Chapter V

TOWARD EXPANDING NON-CANONICAL PROLINE INCORPORATION BY PROLYL-TRNA SYNTHETASE ENGINEERING

5.1 Contributions

Bradley R. Silverman assisted with flow cytometry, FACS, and microscopy; and provided helpful discussions. Katharine Y. Fang initiated the split-GFP approach and performed the cloning of those initial constructs and the first error-prone *proS* library. Alejandro Lopez contributed to cloning M157Q and C443G *proS* mutants. Stephanie L. Breunig prepared all other constructs, performed all experiments, and analyzed all data.

5.2 Abstract

The promiscuity of the translational machinery of *E. coli* has enabled the incorporation of chemically diverse proline analogs into recombinant proteins, and point mutants of the prolyl-tRNA synthetase (ProRS) can increase the residue-specific incorporation for several non-canonical proline residues. However, robust high-throughput methods are needed to identify ProRS variants capable of accepting more chemically diverse proline analogs. Here, we discuss several attempts to develop a screening platform for the directed evolution of the *E. coli* ProRS. First, we describe our efforts involving full-length and split fluorescent protein reporters. Next, we touch on alternative approaches that seek to detect inclusion bodies, or proline analog toxicity. Finally, we detail an approach that relies upon specific labeling of a short tetracysteine motif by the small molecule fluorophore FlAsH.

These efforts highlight the difficulty in engineering the *E. coli* ProRS for improved proline analog incorporation, and lay the groundwork for future ProRS engineering endeavors.

5.3 Introduction

The promiscuity within the translational machinery of *E. coli* has enabled the incorporation of chemically diverse non-canonical amino acids (ncAAs) into proteins in living cells; protein engineering efforts have further expanded ncAA incorporation technologies. In most cases, the gatekeepers of translational fidelity are the aminoacyl-tRNA synthetases (aaRSs),[1] so increased ncAA incorporation is often achieved by aaRS engineering. Amber suppression and related site-specific incorporation approaches require the development of an orthogonal aaRS/tRNA pair.[2] While not required in all cases, residue-specific incorporation has similarly been assisted by aaRS engineering. For instance, separately engineered MetRS variants have enabled the incorporation of the methionine analogs trifluoronorleucine (Tfn)[3] and azidonorleucine (Anl).[4]

To date, a residue-specific incorporation approach is the only method to incorporate proline analogs into recombinant proteins. Many non-canonical proline (ncPro) residues are readily accepted by the endogenous *E. coli* translational machinery,[5] or simply require overexpression of the wild-type prolyl-tRNA synthetase (ProRS), and high NaCl concentrations to promote proline uptake.[6] Point mutations in the ProRS active site have also been reported to improve ncPro incorporation.[7] However, the chemical space about proline is vast, and many ncPro residues of interest cannot be incorporated efficiently into recombinant proteins (for example, see Figure 2.S1 and Table 2.S1).

Directed evolution is a powerful technology to engineer enzymes for applications beyond their endogenous activities;[8] we note numerous examples of aaRS engineering for ncAA incorporation.[2] However, these successes have not involved all the aminoacyl-tRNA synthetases used by biology. Commonly evolved aaRS/tRNA pairs include the PylRS/tRNA_CUA pairs from *Methanosarcina barkeri* and *Methanosarcina mazei*, and the TyrRS/tRNA_CUA pairs from *Escherichia coli* and *Methanocaldococcus janaschii*.[2,9] The sole report in the literature that describes ProRS engineering resulted in orthogonal ProRS/tRNA^Pro pairs,[10] but no site-specific incorporation of proline analogs was reported. Later efforts to expand upon this work[11] point to the difficulties in site-specific ncPro incorporation.

We sought to engineer the *E. coli* ProRS to accept a diverse range of proline analogs, with the goal of incorporating these residues into recombinant proteins in a residue-specific[12] (rather than site-specific) manner. In this chapter, we describe several of these efforts. Most designs are fluorescence-based platforms that might be used to sort large ProRS libraries by fluorescence-activated cell sorting (FACS). We explored a split-GFP system and a FlAsH-based labeling protocol to measure ncPro incorporation. Major challenges encountered with the split-GFP approach include reporter protein solubility and limited dynamic range. FlAsH labeling provided good analytical measurements for ncPro incorporation, but we faced significant hurdles to effectively sort ProRS libraries by FACS. We also describe a few alternative approaches that were similarly unsuccessful. Although we have not yet identified ProRS mutants capable of improved ncPro incorporation by high throughput screening to date, this work informs future ProRS engineering efforts.

5.4 Results and discussion

5.4.1 Full-length fluorescent proteins are poor reporters of proline analog incorporation

In FACS-based screening methods, GFP (or another full-length fluorescent protein) is often used as a reporter for ncAA incorporation. The codon(s) encoding the ncAA of interest are placed at a permissive location in the fluorescent protein, and codon readthrough (presumably by successful translation with the ncAA) leads to complete protein synthesis and successful chromophore formation.

The unique properties of proline, and diversity in the chemical and conformational properties of proline analogs, complicate a fluorescent protein approach for ProRS engineering. Replacing proline with ncPro residues in a full-length protein often alters protein folding and stability, as discussed in Chapter I of this thesis. In fact, global ncPro replacement in fluorescent proteins can lead to unpredictable results: global replacement of proline with 4*S*-fluoroproline (4*S*-F) leads to a fluorescent GFP variant with faster folding kinetics than its proline-containing parent. However, introduction of its diastereomer 4*R*-fluoroproline (4*R*-F) results in an insoluble GFP.[13] Opposite stereospecific effects were obtained with mRFP1; in that case, 4*S*-F incorporation led to an insoluble protein, while 4*R*-F accelerated protein folding.[14]

A more recent report replaced the proline residues in three fluorescent proteins (EGFP, NowGFP, and KillerOrange) with 4*R*-F, 4*S*-F, 4,4-difluoroproline (44-diF) and 3,4-dehydroproline (dhp).[15] In all cases, replacement with 4*S*-F or dhp was tolerated, while 4*R*-F and 44-diF did not lead to fluorescence. For examples in which ncPro and fluorescent protein pairs led to a mature fluorescent protein, the authors measured refolding kinetics

after chemical denaturation. In the context of EGFP, incorporation of 4*S*-F sped refolding, while dhp slowed it. Interestingly, all variants of NowGFP and KillerOrange did not refold once denatured.[15]
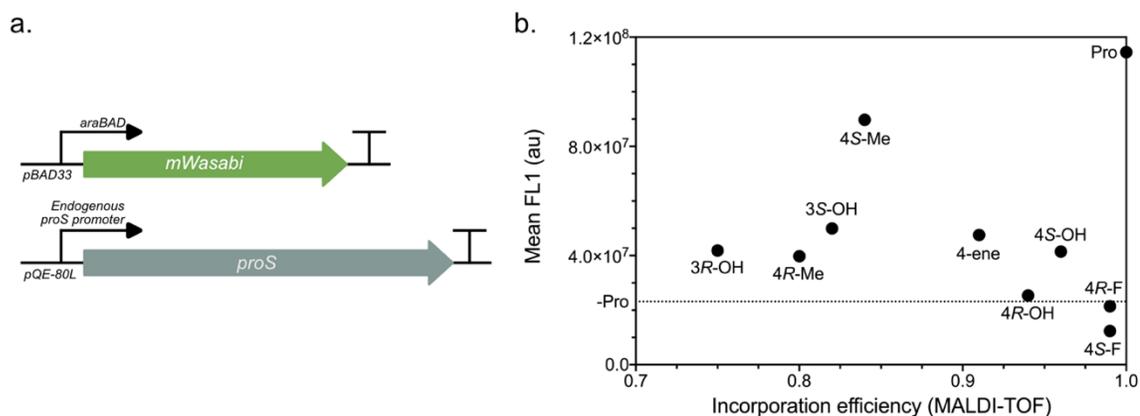

Figure 5.1. mWasabi fluorescence does not correspond to proline analog incorporation efficiencies. **a**. mWasabi expression was placed under control of the arabinose-inducible araBAD promoter, and *proS* was overexpressed from pQE-80L backbone. **b**. mWasabi fluorescence (FL1 by flow cytometry) as a function of incorporation efficiency (as determined by MALDI-TOF). mWasabi was expressed for 2.5 h at 37 °C under ncPro-incorporation conditions, and the resulting fluorescence was assessed by flow cytometry. Incorporation efficiency was determined by MALDI-TOF mass spectrometry after expression of ncPro-containing proinsulins.

Similarly unpredictable behavior was observed in our hands. We expressed mWasabi in *E. coli* under ncPro-incorporation conditions (Figure 5.1a), and assessed fluorescence by flow cytometry. We found that fluorescence is not a good predictor of ncPro incorporation efficiency, as determined by MALDI-TOF (Figure 5.1b). In fact, in some cases (such as 4*S*-F), the measured fluorescence is *less than that of a sample in which no proline has been added* (dotted line), despite the fact that these residues are known to have high incorporation efficiencies in *E. coli*.[5] We suggest that 4*S*-F out-competes residual proline for incorporation into mWasabi, but leads to an improperly folded protein. The result is a

sample that is less fluorescent than the "–Proline" negative control. We also observe that known conformational preferences of the proline analogs tested[16] do not correlate with fluorescence; for instance, 4S-Me and 4R-OH both prefer the *exo* ring pucker and *trans* amide isomer compared to other proline analogs assessed,[16,17] yet lead to divergent fluorescent behaviors when incorporated into mWasabi (Figure 5.1b).

This unpredictable fluorescence prohibits the use of full-length fluorescent proteins as a general screening output for ncPro incorporation. In the directed evolution of the MetRS for Tfn incorporation, methionine residues were first removed from GFP, then reintroduced at permissive locations.[3] We anticipated difficulties in applying a similar approach to obtaining a proline-free fluorescent protein due to the imino acid's unique properties. Further, Pro89 (which exists as the *cis* isomer) and Pro196 are among the most widely conserved residues in GFPs and GFP-like proteins,[18] so changes at those positions will not likely be tolerated. To the best of our knowledge, no reported fluorescent protein contains fewer than two proline residues, and proteins with few proline residues bind exogenous co-factors. We concluded that full-length fluorescent proteins are unable to act as a generalizable, robust approach to determining ncPro incorporation levels, and therefore pursued alternative designs.

