A., and Martiny, A. C. (2013) Present and future global distributions of the marine Cyanobacteria Prochlorococcus and Synechococcus. *Proc Natl Acad Sci U S A* **110**, 9824-9829
4. Scanlan, D. J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W. R., Post, A. F., Hagemann, M., Paulsen, I., and Partensky, F. (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**, 249-299
5. Chisholm, S. W. (2017) Prochlorococcus. *Curr Biol* **27**, R447-R448
6. Biller, S. J., Berube, P. M., Lindell, D., and Chisholm, S. W. (2015) Prochlorococcus: the structure and function of collective diversity. *Nat Rev Microbiol* **13**, 13-27
7. Konstantinidis, K. T., and Tiedje, J. M. (2005) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* **187**, 6258-6264
8. Zhaxybayeva, O., Doolittle, W. F., Papke, R. T., and Gogarten, J. P. (2009) Intertwined evolutionary histories of marine Synechococcus and Prochlorococcus marinus. *Genome Biol Evol* **1**, 325-339
9. Zhaxybayeva, O., Gogarten, J. P., and Doolittle, W. F. (2007) A hyperconserved protein in Prochlorococcus and marine Synechococcus. *FEMS Microbiol Lett* **274**, 30-34
10. Whidden, C. E., DeZeeuw, K. G., Zorz, J. K., Joy, A. P., Barnett, D. A., Johnson, M. S., Zhaxybayeva, O., and Cockshutt, A. M. (2014) Quantitative and functional characterization of the hyper-conserved protein of Prochlorococcus and marine Synechococcus. *PLoS One* **9**, e109327
11. Usher, K. M., Toze, S., Fromont, J., Kuo, J., and Sutton, D. C. (2004) A new species of cyanobacterial symbiont from the marine sponge Chondrilla nucula. *Symbiosis* **36**, 183-192
12. Nowack, E. C., Melkonian, M., and Glockner, G. (2008) Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol* **18**, 410-418
13. Dufresne, A., Garczarek, L., and Partensky, F. (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* **6**, R14
14. Komarek, J., Kastovsky, J., Mares, J., and Johansen, J. R. (2014) Taxonomic classification of cyanoprokaryotes (cyanobacterial genera) 2014, using a polyphasic approach. *Preslia* **86**, 295-335
15. Meszaros, B., Erdos, G., and Dosztanyi, Z. (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* **46**, W329-W337
16. Shen, Y., and Bax, A. (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR* **56**, 227-241
17. Holm, L., and Rosenstrom, P. (2010) Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**, W545-549
18. Lu, R., and Wang, G. G. (2013) Tudor: a versatile family of histone methylation 'readers'. *Trends Biochem Sci* **38**, 546-555
19. Chen, C., Nott, T. J., Jin, J., and Pawson, T. (2011) Deciphering arginine methylation: Tudor tells the tale. *Nat Rev Mol Cell Biol* **12**, 629-642
20. Musselman, C. A., Avvakumov, N., Watanabe, R., Abraham, C. G., Lalonde, M. E., Hong, Z., Allen, C., Roy, S., Nunez, J. K., Nickoloff, J., Kulesza, C. A., Yasui, A., Cote, J., and