5.4.2 Split-GFP as a reporter for proline analog incorporation

The presence of ncPro residues within any reporter protein will likely interfere with the function of that reporter. A split-GFP approach, in which one GFP fragment does not contain proline residues, might circumvent these issues.[19] In this case, a larger, proline-containing unit (here, GFP1-10) would be expressed in proline-containing medium. Proline codon(s) would be placed N-terminal to the proline-free fragment (GFP11), which would

be expressed in ncPro medium after a medium shift. As a result, expression of the second, proline-free fragment would depend upon readthrough of the proline codons, yet the fluorescent output should not be affected by the conformational properties of a particular ncPro (Figure 5.2a).
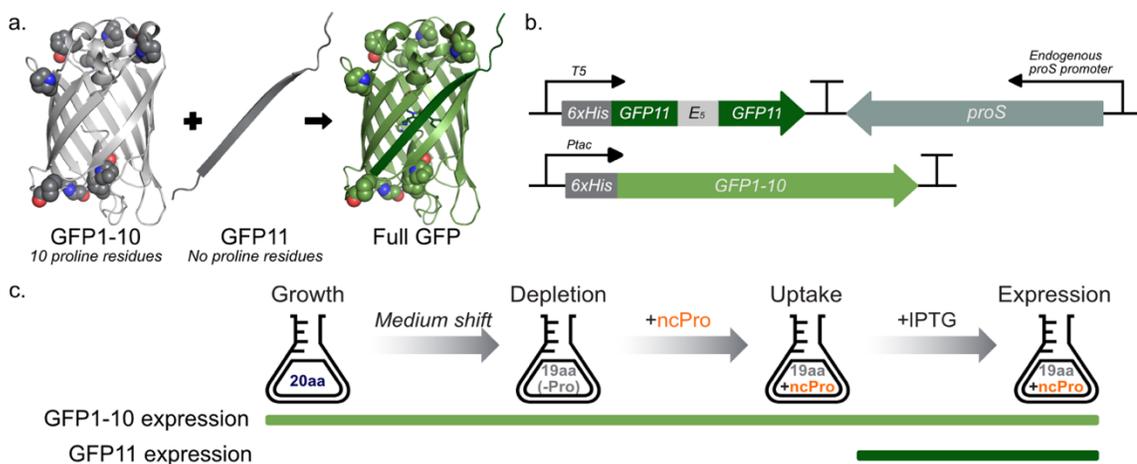

Figure 5.2. Design of split-GFP to measure proline analog incorporation. **a**. GFP1-10 (which contains ten proline residues, rendered as spheres) is expressed in proline-containing medium. Proline-dependent expression of the GFP11 fragment is induced after a medium shift into ncPro-containing medium. Complementation of the two fragments enables chromophore formation, providing a fluorescent output for ncPro incorporation (PDB: 2B3P). **b**. The initial construct design, as previously described.[19] **c**. Experimental expression scheme for measuring ncPro incorporation by GFP1-10/11 complementation. GFP1-10 is under the control of a strong constitutive promoter, so is expressed throughout the duration of the experiment; expression of GFP11 is induced only after a medium shift into ncPro medium.

Our first efforts[19] utilized split GFP1-10/11 (Figure 5.2b), a pair of protein fragments initially used to screen for protein solubility.[20] With this approach, the proline-containing GFP fragment GFP1-10 (containing the first ten β strands of GFP) was constitutively expressed with a modified version of the strong *tac* promoter[21] (missing the LacI binding site) on a pBAD33 backbone. Two proline-free GFP11 strands, separated by an elastin-like

protein linker, were placed under control of the IPTG-inducible T5 promoter on a pQE-80L backbone. One proline codon preceded the first GFP11 domain. With the five proline codons in the elastin linker, expression of the full GFP11-elastin-GFP11 fusion protein (which will henceforth generally be referred to as "GFP11") was dependent upon readthrough of six proline codons. Complementation of the two GFP fragments, an association reported to occur with sub-picomolar affinity,[22] should form the completed β-barrel of GFP. Subsequent chromophore formation would provide a fluorescent readout of ncPro incorporation. The scheme outlining initial expression conditions is shown in Figure 5.2c. Finally, *proS*, the gene encoding the *E. coli* ProRS, was placed under control of its endogenous promoter and on the same plasmid backbone as GFP11 (Figure 5.2b). We envisioned replacing this wild-type ProRS with a library of mutants when screening for improved variants.
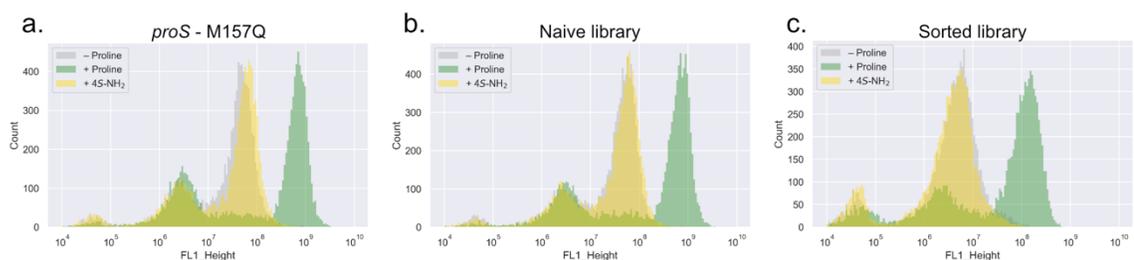

Figure 5.3. 4S-NH$_2$ incorporation does not improve after sorting. Flow cytometry after a split-GFP1-10/11 expression experiment for *E. coli* strain SLB2001 overexpressing the following ProRS variants or libraries: **a**. ProRS-M157Q; **b**. Error-prone PCR library. **c**. 1x sorted library. The fluorescence of the population treated with 4S-NH$_2$ did not increase relative to the negative (–Proline) control after sorting.

Initial efforts appeared promising: separation could be discerned between +/– proline samples when using *E. coli* strain SLB2001, a proline auxotroph derived from DH10B. Further, low levels of 4S-aminoproline (4S-NH$_2$) incorporation could be observed as a

slight increase in fluorescence over background (Figure 5.3a) when overexpressing the M157Q ProRS mutant, which is known to modestly improve 4S-NH$_2$ incorporation.[19] However, sorting an error-prone PCR library of ProRS variants for incorporation of 4S-NH$_2$ did not improve relative fluorescence (Figure 5.3b-c).

5.4.3 Efforts to improve split-GFP dynamic range reveals poor GFP1-10 solubility

Several control experiments gave rise to concern. First, we noted high background fluorescence without adding proline or a ncPro after the medium shift, or inducing GPF11 expression (Figure 5.4a). We did not detect an increase in fluorescence over background for strains containing the GFP1-10 gene alone (the –GFP11 population is representative of *E. coli* autofluorescence only; data not shown), suggesting that these results are due to leaky GFP11 expression, and proline codon read-through. We did not detect incorporation of any other canonical amino acid into GFP11 in the absence of proline by mass spectrometry (Figure 5.4b-c). Increasing wash step stringency during the medium shift did not have any effect on background fluorescence (data not shown).

Residual proline incorporation appears to be a general phenomenon in all proline replacement approaches described in this thesis. Low levels of proline-containing proinsulin are nearly always observed by mass spectrometry, even when using proline analogs with high incorporation efficiencies (Figure 2.2a-d, Figure 3.2a-d), and low levels of proinsulin can be observed by SDS-PAGE even in the absence of proline (Figure 2.S1).
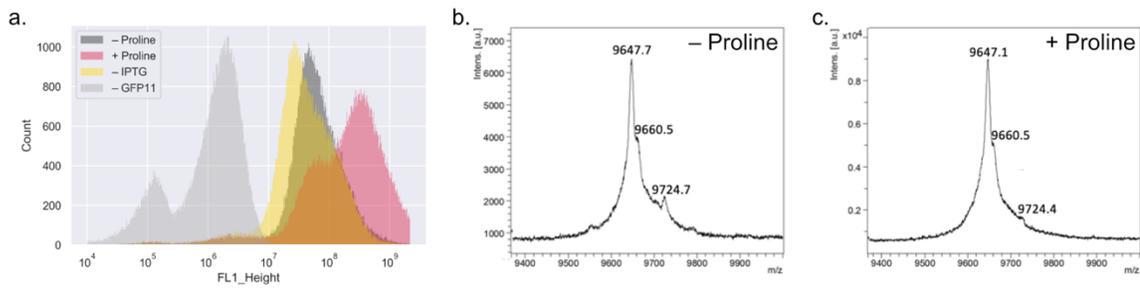
Figure 5.4. Leaky GFP11 expression and residual proline result in high background fluorescence. **a**. Fluorescence of the split-GFP system as measured by flow cytometry. Negative controls include a strain lacking the GFP11 gene (–GFP11), no induction of GFP11 expression (–IPTG), and no addition of proline (–Proline). **b-c**. MALDI-TOF spectra of purified GFP11 expressed in the absence (b) or presence (c) of proline. Expected m/z of proline-containing GFP11 fragment = 9647.5. No significant levels of amino acid misincorporation were detected.
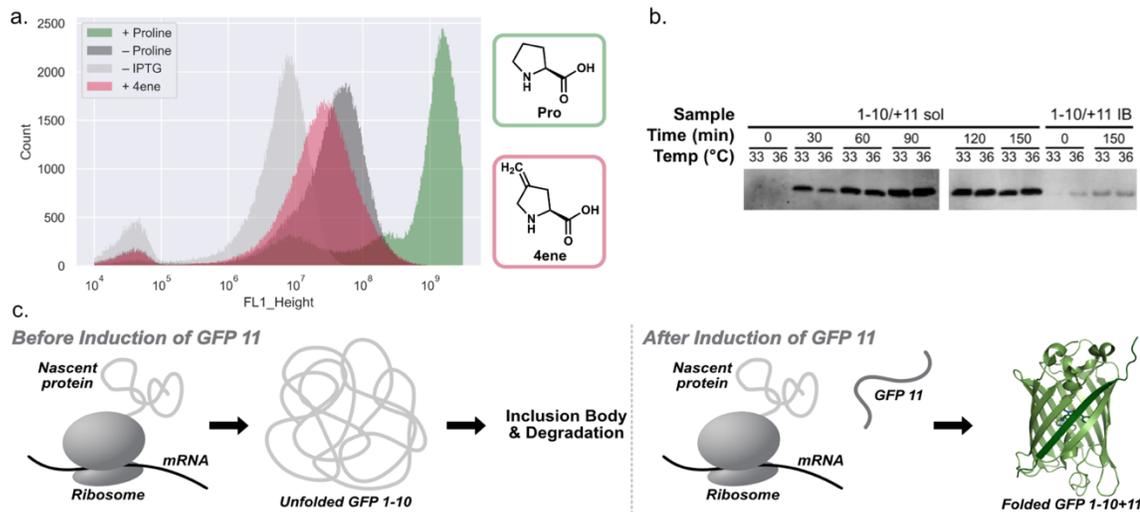

Figure 5.5. Poor solubility impairs split-GFP performance. **a**. Split-GFP flow cytometry results. 4-ene (~90% incorporation efficiency, Chapter II) is less fluorescent than the "–Proline" control. **b**. Western blot detecting the presence of GFP1-10 in the soluble (sol) or inclusion body (IB) fraction over time after inducing GFP11 expression; similar effects were observed at both 33 and 36 °C. **c**. Proposed model for GFP1-10/11 expression in *E. coli*. In the absence of GFP11, GFP1-10 is insoluble and unstable in *E. coli*. In the presence of GFP11, protein stability and chromophore formation are achieved.

Tighter control of GFP11 expression with the *araBAD* promoter[23] moderately improved the resolution between positive and negative controls (+/– Proline, Figure 5.5a). However, treatment with the proline analog 4ene, which exhibits high levels of incorporation during proinsulin expression (see Chapter II of this thesis), did not result in a fluorescent population commensurate with the expected incorporation efficiency. In fact, this population was *darker* than the negative control, a sample which did not contain any additional proline (Figure 5.5a). This result was reminiscent of the mWasabi example above (Figure 5.1b).

A western blot analysis of the soluble and insoluble fractions illustrated a significant difference in protein solubility in the presence and absence of GFP11 (Figure 5.5b). Before induction of GFP11 expression (time 0), no GFP1-10 was detected by western blot at either temperature tested. However, robust expression could be detected in the soluble fraction 30 minutes after inducing GFP11 expression, and grew more prominent over time. Faint bands could be detected in the insoluble, inclusion body fraction before and after GFP11 expression.

We propose that in the absence of GFP11, GFP1-10 does not fold well, and either is actively degraded by the cell, or is insoluble. As a result, proline-containing GFP1-10 expressed before the medium shift is unavailable for GFP11 complementation after ncPro incorporation. In the presence of GFP11, newly synthesized GFP1-10 remains soluble by associating with GFP11, and is capable of chromophore formation (Figure 5.5c). Importantly, the only GFP1-10 able to associate with GFP11 is that translated *after* the medium shift and *in the presence of the ncPro*, so the advantage of using split-GFP over full-length fluorescent proteins is lost. While many others have used this particular split-

GFP pair in a wide range of contexts,[24–31] we are not aware of any studies that require temporal separation of each GFP fragment. In fact, in manuscripts describing uses of GFP1-10/11, GFP1-10 is either expressed after the GFP11 fragment, or is purified from the inclusion body fraction and then refolded.[25] We have also observed that strains constitutively expressing GFP1-10 under control of the modified *tac* promoter exhibit substantially slower growth rates (data not shown), suggestive of GFP1-10 toxicity.

5.4.4 Solubility tags improve GFP1-10 solubility

Decreasing expression temperature to 25°C did restore some GFP1-10 solubility in the absence of GFP11, albeit at reduced levels (Figure 5.6a). To better control protein expression, the constitutive *tac* promoter previously used to drive GFP1-10 expression was replaced with the IPTG-inducible *T5* promoter (Figure 5.6b). We envisioned inducing GFP1-10 expression in proline-containing medium at mid-log phase. After incubation for 3 h at 25°C, cells would be shifted into a proline-free medium. The ncPro of interest was added under osmotic stress conditions, and expression of the GFP11 fragment induced (Figure 5.6c). However, no significant increase in fluorescence could be detected with the addition of proline (Figure 5.6d). These data suggest that, despite the apparent increased solubility at lower temperatures, there are still insufficient quantities of soluble GFP1-10 available for complementation after the medium shift.
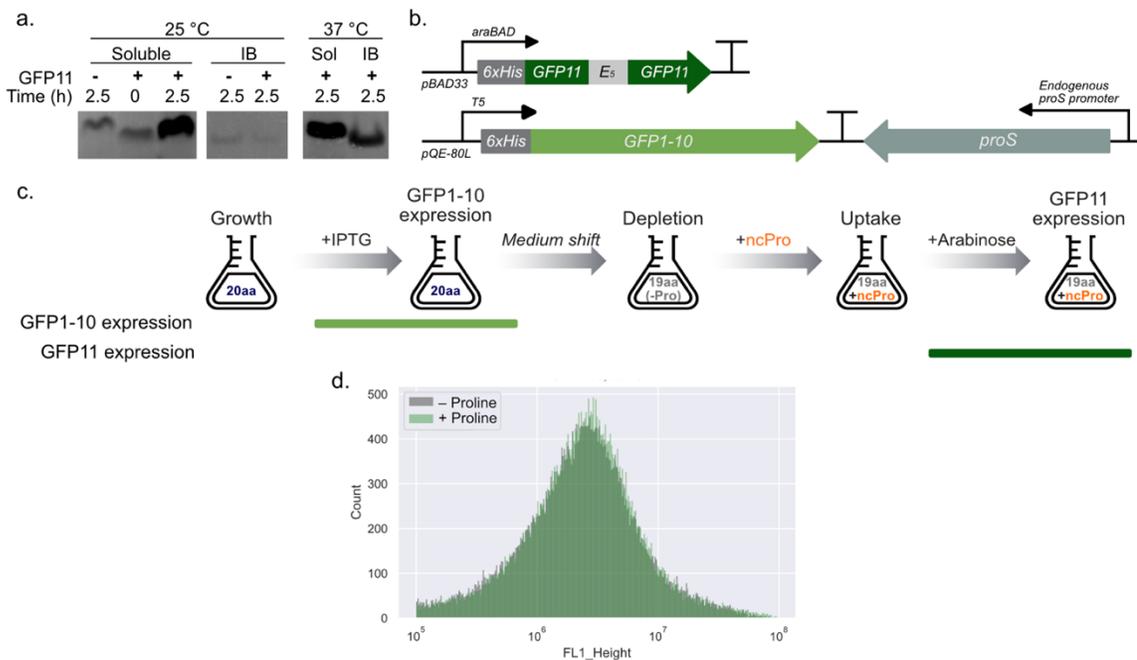
Figure 5.6. Staggered expression of split-GFP at 25°C. **a**. Western blot analysis of GFP1-10 in the soluble and inclusion body (IB) fractions during split-GFP expression. GFP1-10 was expressed constitutively, and expression of GFP11 was induced at time=0 at 25 and 37°C. **b**. Redesigned plasmid scheme that includes inducible control of expression for both GFP fragments. **c**. Staggered expression protocol. Expression of GFP1-10 is induced at mid-log phase in proline-containing medium prior to a medium shift to proline-free medium. GFP11 expression is then induced in the presence of the ncPro. Association and fluorescence are dependent upon the ability of soluble GFP1-10 to persist throughout the experiment. **d**. Fluorescence in the presence and absence of proline, as measured by flow cytometry.

"Supercharging" proteins is an approach to solubilizing proteins that involves replacing nonpolar surface residues with charged amino acids.[32] Both super-positive and super-negative versions of GFP with net charges of +36 and -30, respectively, have been developed. We cloned the corresponding split super-positive and -negative versions of GFP1-10 (referred to as "spGFP" and "snGFP") and assessed their ability to complement GFP11. However, even simultaneous expression of both GFP1-10/11 fragments did not result in an increase in fluorescence, as measured by plate reader (Figure 5.7a,b). These

data indicate that chromophore formation is inhibited for these charged GFP variants, perhaps due to decreased fragment association.
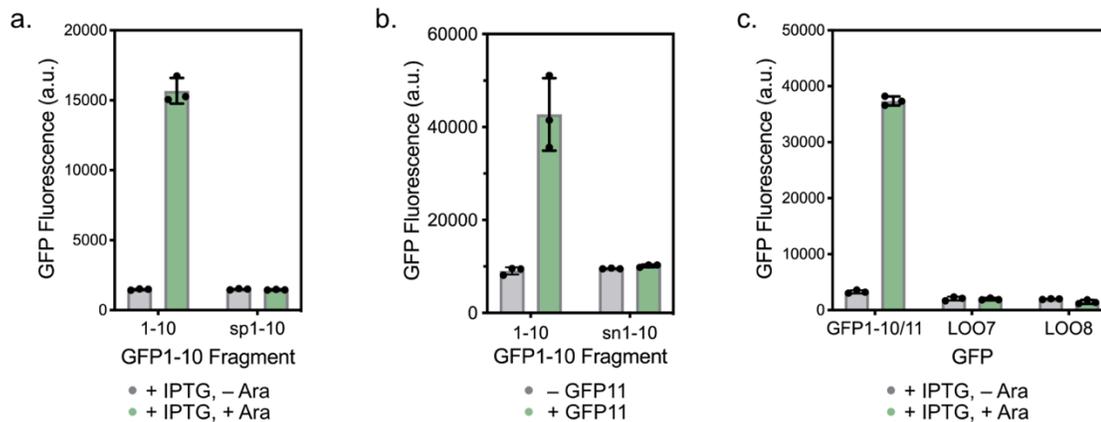

Figure 5.7. Fluorescence of split supercharged and permuted GFP variants. IPTG and arabinose induce expression of the 1-10 and 11 fragments, respectively. Complementation was assessed by co-expressing each GFP1-10/11 fragment. **a**. sp1-10/GFP11 complementation was assessed in rich medium after inducing expression of each GFP fragment. **b**. sn1-10/GFP11 complementation was assessed in rich medium by measuring fluorescence of strains containing or lacking the GFP11 gene. **c**. Complementation of circularly permuted GFP fragments was assessed in minimal medium after inducing expression of each GFP fragment.

"LOO7" and "LOO8" are versions of split-GFP in which either the 7th or 8th strand of the β-barrel of a circularly permuted GFP variant is omitted.[33] Neither the 7th or 8th strand of GFP contains proline residues, and the resulting large fragment was reported to be either more soluble (LOO7) or lead to greater fluorescence (LOO8) than the split-GFP1-10/11 version.[33] However, we did not observe fluorescence upon co-expression of both fragments for either split-GFP variant tested (Figure 5.7c).