Kutateladze, T. G. (2012) Molecular basis for H3K36me3 recognition by the Tudor domain of PHF1. *Nat Struct Mol Biol* **19**, 1266-1272
21. Gong, W. B., Wang, J. F., Perrett, S., and Feng, Y. G. (2014) Retinoblastoma-binding Protein 1 Has an Interdigitated Double Tudor Domain with DNA Binding Activity*. *J Biol Chem* **289**, 4882-4895
22. Lancelot, N., Charier, G., Couprie, J., Duband-Goulet, I., Alpha-Bazin, B., Quemeneur, E., Ma, E., Marsolier-Kergoat, M. C., Ropars, V., Charbonnier, J. B., Miron, S., Craescu, C. T., Callebaut, I., Gilquin, B., and Zinn-Justin, S. (2007) The checkpoint Saccharomyces cerevisiae Rad9 protein contains a tandem tudor domain that recognizes DNA. *Nucleic Acids Res* **35**, 5898-5912
23. Ponting, C. P. (1997) Tudor domains in proteins that interact with RNA. *Trends in Biochemical Sciences* **22**, 51-52
24. Pek, J. W., Anand, A., and Kai, T. (2012) Tudor domain proteins in development. *Development* **139**, 2255-2266
25. Shimojo, H., Sano, N., Moriwaki, Y., Okuda, M., Horikoshi, M., and Nishimura, Y. (2008) Novel structural and functional mode of a knot essential for RNA binding activity of the Esa1 presumed chromodomain. *J Mol Biol* **378**, 987-1001
26. Gonzalez, G. M., Hardwick, S. W., Maslen, S. L., Skehel, J. M., Holmqvist, E., Vogel, J., Bateman, A., Luisi, B. F., and Broadhurst, R. W. (2017) Structure of the Escherichia coli ProQ RNA-binding protein. *RNA* **23**, 696-711
27. Yan, Y., Zhang, D., Zhou, P., Li, B., and Huang, S. Y. (2017) HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res* **45**, W365-W373
28. Charier, G., Couprie, J., Alpha-Bazin, B., Meyer, V., Quemeneur, E., Guerois, R., Callebaut, I., Gilquin, B., and Zinn-Justin, S. (2004) The Tudor tandem of 53BP1: a new structural motif involved in DNA and RG-rich peptide binding. *Structure* **12**, 1551-1562
29. Smirnov, A., Wang, C., Drewry, L. L., and Vogel, J. (2017) Molecular mechanism of mRNA repression in trans by a ProQ-dependent small RNA. *EMBO J* **36**, 1029-1045
30. Markley, J. L., Bax, A., Arata, Y., Hilbers, C. W., Kaptein, R., Sykes, B. D., Wright, P. E., and Wuthrich, K. (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union Task Group on the Standardization of Data Bases of Protein and Nucleic Acid Structures Determined by NMR Spectroscopy. *J Biomol NMR* **12**, 1-23
31. Herrmann, T., Guntert, P., and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE-identification in the NOESY spectra using the new software ATNOS. *J Biomol NMR* **24**, 171-189
32. Herrmann, T., Guntert, P., and Wuthrich, K. (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *Journal of molecular biology* **319**, 209-227
33. Guntert, P. (2004) Automated NMR structure calculation with CYANA. *Methods in molecular biology* **278**, 353-378
34. Guntert, P., Mumenthaler, C., and Wuthrich, K. (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. *Journal of molecular biology* **273**, 283-298
35. Nederveen, A. J., Doreleijers, J. F., Vranken, W., Miller, Z., Spronk, C. A., Nabuurs, S. B., Guntert, P., Livny, M., Markley, J. L., Nilges, M., Ulrich, E. L., Kaptein, R., and Bonvin, A. M. (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* **59**, 662-672
36. Laskowski, R. A., Rullmannn, J. A., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* **8**, 477-486

37. Koradi, R., Billeter, M., and Wuthrich, K. (1996) MOLMOL: a program for display and analysis of macromolecular structures. *J Mol Graph* **14**, 51-55, 29-32
38. Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T., Geng, W., Krasny, R., Wei, G. W., Holst, M. J., McCammon, J. A., and Baker, N. A. (2018) Improvements to the APBS biomolecular solvation software suite. *Protein Sci* **27**, 112-128
39. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402
40. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948
41. Berube, P. M., Biller, S. J., Hackl, T., Hogle, S. L., Satinsky, B. M., Becker, J. W., Braakman, R., Collins, S. B., Kelly, L., Berta-Thompson, J., Coe, A., Bergauer, K., Bouman, H. A., Browning, T. J., De Corte, D., Hassler, C., Hulata, Y., Jacquot, J. E., Maas, E. W., Reinthaler, T., Sintes, E., Yokokawa, T., Lindell, D., Stepanauskas, R., and Chisholm, S. W. (2018) Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. *Sci Data* **5**, 180154
42. Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., and Stuckey, P. J. (2011) Automatic generation of protein structure cartoons with Pro-origami. *Bioinformatics* **27**, 3315-3316