A common approach to increase protein solubility is fusion to a solubility tag. We translationally fused MBP, TrxA, SUMO, or NusA to the N-terminus of the GFP1-10 fragment. In initial experiments, the constitutive *tac* promoter described above was used to

drive expression of the GFP1-10 fragments; GFP11 expression was controlled by the T5

promoter (Figure 5.2b). We probed the solubility of each GFP1-10 fusion protein by

western blot in the presence and absence of GFP11. All fusion proteins were found in the

soluble fraction when expressed in the presence of GFP11. In the absence of GFP11, small

but detectable levels of 1-10 fusion proteins could be observed in the soluble fraction
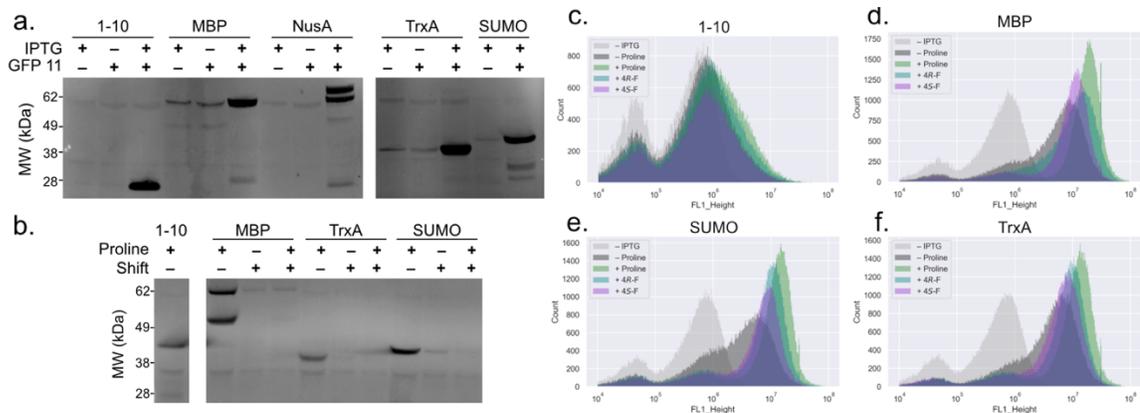
(Figure 5.8a).

Figure 5.8. GFP1-10 solubility tag fusions improve resolution by flow cytometry. **a**. Western blot analysis of constitutively expressed GFP1-10 fragments in the soluble fraction at 37°C in the presence or absence of the GFP11 gene, and with or without inducing GFP11 expression with IPTG. **b**. Western blot analysis of GFP1-10 in the soluble fraction during expression in minimal medium at 25°C. Samples were collected either before or after the medium shift; GFP1-10 without a solubility tag was expressed in rich (LB) medium. **c-f**. Flow cytometry analysis of GFP1-10, or GFP1-10 fusion proteins. Flow cytometry results were obtained from staggered expression of each GFP1-10/11 fragment: GFP1-10 was first expressed in proline-containing medium, and GFP11 was expressed after a medium shift.

To measure protein solubility under conditions that more closely mimic ncPro

incorporation, we grew cells in M9 medium, performed a medium shift, and controlled

expression of both split-GFP fragments with the pBAD/pQE80 construct design (Figure

5.6b). We observed soluble GFP1-10 at 25°C even in the absence of GFP11 (Figure 5.8b),

though levels of each GFP1-10 fusion protein in the soluble fraction decreased substantially after a medium shift.

Tagged split-GFP constructs were assessed by flow cytometry. In the absence of a solubility tag, minimal levels of fluorescence over background were observed (Figure 5.8c), consistent with poor GFP1-10 solubility even at lowered temperatures. For the tagged GFP1-10 constructs, we did observe an increase in fluorescence for the +Proline positive control. Further, samples treated with 4R-F and 4S-F (proline analogs with distinct conformational behaviors[5]) both exhibited fluorescence (Figure 5.8d-f), suggesting that this may be a general approach to assess ncPro incorporation. However, the resolution between the positive (+Proline) and negative (–Proline) controls was low, and likely insufficient for effective library sorting by FACS.

5.4.5 Changing expression protocols does not improve screen performance

Based on the observation that the medium shift reduces the levels of GFP1-10 in the soluble fraction (Figure 5.8b), we hypothesized that performing a medium shift increases protein degradation in *E. coli* and leads to reduced dynamic range. We explored alternative GFP1-10/11 expression protocols, in which the expression of GFP1-10 would be induced *after* the expression of GFP11 (Figure 5.9a-b). This protocol would rely upon the stability of GFP11, rather than GFP1-10. However, these approaches did not improve screen performance (Figure 5.9c-d).
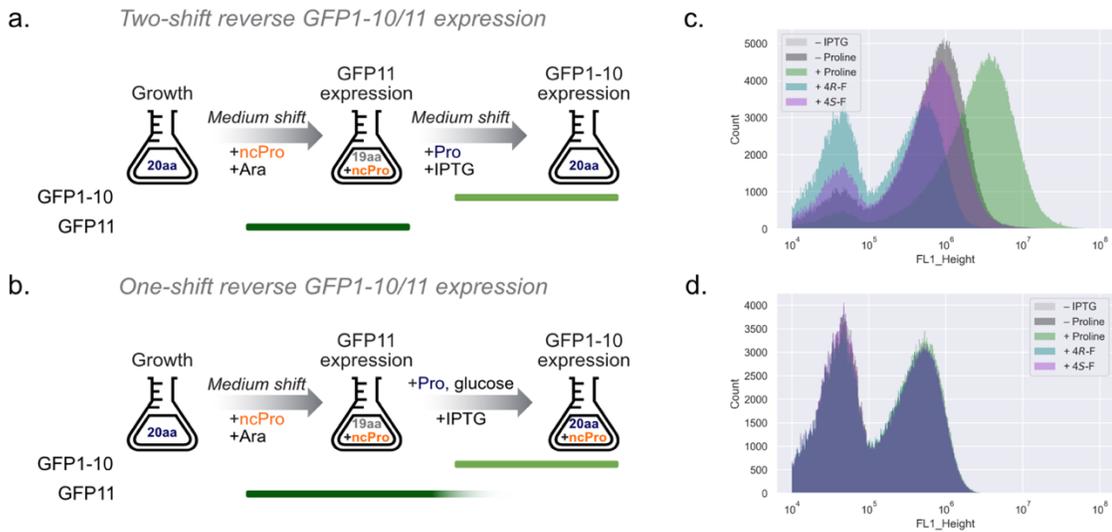
Figure 5.9. GFP1-10/11 reverse expression. **a**. Expression protocol for a two-shift reverse GFP1-10/11 expression protocol. GFP11 expression is induced after the first shift to ncPro medium; GFP1-10 expression induced after a second shift to proline medium that contained glucose as a carbon source to repress GFP11 expression. **b**. One-shift reverse protocol: proline (to out-compete the ncPro) and glucose (to repress GFP11 expression) were added before inducing expression of GFP1-10. **c-d**. Flow cytometry results for two-shift (c) and one-shift (d) ncPro incorporation experiments.
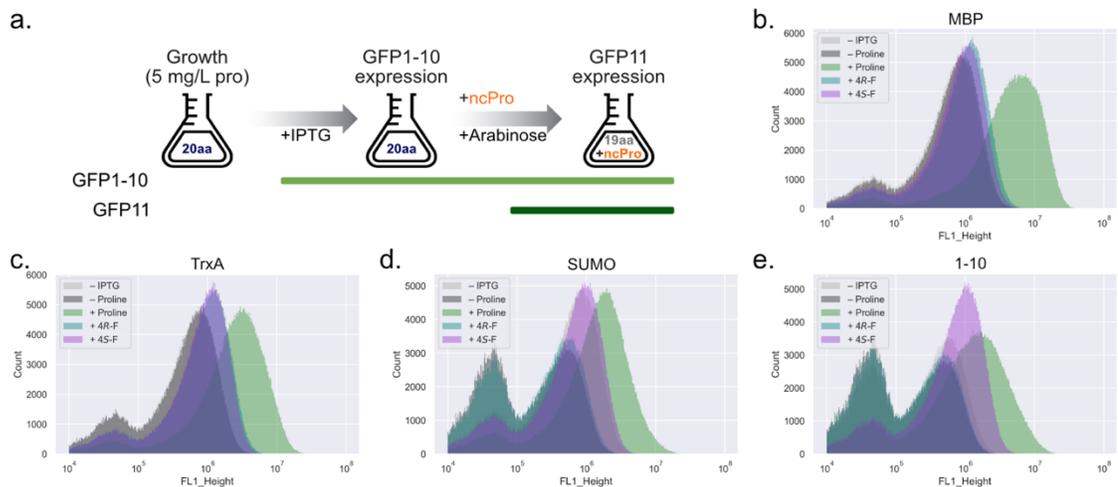

Figure 5.10. SPI expression of GFP1-10/11. **a**. Expression protocol for SPI-based expression of split-GFP. **b-e**. Flow cytometry results for ncPro incorporation experiments with MBP–GFP1-10 (b) TrxA–GFP1-10 (c), SUMO–GFP1-10 (d), or GFP1-10 (e).

An alternative method used for residue-specific incorporation of ncAAs is selective pressure incorporation (SPI).[34] Rather than use a medium shift, this approach relies upon bacterial growth in amino acid limiting medium, before adding the ncAA. In the context of GFP1-10/11 expression (Figure 5.10a), SPI did improve resolution between the +Proline and –Proline controls by drastically reducing background in the absence of proline. However, samples treated with the translationally active 4-fluoroprolines tested were barely fluorescent above background (MBP–1-10 and TrxA–1-10, Figure 5.10b-c), or did not demonstrate similar levels of fluorescence between the two diastereomers (SUMO–1-10 and GFP1-10, Figure 5.10d-e).

5.4.6 Alternative GFP11 designs do not improve performance

XTEN is an unstructured protein commonly used to increase the half-life of therapeutic proteins and peptides.[35] This domain contains many proline residues, and its length can be easily changed to tune screening stringency. We hypothesized that, because of its disordered, hydrophilic nature, ncPro incorporation would not affect its solubility. We constructed two versions of XTEN-GFP11 fusion proteins: XTEN72-GFP11, which contains a 72-residue XTEN domain with 13 proline codons N-terminally fused to the GFP11 fragment; and XTEN144-GFP11, which contains 25 residues N-terminal to GFP11. These genes were installed in the pBAD33 plasmid, replacing the GFP11-elastin-GFP11 gene shown in Figure 5.6b. We expressed the MBP–GFP1-10 fusion protein for 3 h at 25°C. After a medium shift and addition of ncPro residues, XTEN-GFP11 expression was induced with arabinose. We measured the fluorescence of each construct after 135 min of GFP11 fragment expression, and compared it to the original GFP11-elastin-GFP11 design (Figure 5.11).

As expected, we find that including more proline codons upstream of the GFP11 gene does improve the resolution between the –Proline and +Proline controls. The 25 proline codons in XTEN144 are too stringent for ProRS evolution efforts: even 4*R*-F, which is incorporated with high efficiency,[5] is barely discernable over background for this strain. Fluorescence of the 4*S*-F treated XTEN samples is most concerning, however. Compared to the original GFP11-elastin-GFP11 design, in which both 4*R*-F and 4*S*-F display fluorescence levels between the positive and negative controls (Figure 5.11a), the 4*S*-F–treated XTEN samples are not fluorescent over background (Figure 5.11b-c). These results suggest that 4*S*-F incorporation is not well tolerated by XTEN, perhaps due to low protein solubility after incorporation, and demonstrate that this design is not a good general approach to measure ncPro incorporation.

Other attempts at improving resolution of our split-GFP system using alternative proline auxotrophic strains of *E. coli*, or knocking out the *lon* protease, were similarly unsuccessful (data not shown).
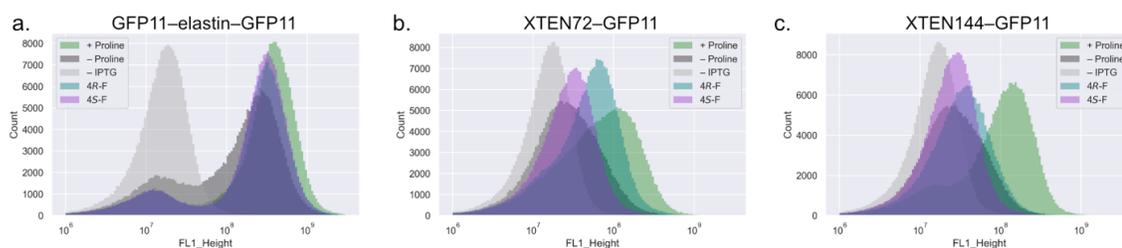
Figure 5.11. XTEN-GFP11 fusion proteins to measure proline analog incorporation. Cells were grown in minimal medium until early log phase, at which point expression of the MBP–GFP1-10 fusion protein was induced for 3 h at 25°C. After a medium shift and ncPro addition, the original GFP11 fragment (a), or GFP fused to the C-terminus of XTEN72 (b) or XTEN144 (c), was expressed for 135 min. Fluorescence was assessed by flow cytometry.

5.4.7 Inclusion body analysis by flow cytometry

As detailed earlier in this thesis, we are interested in applying ncPro mutagenesis to study and modulate the behavior of the peptide therapeutic insulin. To prepare our insulin variants, we first express proinsulin, a precursor to insulin, in *E. coli*. Proinsulin is isolated from the inclusion body fraction before refolding and maturation. Inclusion body formation in *E. coli* has been monitored by measuring the scattering properties via flow cytometry;[36] we wondered if a similar technique could be used to assess proinsulin production, which correlates well with ncPro incorporation efficiency (Figure 2.S1, Table 2.S1). While the resolution and dynamic range of such an approach is expected to be low, it would have the advantage of directly observing proinsulin production (rather than monitoring the expression of a reporter protein). To assess feasibility, we measured the scattering properties of *E. coli* by flow cytometry after a typical proinsulin expression experiment. However, no difference in scattering was observed after inducing proinsulin expression (Figure 5.12).
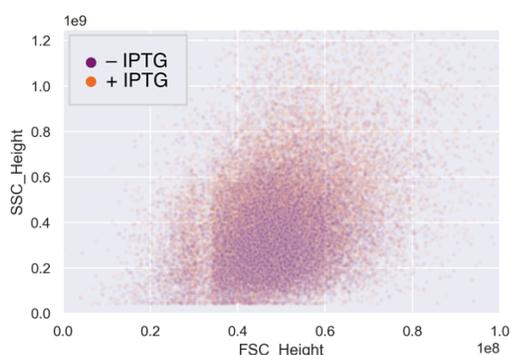

Figure 5.12. Interrogating inclusion body formation by flow cytometry. Forward-scatter (FSC) vs side-scatter (SSC) properties of *E. coli* with (+IPTG) or without (–IPTG) inducing proinsulin expression, as measured by flow cytometry.

5.4.8 Survival on proline analog-containing medium

Because our ncPro incorporation strategy relies upon a residue-specific approach that leads to proteome-wide proline replacement, we hypothesized that extended ncPro exposure might lead to toxicity. Differential growth between high- and low-ncPro incorporating strains, determined after replica plating, might be used to screen for improved ncPro incorporation. Here, inhibited growth in the presence of the ncPro would indicate higher levels of incorporation.

Strains overexpressing either the wild-type ProRS, or M157Q mutant (which increased incorporation of the ncPro 4S-NH$_2$, Ref. 19) were plated onto agar plates with rich (LB) medium. We transferred colonies by replica plating to M9 medium agar plates that contained low (1-10 μM) concentrations of proline, and increasing (0-500 μM) concentrations of 4S-NH$_2$. However, we did not observe differences in colony formation as a function of known 4S-HN$_2$ incorporation efficiency (data not shown).

5.4.9 A small molecule alternative to fluorescent reporter proteins

Small molecule organic dyes possess several advantageous properties compared to fluorescent proteins, such as increased brightness and photostability.[37] Since a small molecule dye is not itself genetically encoded, but can be designed to specifically label a protein of interest, judicious choice of a labeling target might avoid the ncPro-specific effects on full-length fluorescent proteins.

FlAsH-EDT$_2$ is a small organoarsenic molecule derived from the fluorophore fluorescein (Figure 5.13a). In its unbound state, FlAsH-EDT$_2$ is not fluorescent. However, upon binding to a tetra-cysteine (TC) peptide motif, increased conformational restriction leads

to fluorescence upon irradiation.[38] FlAsH is especially useful for protein localization studies: compared to fusing a protein of interest to a large fluorescent protein, which can lead to artifacts in imaging and protein behavior,[39] the FlAsH system requires only a short peptide sequence.[38] Despite these advantages, FlAsH is not widely used, perhaps due to high background labeling.[40]
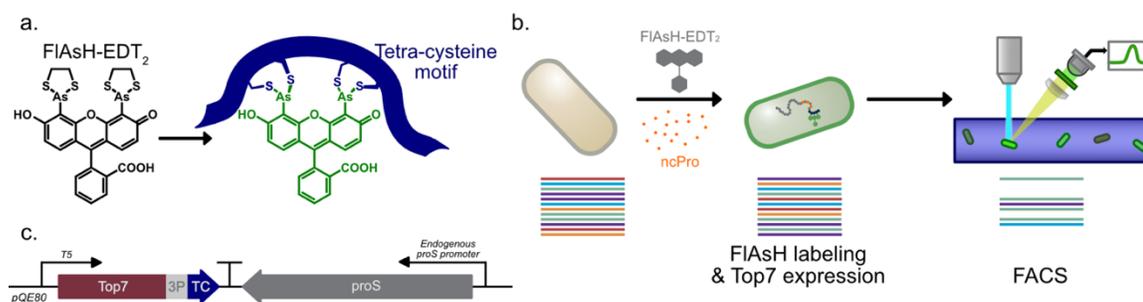

Figure 5.13. A FlAsH-based labeling strategy to measure proline analog incorporation. a. The chemical structure of the biarsenical fluoresceine derivative FlAsH-EDT$_2$. FlAsH fluoresces after binding to a tetra-cysteine (TC) motif, which can be translationally fused to a protein of interest. **b**. Proposed protocol for ProRS engineering by FlAsH labeling. A library of ProRS variants (represented by the colored lines) are expressed in *E. coli*. A TC motif is translationally fused to a soluble, proline-free protein and proline-containing linker. FlAsH labeling detects TC motif expression after ncPro incorporation, which can be detected by flow cytometry. FACS would sort cells based on fluorescence, leading to enrichment of ProRS library members capable of enhanced ncPro incorporation. **c**. Plasmid design. The expression of Top7-TC3 is controlled by the IPTG-inducible T5 promoter; the ProRS library is installed on the same pQE-80L plasmid backbone.

We envisioned applying a FlAsH-based labeling approach to measure ncPro incorporation, and took several design considerations into account. First, we identified Top7 as a candidate carrier protein to which the TC motif could be fused. Top7 is the result of a computational protein design endeavor, and, importantly, contains no proline residues.[41] Its behavior (i.e., solubility or susceptibility to degradation) should not be dependent upon

the ncPro residue of interest. Second, we appended a proline-containing linker sequence to the C-terminus of Top7 that would connect the carrier protein to the TC motif. Readthrough of the proline codons in this linker, presumably by ncPro incorporation, would allow for translation to continue to the TC motif. Finally, we considered the TC motif itself. The generic TC motif is the amino acid sequence CCXXCC, where X represents any amino acid. However, an internal proline-glycine dipeptide (as in the TC motif CCPGCC) has been found to most effectively engage FlAsH, since the fluorophore binds best to β-hairpin turns.[42] Because fluorescence after FlAsH binding to a CCPGCC TC motif depends on the ncPro incorporated (Figure 5.S1), we elected to search for alternate TC motifs that do not contain proline residues. Sorting a library of proline-free TC motifs by FACS (Figure 5.S2) identified TC3, whose amino acid sequence is YCCGVCCI.

Typical FlAsH labeling experiments proceeded as follows (Figure 5.13b). An *E. coli* proline auxotrophic strain harboring plasmid pQE80_Top7-TC3_proS, which contains an IPTG-inducible Top7-TC3 gene and the *E. coli proS* under the control of its endogenous promoter (Figure 5.13c), was grown at 37°C to mid-log phase in minimal medium that contained proline. Cells were washed with cold 0.9% NaCl, and resuspended in medium lacking proline. FlAsH was added at this stage, since labeling cells after Top7 expression led to poor fluorophore uptake and significant heterogeneity (Figure 5.S3). After 30 min incubation to deplete residual proline, the ncPro of interest was added to the culture. High (0.3 M) levels of sodium chloride were also added to facilitate ncPro uptake.[6] After inducing expression of Top7-TC3 with IPTG, samples were washed and analyzed by flow cytometry; fluorescence should correlate with ncPro incorporation.