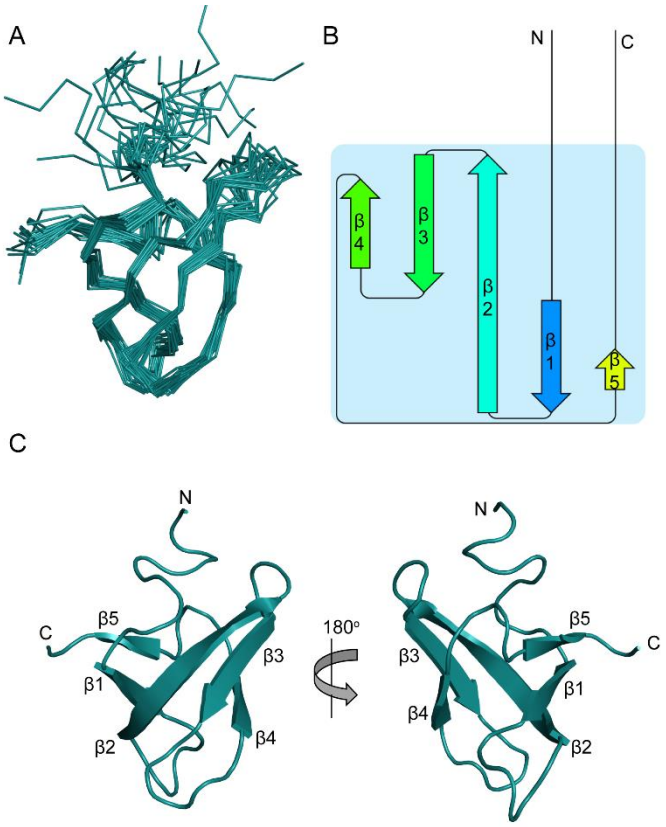

Figure 1. The structure of PSHCP. A) Ribbon diagram of $PSHCP_{1-56}$ showing the overlay of the 20 lowest energy structures. B) Topology diagram of PSHCP generated using the pro-origami webserver (42). C) Cartoon representation of PSHCP shown in two different orientations. Each β-strand in the structure is labeled and numbered. The N and C termini are also labeled.

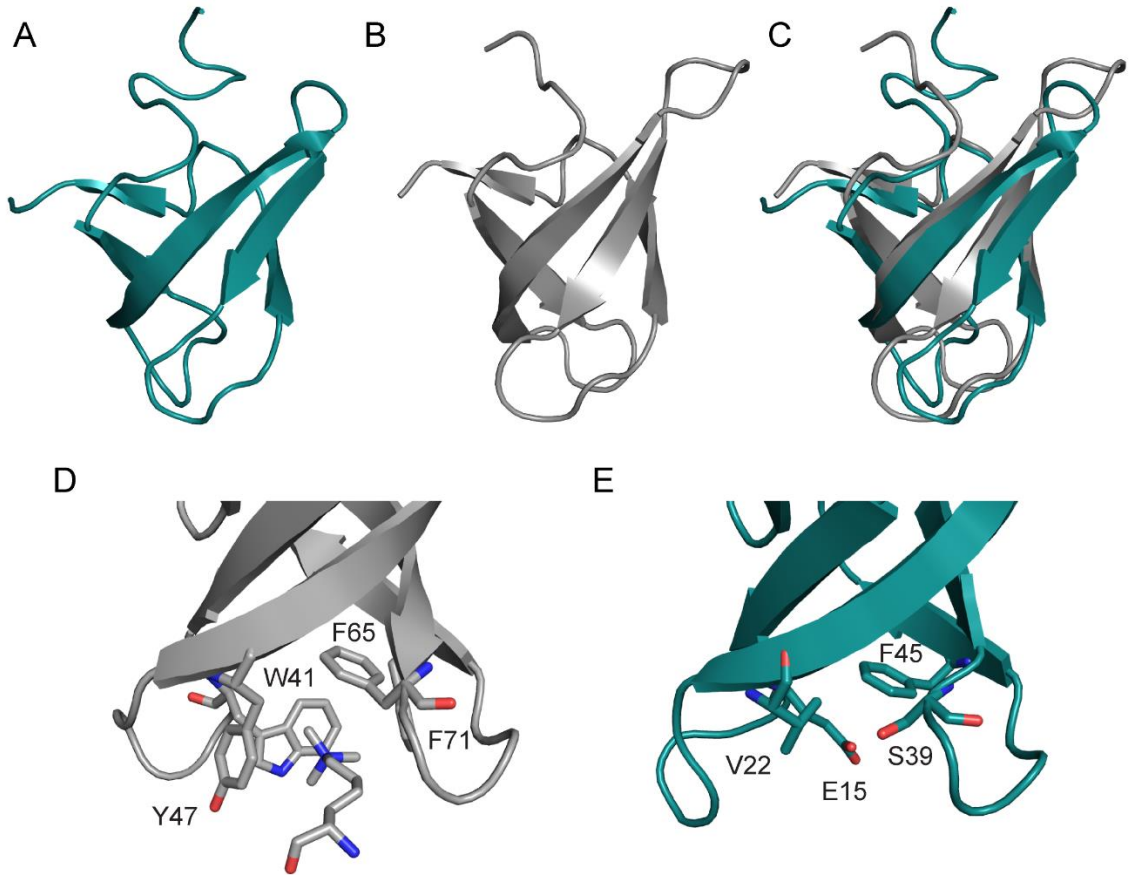

Figure 2. PSHCP contains a Tudor domain. A) Cartoon representation of PSHCP. B) Cartoon representation of the Tudor domain from human Tudor domain-containing protein 3 (PDBID: 3PMT). C) Overlay of the structures from panels A and B. D) A close-up view of the aromatic cage from the Tudor domain of PHF1 (PDBID: 4HCZ) bound to trimethylated lysine. E) A close-up view showing the residues in PSHCP that are located in the position of the aromatic cage residues in typical Tudor domains. In panels D and E all residues that make up the aromatic cage are shown as stick representations and labeled.