5.4.10 Optimization of FlAsH labeling conditions does not improve sorting performance

We tested a range of proline auxotroph strains, and found that all yielded similar results by flow cytometry (Figure 5.14a-e). We chose to continue with strain CAG18515, since we use this strain for proinsulin expression. The bright tail present in many samples was removed through more stringent wash steps that use 3% DMSO and ethanedithiol (EDT); representative histograms of Top7 expression and optimized FlAsH labeling and wash conditions are shown in Figure 5.14f.
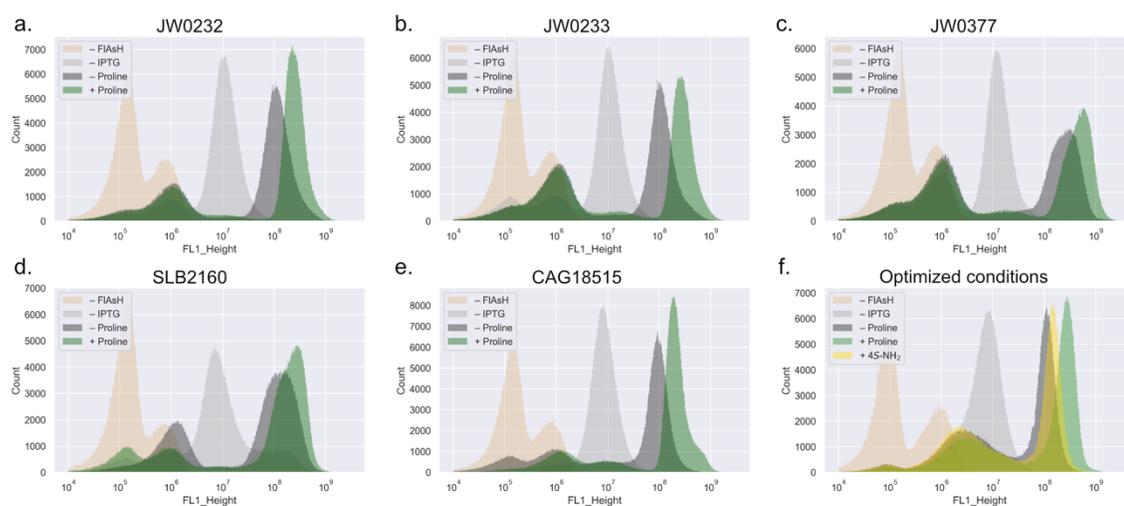
Figure 5.14. Optimization of FlAsH labeling and protein expression conditions. **a-e**. Flow cytometry histograms of proline auxotrophic strains of *E. coli* after Top7-TC3 expression and non-optimized FlAsH labeling. Strain SLB2160 is ampicillin resistant, so plasmid pQE80-c_Top7-TC3_proS, which confers resistance to chloramphenicol, was used in that case. **f**. Lower concentrations of FlAsH added during labeling, and more stringent washes, improves fluorescence histogram appearance; shown are optimized labeling conditions for strain CAG18515.

We cloned a site-saturation mutagenesis library by standard restriction enzyme cloning approaches, targeting five residues in the ProRS active site (Figure 5.15a). Extreme care was taken to purify correctly-sized plasmid DNA by agarose gel. Nevertheless, sorting the

brightest ~0.1% of cells after treatment with 4S-NH$_2$ did not result in increased fluorescence, and instead led to a more heterogeneous population (Figure 5.15b-c). Preliminary analysis of the sorted plasmid DNA suggested loss of the *proS* gene (data not shown). We also noted low rescue efficiencies in our sorting experiments: after rescuing sorted cells in rich medium at 37°C, <1% of our sorted events led to colony-forming units (CFUs) when plated onto selective agar plates.
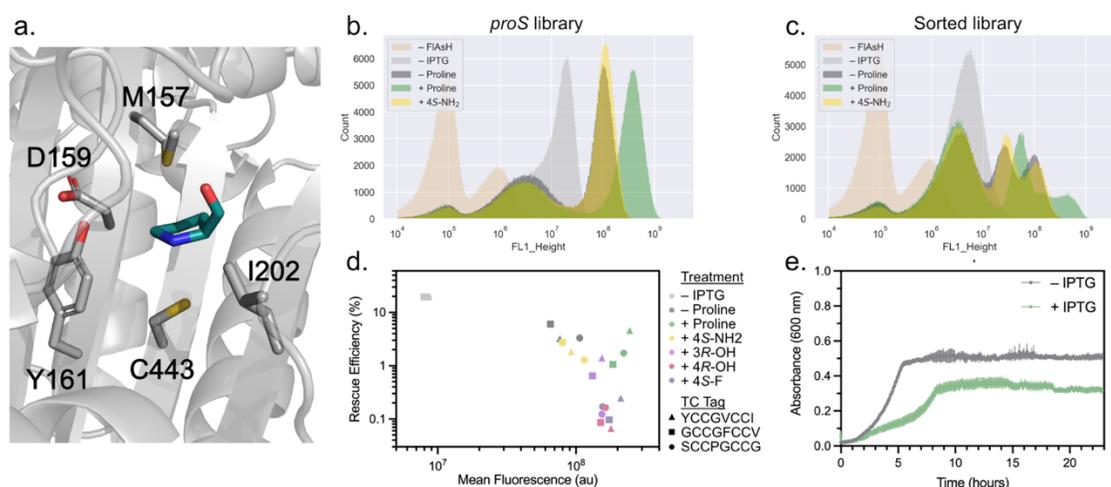
Figure 5.15. Sorting after FlAsH labeling reveals poor rescue efficiency and Top7 toxicity. **a**. Homology model of the *Ec*ProRS bound to prolinol (teal), an analog of proline used in the initial crystal structure. The model was created using SWISS-MODEL[43] using the structure of the ProRS from *Enterococcus faecalis* (PDB: 2J3M, 42% homology) as the template. **b**. Flow cytometry histograms after Top7-TC3 expression and FlAsH labeling. The top ~0.1% of the sample treated with 4S-NH$_2$ was sorted, rescued in rich medium, and was subjected to a second round of analysis by flow cytometry (**c**.). **d**. The entire bright population of a clonal sample overexpressing the wild-type ProRS was sorted after Top7 expression and FlAsH labeling. Sorted cells were rescued in rich medium for 1 h at 37°C, then plated onto selective agar plates. Rescue efficiency was defined as CFUs divided by total events sorted. **e**. Growth with (+IPTG) and without (–IPTG) inducing Top7-TC3 expression. Cells were diluted into M9 medium with 0.3 M NaCl (to simulate ncPro incorporation conditions), and growth was monitored in a 96-well plate at 37°C with shaking.

We sorted the entire bright population (FL1 > ~10^7) under a variety of expression conditions, and found that rescue efficiency was inversely correlated with fluorescence across three TC tags assessed (Figure 5.15d). Especially notable is the decreased rescue efficiency between –Proline (which should result in *lower* viability for this proline auxotrophic strain of *E. coli*), and +Proline conditions. Further, Top7 expression substantially slowed growth (Figure 5.15e). We hypothesized that increased Top7 expression (which should indicate higher levels of ncPro incorporation) hindered growth and survival after sorting, an undesired property for a protein engineering endeavor. These effects might enable out-competition by any cells containing undesired plasmids; for instance, those found here lacking the *proS* gene.

We identified eight other soluble, proline-free proteins, tested their toxicity when expressed in *E. coli* (Figure 5.16a-h), and analyzed select strains by flow cytometry after FlAsH labeling (Figure 5.16i-l). The protein DHR14, a highly stable, computationally-designed helical repeat protein[44] performed best. We note here that sample heterogeneity (in particular, the darkest population for each FlAsH-labeled sample) was present in all strains assessed by flow cytometry (Figure 5.16i-l). We find that dark cells were physiologically distinct: their scattering properties differed from brighter cells (Figure 16m), and appeared elongated by confocal microscopy (Figure 16n-p). Shown are images for strains expressing DHR14-TC3; Top7-TC3 strains behaved similarly (data not shown).
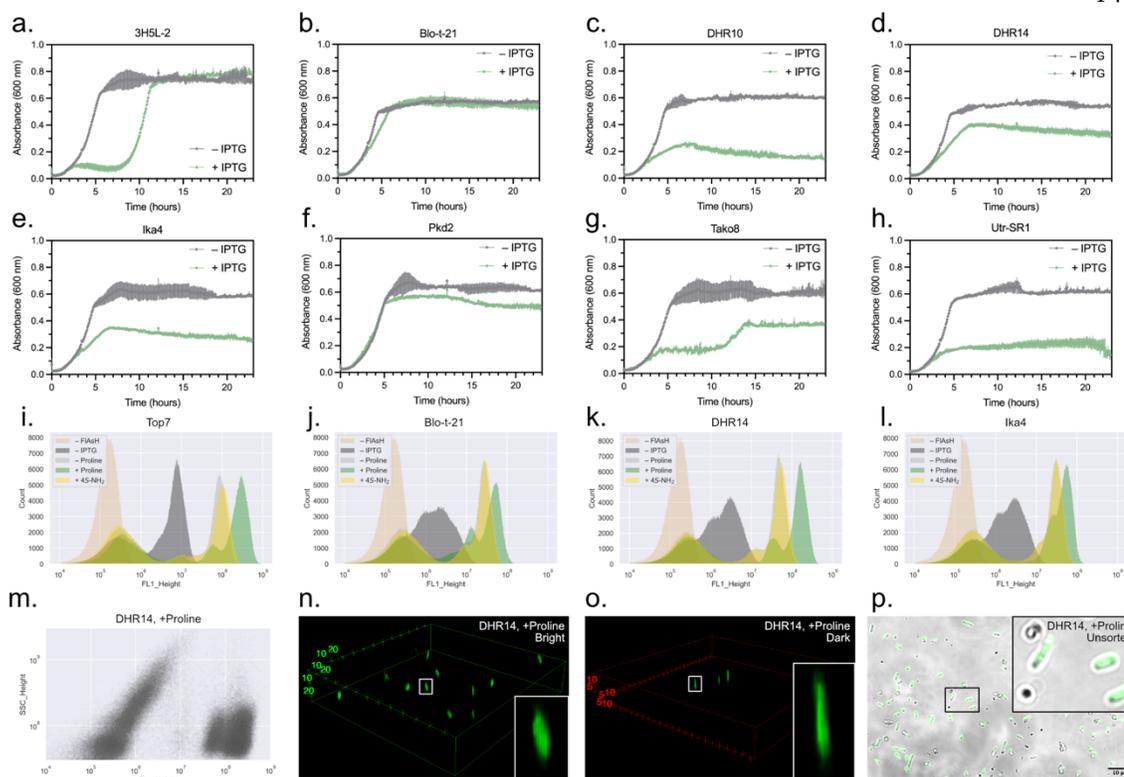
Figure 5.16. Alternative proline-free target proteins. **a-h**. Eight alternative proline-free proteins were identified, and replaced the Top7 gene in plasmid pQE80_Top7-TC3_proS. Growth of strain CAG18515 carrying each of these plasmids in M9 medium +0.3 M NaCl was monitored by tracking absorbance at 600 nm in a plate reader at 37°C with shaking. **i-l**. Flow cytometry histograms of select proline-free proteins after ncPro incorporation and FlAsH labeling. We chose carrier proteins whose expression limited growth the least; Pkd2 was omitted because we could not detect its expression by SDS-PAGE (data not shown). Note that the shoulder of the brightest population is likely associated with oxidized EDT: its presence disappears with fresh EDT (e.g., Fig. 5.14f). **m**. Fluorescence (FL1) versus side-scatter (SSC) of proline-treated DHR14-TC3 in panel k; all other FlAsH-labeled samples behaved similarly. **n-p**. Confocal microscopy images after DHR14-TC expression and FlAsH labeling for proline-treated samples. Bright (n) and dark (o) populations were sorted, then imaged with different microscope settings that optimized dynamic range for each sample. Unsorted cells (p) illustrate the difference in fluorescence and morphology; here, a brightfield image is overlaid with the fluorescence z-stack image.
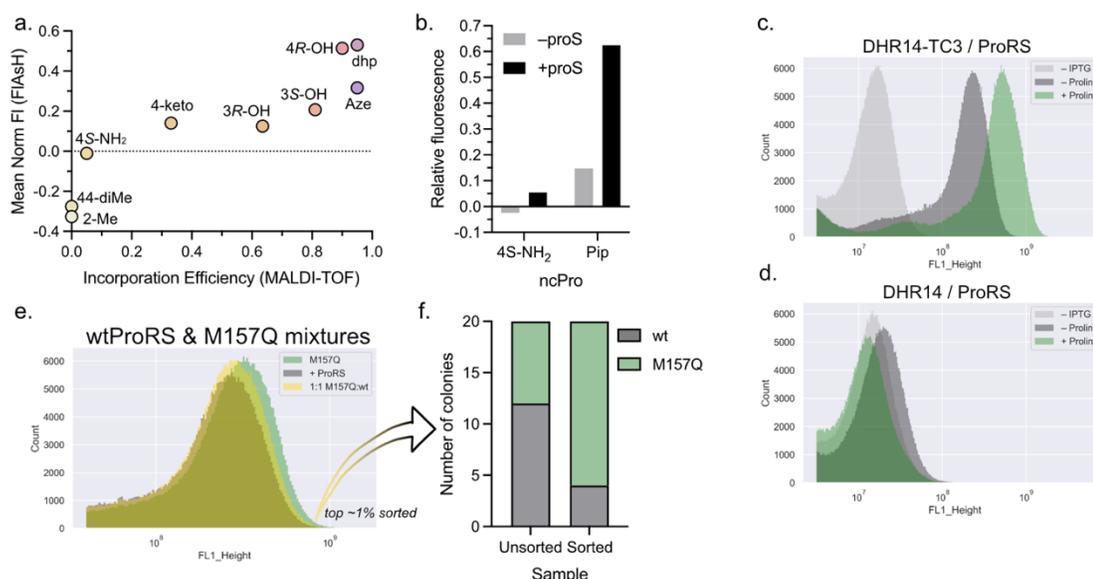
Figure 5.17. Validation of the DHR14 approach. **a**. Fluorescence, as determined by flow cytometry after DHR14-TC3 expression and FlAsH labeling, as a function of incorporation efficiency, as determined by MALDI-TOF after proinsulin expression. Flow cytometry histograms are shown in Figure 5.S4. **b**. ProRS overexpression increases fluorescence of ncPro-treated samples relative to –Proline and +Proline controls. Flow cytometry histograms are shown in Figure 5.S5. **c-d**. +Proline, –Proline, and –IPTG controls for DHR14 constructs translationally fused to TC3 (c), or lacking a TC tag (d). **e-f**. Strains overexpressing the wild-type ProRS, or mutant M157Q (which is known to increase incorporation of 4S-NH$_2$, ref. 19) were mixed 1:1 after DHR14-TC3 expression and FlAsH labeling. The brightest ~1% of cells were sorted (e). Plasmid DNA was purified and used to transform fresh stocks of strain CAG18515. Individual colonies were cultured, and plasmid DNA was sequenced to assess the presence of the M157Q mutation (f).

Nevertheless, we were hopeful that reduced toxicity associated with DHR14 might facilitate sorting by reducing the selective pressure against bright cells. We were encouraged by the observation that fluorescence correlated well with known ncPro incorporation efficiencies (R^2 = 0.834; Figure 5.17a, Figure 5.S4). Further, we observed decreased fluorescence for ncPro samples in the absence of ProRS overexpression (Figure 5.17b, Figure 5.S5), and no fluorescence over background was detected in the absence of the TC3 tag (Figure 5.17c-d). In an effort to reduce differences in physiology and rescue

efficiency between library members, we chose to lyse cells after sorting and transform fresh CAG18515 stocks using the purified DNA[45,46] before the next rounds of sorting and analysis. Encouragingly, we could enrich the ProRS mutant M157Q after 4S-NH$_2$ treatment from a 1:1 mixture with the wild-type ProRS using this protocol (Figure 5.17e-f). We constructed a separate site-saturation mutagenesis library, this time only targeting four locations (excluding position 202; 4SSM) to reduce library size and cloning burden. Separately, we cloned an error-prone PCR library (ePCR), targeting an average of 2 mutations per *proS* gene (Table 5.S1).

Despite these changes, we still observed significant increases in sample heterogeneity after sorting across the three ncPro analogs and two libraries assessed (Figure 5.18a-b, Figure 5.S6). This darker population retained resistance to ampicillin (the plasmid selection marker; Figure 5.18c), so were not the result of non-transformed cheaters. Sequencing revealed small plasmid products that were perhaps the result of DNA recombination. These undesired plasmids lacked both the DHR14-TC3 and *proS* genes. We noted the loss of the *proS* gene in the Top7-TC3 case described earlier; presence of the Top7-TC3 gene was not assessed. Though we did not perform whole-plasmid sequencing after this earlier sorting attempt, it seems likely that similar recombination events are occurring in both cases. Culturing library strains in the absence of FlAsH led to increasing plasmid heterogeneity over multiple transformations (Figure 5.18d), suggesting that the issue is not inherent to fluorophore labeling. At this stage, because of these pervasive issues with ProRS engineering, we elected to limit our efforts in insulin modification to those ncPro residues known to incorporate well with existing methods (see Chapters II & III of this thesis).
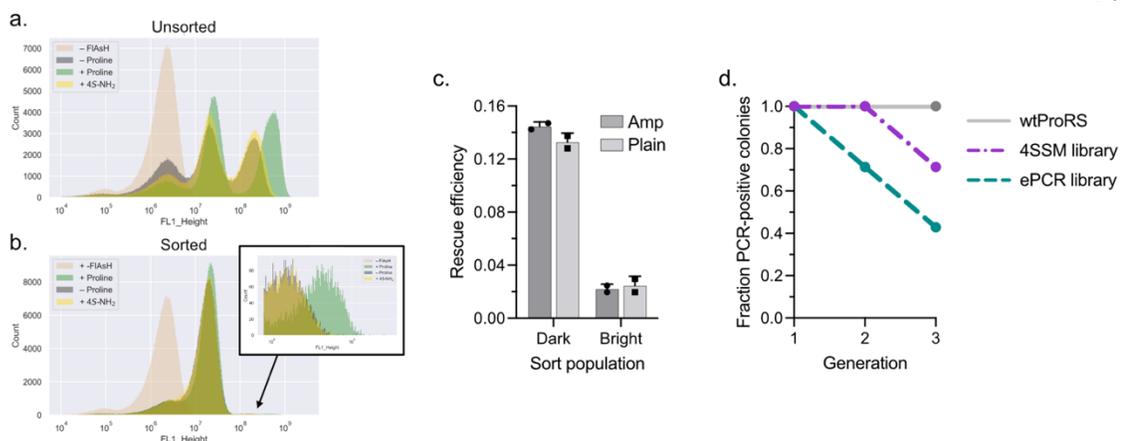
Figure 5.18. Significant problems persist. **a-b**. Flow cytometry histograms of a culture harboring a ProRS ePCR library before (**a**) and after (**b**) sorting for 4S-NH$_2$ incorporation. **c**. Rescue efficiencies of "Dark" ($\sim 10^7 <$ FL1 $< \sim 10^8$) and "Bright" (FL1 $> \sim 10^8$) 4SSM, +Proline library populations. Sorted cells were plated onto nonselective ("Plain") or selective ("Amp") LB-agar plates, and rescue efficiency determined after overnight growth at 37°C. **d**. Loss of the *proS* gene over multiple generations (transformation, antibiotic selection, overnight growth, and plasmid purification). The presence of the plasmid-borne *proS* gene was assessed by colony PCR in each generation.

5.4.11 Considerations for future ProRS engineering

The most pressing issue with the FlAsH-based labeling described here is the fact that undesired plasmid species are out-competing legitimate library members. This occurs independent of recovery method after sorting (outgrowth as in Figure 5.15b-c, or plasmid purification and transformation, Figure 5.18a-b), and was found with five separate library preparations cloned by both restriction enzyme and Gibson Assembly approaches. The presence of this dark population is observed by flow cytometry across the proline-free target proteins we screened (Figure 5.16i-l), and mutagenesis seems to be occurring even in the absence of FlAsH labeling (Figure 5.18d). The exact relationship between cellular morphology (Figure 5.16n-p) and the presence of undesired plasmids is not yet clear.

The significant depletion of the desired plasmid during our sorting experiments is perplexing: we have sorted the brightest cells, yet it is this dark population that is instead consistently enriched after sorting. Mutagenesis might be occurring at a basal level throughout these experiments, in which case undesired plasmid species might continually be generated. FACS sorting is likely imperfect, which would also enable the persistence of undesired plasmids. To exacerbate the problem, strains carrying the smaller plasmids tend to grow faster, and undesired plasmids seem to lead to higher transformation efficiencies. We also note that the pQE-80L vector used here is a high copy plasmid. If DNA recombination occurs *after* target protein expression and FlAsH labeling, a fraction of plasmids in a given bright cell might actually be undesired plasmid products, leading to those plasmids' persistence after sorting.

We note that our mock sorting experiment was successful in enriching a better ProRS variant after ncPro treatment (Figure 5.17e-f), and undesired plasmid species were not observed. The success of this test case compared to the failure in library sorting is puzzling. In the mock sorting case, starter cultures were grown from single colonies, whereas library cultures are necessarily heterogeneous. We did not observe plasmid instability across generations in the clonal sample (Figure 5.18d), though the relatively small number of samples and generations tested may have prevented detection if levels are initially very low. It seems likely that our library generation protocol may exacerbate the plasmid instability issue, despite the extensive quality control measures and multiple cloning approaches taken.