A

1 MELDLQPGDVVKVLESAALGWVRARVIRVKS 31

32 GGRVVVQSDQGREFTARGNQVRLIEPAGFRP 62

B

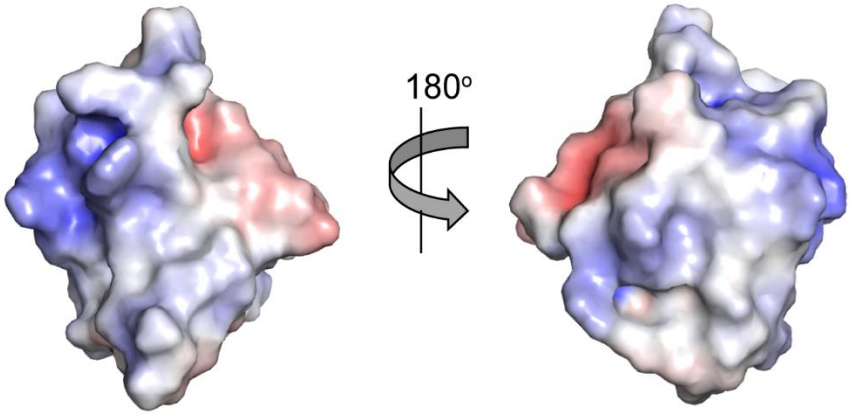

Figure 3. Electrostatics of PSHCP. A) The amino acid sequence of the PSHCP from *P. marinus* strain CCMP1375. All positively charge residues are highlighted in yellow. B) Electrostatic surface representation of PSHCP shown in two orientations. The electrostatic surface was generated using APBS (38), with blue and red representing positively charged and negatively charged surfaces, respectively.

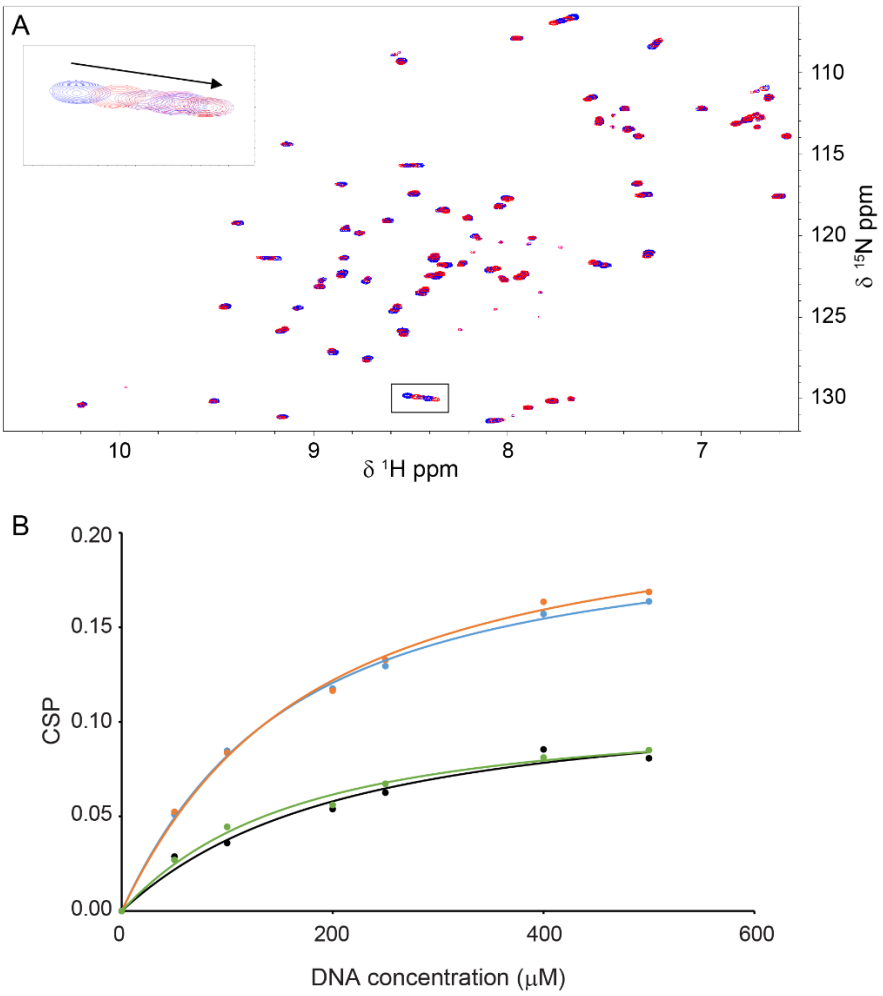

Figure 4. PSHCP weakly binds to dsGC DNA. A) ^{1}H-^{15}N HSQC spectra of 50 μM PSHCP in the presence of increasing concentrations of dsGC DNA. The ^{1}H-^{15}N HSQC spectrum of PSHCP in the absence of dsGC DNA is shown in blue. The inset shows the chemical shift perturbation for E15 with the arrow pointing in the direction of the shift with increasing DNA concentration. C) CSP data for E15 (cyan), G20 (orange), V22 (black) and T46 (green) as a function of DNA concentration. All data points are shown as circles. The fits to the data are shown as lines of the same color as the datapoints.

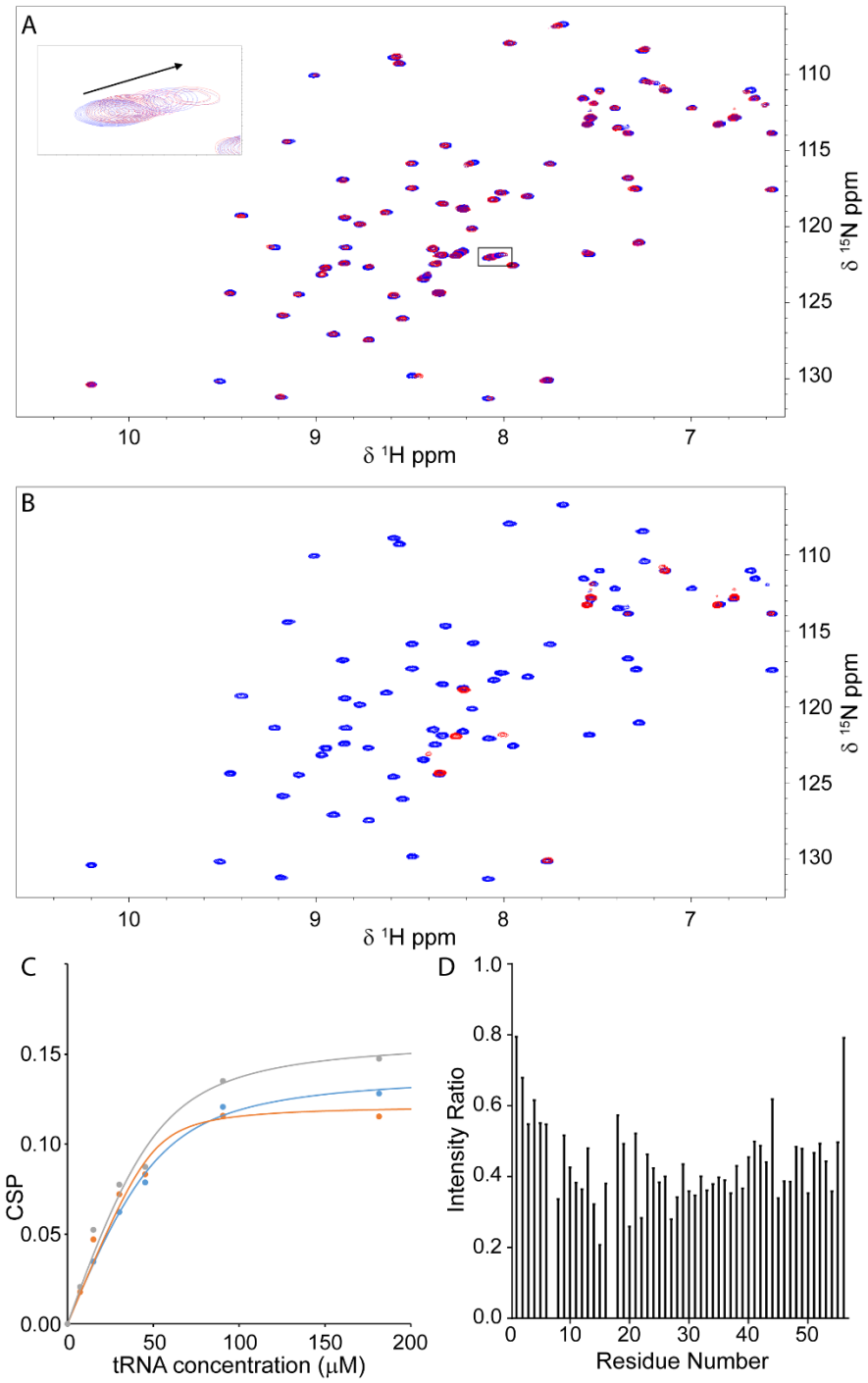

Figure 5. PSHCP binding to *E. coli* tRNA. A) ^{1}H-^{15}N HSQC spectra of 50 μM PSHCP with increasing concentrations of *E. coli* tRNA. The inset shows the chemical shift perturbation for L3 with the arrow showing the direction of the shift with increasing tRNA concentration. B) ^{1}H-^{15}N HSQC spectra of 50 μM PSHCP in the absence of *E. coli* tRNA (blue) and in the presence of 90.7 μM *E. coli* tRNA (red). C) CSP data for E2 (cyan), L3 (orange), and D4 (grey) as a function of tRNA concentration. D) Bound/unbound intensity ratios for individual PSHCP residues in the presence of 7.4 μM tRNA.

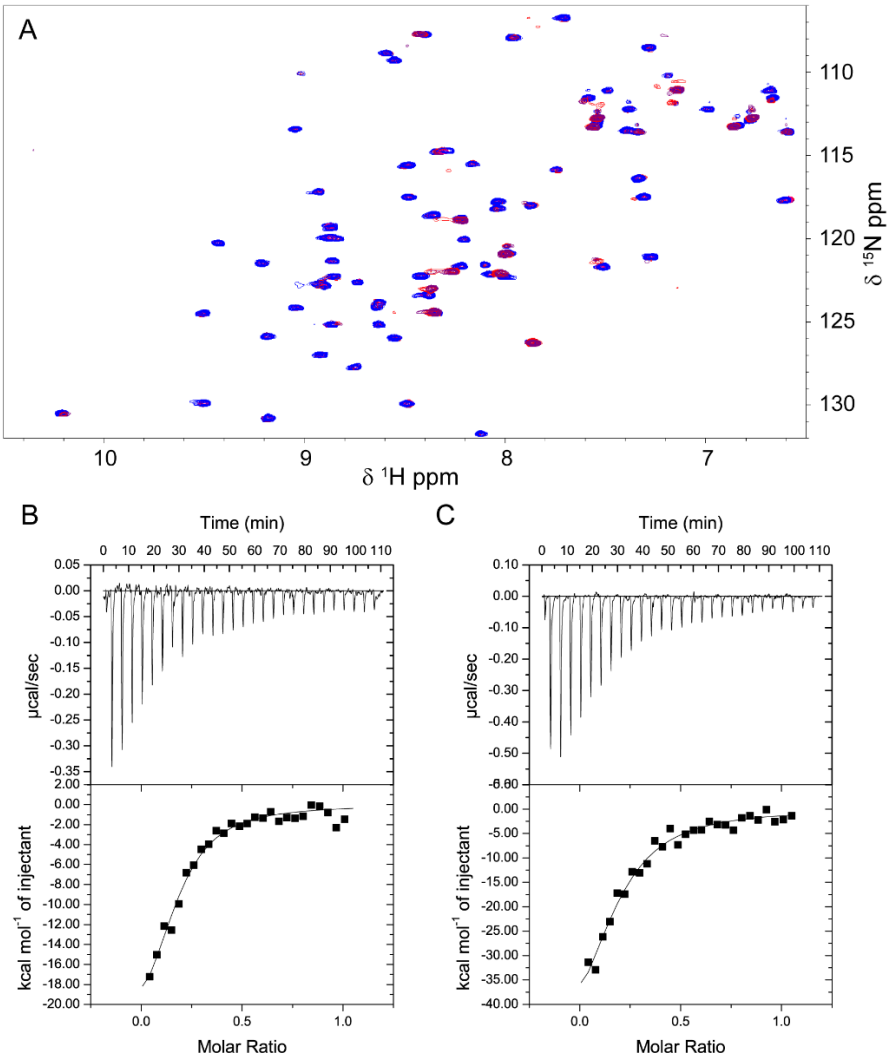

Figure 6. Residues 57-62 do not alter tRNA binding. A) ^{1}H-^{15}N HSQC spectra of 50 µM PSHCP$_{1-62}$ with increasing concentrations of *E. coli* tRNA. B) ITC thermogram (top) and binding isotherm (bottom) for tRNA titrated into pSHCP$_{1-56}$. C) ITC thermogram (top) and binding isotherm (bottom) for tRNA titrated into pSHCP$_{1-62}$.

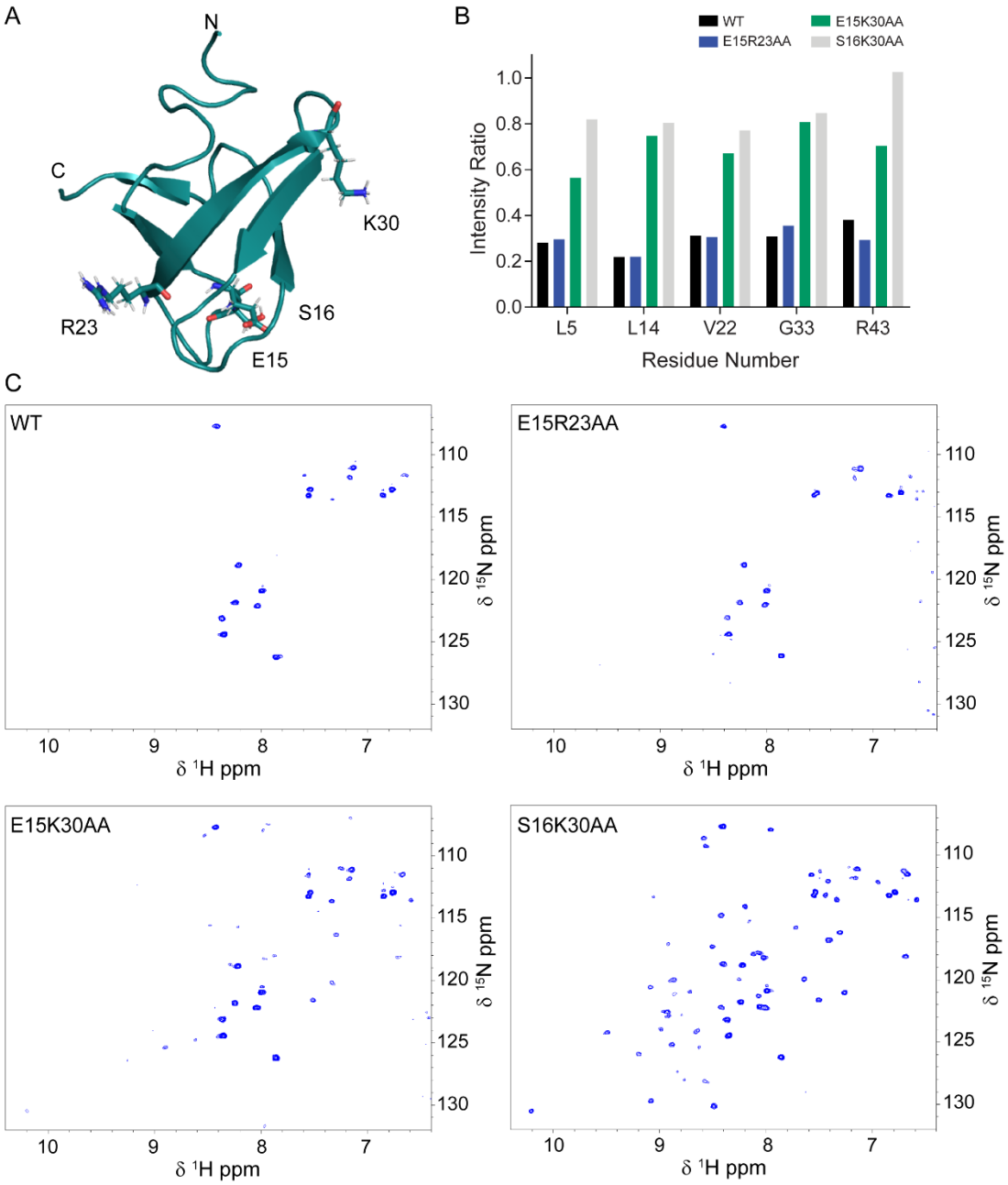

Figure 7. K30 is essential for tRNA binding. A) Cartoon representation of PSHCP showing the four different residues which were mutated as sticks. Each residue is labeled. B) The intensity ratio between $PSHCP_{1-62}$ in the presence of 7.4 µM tRNA and unbound $PSHCP_{1-62}$ was measured for L5, L14, V22, G33 and R43. WT PSHCP, E15R23AA, E15K30AA and S16K30AA are shown in black, blue, green and grey respectively. C) ^{1}H-^{15}N HSQC spectra of 50 µM PSHCP in the presence of 90.7 µM *E. coli* tRNA. The spectra for the WT and three mutant proteins are shown in separate panels. Each spectrum is adjusted to an identical contour level.

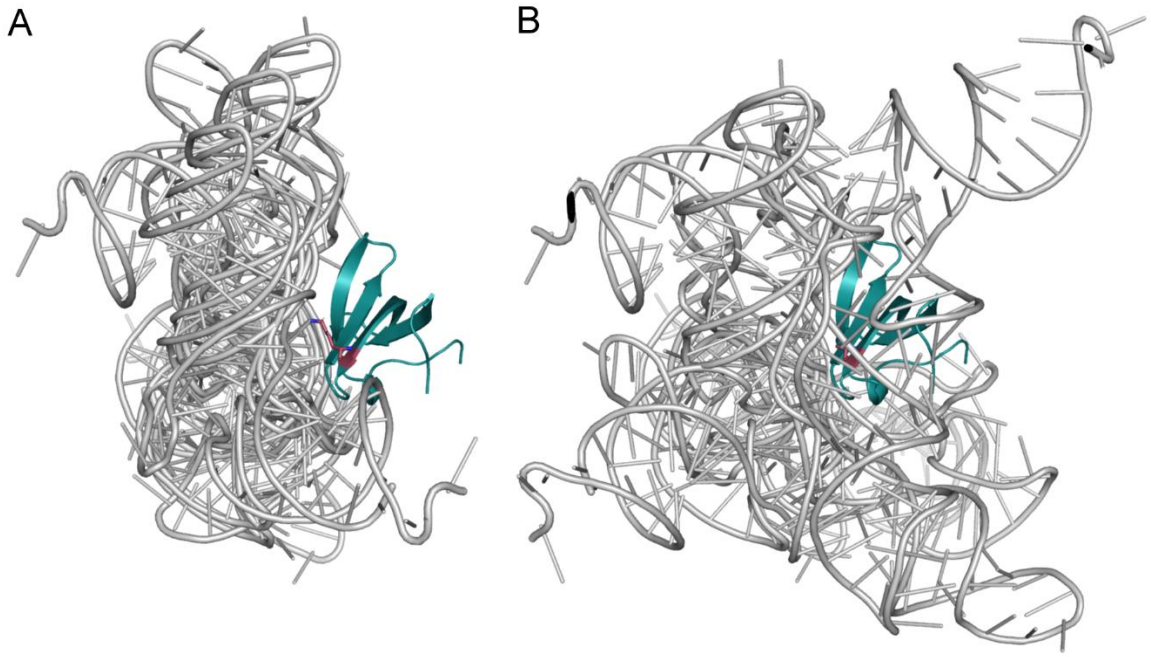

Figure 8. Docking reveals a potential tRNA binding site on PSHCP. A) The top five models of *E. coli* fMet initiator tRNA (PDB: 5L4O; in grey) docked onto PSHCP WT (blue). K30 is shown as a stick representation in red. B) The top five models of *E. coli* fMet initiator tRNA (grey) docked onto PSHCP K30A mutant (blue). K30A is shown as a stick representation in red. Both panels A and B are shown at the same orientation of PSHCP.

Table 1. Structural statistics of $PSHCP_{1-56}$.

	$PSHCP_{1-56}$
NMR distance and dihedral constraints	
Distance constraints	
Total NOE	1195
Intraresidue	282
Interresidue	913
Sequential (|i-j| = 1)	311
Medium range (1<|i-j|<=4)	148
Long range (|i-j|>= 5)	454
Dihedral angle constraints	
φ	22
ψ	22
Structure statistics	
Violations (mean ± SD)	
Distance constraints (Å)	0.036 ± 0.004
Dihedral angle constraints (°)	0.028 ± 0.069
Deviations from idealized geometry (mean ± SD)	
Bond lengths (Å)	0.015 ± 0.001
Bond angles (°)	1.724 ± 0.054
Impropers (°)	1.731 ± 0.158
Ramachandran statistics	
Most favored	82.2%
Additional allowed	16.3%
Generously allowed	1.1%
Disallowed	0.4%
Average pairwise root mean square deviation (Å)	
Heavy (4-29, 35-56)	1.32 ± 0.190
Backbone (4-29, 35-56)	0.72 ± 0.150

Table 2. DNA Binding Data

Residue	**Calculated K_d (μM)**	**R^2**
E15	121.3	0.99837
G20	138.4	0.99680
V22	173.3	0.97596
T46	129.2	0.99165

Table 3. tRNA Binding Data

Residue	**Calculated K_d (μM)**	**R^2**
E2	10.4	0.99510
L3	2.9	0.97705
D4	9.4	0.98226