The strain CAG18515 may be the culprit, especially considering that mutagenesis occurred in the absence of FlAsH labeling (Figure 5.18d); assessing plasmid stability in other proline

auxotrophs under these conditions might help to address the issue. Extended exposure to the organoarsenic compound FlAsH, or competition molecule EDT, may also contribute to plasmid instability. For instance, inorganic arsenic is known to disrupt DNA repair pathways,[47] and could conceivably be formed as a result of FlAsH metabolism in *E. coli*. FlAsH likely binds to other off-target intracellular molecules (i.e., endogenous cysteine residues), which may result in unintended consequences. EDT could similarly interfere with endogenous intracellular components: for instance, reaction of EDT with carbonyl functionalities can yield 1,3-dithiolanes, and EDT is known to form metal complexes.

Our first attempts to implement this screening approach included a short FlAsH labeling period after target protein expression. This protocol led to poor and very heterogeneous labeling (Figure 5.S3), a problem that was resolved by adding FlAsH and EDT to cultures before inducing target protein expression (Figure 5.14f). While improving flow cytometry results from an analytical perspective, this approach led to extended incubation of cells with FlAsH and EDT, which may be impacting cellular physiology and library sorting. There may be alternative labeling protocols that both limit exposure to FlAsH and EDT, and result in good protein labeling.

Driving ProRS expression with the endogenous promoter might also hinder assay performance. We found that mutated plasmids generally lacked the *proS* gene, and its absence results in faster growth (Figure 5.S8a). Strains overexpressing different ProRS variants grow at different rates in M9 medium (Figure 5.S8a), which contributes an additional variable to the evolution experiment. Growing cells in rich medium (Figure 5.S8b) and controlling ProRS expression with the *araBAD* promoter (Figure 5.S8c) reduces growth differences. While we do not believe that this is the most prevalent problem

(increasing plasmid stability should be the immediate focus), we would recommend that future work drive *proS* expression with an inducible promoter.

One workaround to the presence of undesired plasmid species may be to PCR-amplify the *proS* gene from sorted cells, and use this linearized PCR product to install the sorted ProRS variants into new pQE80_DHR14-TC3_proS vectors. A drawback to this solution is the potential accumulation of PCR biases[48] which may skew representation in a sorted library.

An alternative screening modality to FACS that avoids culturing library members together might reduce the effect of differential growth. For instance, plate reader-based assays, while lower-throughput compared to FACS, would individually culture library members in separate wells. In this case, we might be able to more easily identify high performers, even if they grow more slowly. Finally, revisiting the split-GFP system may prove more fruitful than continuing to pursue the FlAsH approach. Sample heterogeneity was present by flow cytometry with this approach when expressing GFP1-10/11 in strain SLB2001 (Fig. 5.3), and this dark population was not analyzed. However, unlike performing FACS with FlAsH labeling, heterogeneity was not enriched after sorting.

5.5 Conclusion

In conclusion, we have described our efforts to develop a robust and general method to screen for ncPro incorporation, and to implement it in a ProRS engineering endeavor. From an analytical perspective, our FlAsH labeling approach seemed the most promising of all methods tested: we were able to achieve good correlation between known incorporation efficiencies as measured by mass spectrometry, and fluorescence as measured by flow cytometry (Figure 5.17a). Ultimately our attempts to apply this approach to ProRS

evolution were hindered by out-competition with undesired plasmid products. Together, the experiments described in this chapter lay the groundwork for future ProRS engineering endeavors. Though all approaches detailed here have yet to lead to improved ncPro incorporation, we are hopeful that future attempts at ProRS engineering can leverage this information to more successfully design and implement ProRS screening methods.

5.6 Materials and methods

5.6.1 Chemicals

All chemicals were purchased from MilliporeSigma unless otherwise indicated. FlAsH-EDT$_2$ was purchased from Santa Cruz Biotechnology, Carbosynth, and Toronto Research Chemicals. 1,2-ethanedithiol (EDT) was purchased from Fisher Scientific.

4-methyleneproline (4ene) was purchased as the N-boc protected version from Acros Organics, and deprotected with trifluoroacetic acid (TFA) in dichloromethane. 4ene was extracted with H$_2$O and lyophilized; complete deprotection and >95% purity was verified by ^{1}H NMR. All other proline analogs were used as received: 4S-aminoproline (4S-NH$_2$) was purchased as the dihydrochloride salt from Toronto Research Chemicals. 4R-fluoroproline (4R-F) and 4S-fluoroproline (4S-F) were purchased from Bachem. 2-methylproline (2-Me) was purchased as the hydrochloride salt from Advanced ChemBlocks. 4,4-dimethylproline (44-diMe) was purchased from J&W Pharmlab. 4-oxoproline (4-keto) was purchased as the hydrobromide salt from Sigma Aldrich. 3R-hydroxyproline (3R-OH) and 4R-hydroxyproline (4R-OH) were purchased from Sigma Aldrich. 3S-hydroxyproline (3S-OH) was purchased from Ark Pharm. 3,4-dehydroproline

(dhp) was purchased from Combi-Blocks. Azetidine-2-carboxylic acid (Aze) was purchased from Bachem.

5.6.2 Enzymes

Restriction enzymes, kinases, and ligases were purchased from New England Biolabs (NEB). 2x Q5 master mix (NEB) was used for all PCR applications, unless otherwise indicated. DNA phosphorylation was performed with T4 PNK, and ligation with T4 ligase. Gibson assembly was performed with the Repliqa HiFi assembly mix from Quantabio, or HiFi DNA Assembly Mix (NEB).

5.6.3 Strains

E. coil strains CAG18515, JW0232, JW0233, and JW0377 were obtained from the Coli Genetic Stock Center (CGSC) at Yale University.

Proline auxotrophs generated in this work include strains SLB2001 and SLB2160. Lambda red recombination[49] was used to knock out *proA* and *proB* in DH10B (SLB2001) or BL21(DE3) (SLB2160). Briefly, pKD4 was PCR-amplified with primers delPro_fwd & delPro_rev and used to transform parent strains harboring plasmid pKD46 and expressing the recombination machinery (induced with 10 mM L-arabinose). We were unable to transform BL21 with plasmid pKD46 ourselves, so obtained strain BL21(DE3)/pKD46 from the CGSC. Transformants were plated on selective (kanamycin) LB-agar plates, and gene knockout was verified by colony PCR. pKD46 was cured growing cells at 37°C; strain SLB2160 unexpectedly retained ampicillin resistance. Proline auxotrophy of each strain was verified by comparing growth in the presence and absence of proline.

Strain DH10B was used for standard cloning operations. Electrocompetent DH5α or 10-

beta stocks were obtained from NEB, and were used for library generation.

5.6.4 DNA oligos

DNA oligos were purchased from Integrated DNA Technologies (IDT).

delPro_fwd: TAAAACGTTCGTTTGATATCATTTTTCCTAAAATTGAGTGTAGGCTGGAGCTGCTTC

delPro_rev: GTCAATGGCCTTGTGAATCAAATGGCTACTTTTGCATATGGGAATTAGCCATGGTCC

proS_NheI_fwd: CTAGCTAGCTCAGCCTTTAATCTGTTTCACC

proS_SacII_rev: CGCGTGTGGTAGCCGCGGCGATTGAGC

proS_SacII_fwd: GCTCAATCGCCGCGGCTACCACACGCG

proS_KpnI_rev: CCAGCTGGGTACCAAGTACTCCG

proS_PstI_fwd: AAACTGCAGGGATTCCTGAGAAGTATGGAAAG

proS_NotI_rev: GACGAAGTGCGGCCGCGTTTCG

proS_NotI_fwd: CGAAACGCGGCCGCACTTCGTC

proS_SacI_rev: CCTGATTCGTAACGAGCTCAGCTCTTAC

del_11E11_fwd: TCGAGGGACTCCTGTTGATAGATCCAGTAATGACCTCAGAACTCCG

del_11E11_rev: CTAGCGGAGTTCTGAGGTCATTACTGGATCTATCAACAGGAGTCCC

11E11_KpnI_fwd: CGGGGTACCCTATGAGAGGATCGCATCACCATCACC

11E11_PstI_rev: AACTGCAGGCTTAATGCATAGAGCCTGAACCTGTG

11E11_RBS_fwd: CCCATTAAAGAGGAGAAATTAACTATGAGAGGATCGCATCACCATCACC

11E11_RBS_rev: TTAATTTCTCCTCTTTAATGGGGGTACCGAGCTCGAATTCGCTAGC

AL01004_fwd: ATACCGTAGCCACCCATCGTCAGG

AL01004_rev: CGGGGTAACGCGTGTGGT

AL01005_fwd: AGCATCTTTCTGCAGGAATTCGC

AL01005_rev2: TACTCTTTCCATACTTCTCAGGAATCC

SLB2110_fwd: GATTACAGGATCCTCTATGAAAATCGAAGAAGGTAAACTGGTAATCTGG

SLB2110_rev: GTATGGATCCCGAGCCAGAGCCAGAGCCAGTCTGCGCGTCTTTCAGGGC

SLB2111_fwd: GATTACACGGATCCGCTAGCAAAGGTGAAGAGCTGTTTGAC

SLB2111_rev: TACGGGGTACCACTAGTTTAGTCTTCGTTCGGATCTTTGCTCAGG

SLB2115_ins_fwd: CACCATCACCATCACGGATCCGCTAGCAAAGGTGAACGTCTG

SLB2115-b_ins_rev: CCAAAACAGCCAAGCTTGGTACCTTATTTTTCTTTCGGATCTTTGCTCAGTTTAGAACG

SLB2121_rev: CCAAAACAGCCAAGCTTGGTACCAC

SLB2121-b_rev: GCCAAAACAGCCAAGCTTGGTAC

SLB2122_fwd: GATTACAGCATGCATGTCCAAAGGAGAAGAACTGTTTACCGG

SLB2128_fwd: GATTACAGCATGCATGAAAATCGAAGAAGGTAAACTGGTAATCTGG

SLB2139_ins_fwd: CATCACCATCACCATCACGGATCCGGCTCCCCGGCTGGC

SLB2139_ins_rev: GCCCGGCCGACTAGTACCCGGCGCGCTACCCTCAG

SLB2139_vec_fwd: CTGAGGGTAGCGCGCCGGGTACTAGTCGGCCGGGC

SLB2139_vec_rev: GCCAGCCGGGGAGCCGGATCCGTGATGGTGATGGTGATG

SLB2149_ins_fwd: GGTGAGAATCCAAGCTAGCTTGGCGGACTCCTGTTGATAGATCCAGTAATG

SLB2149_ins_rev: GCTTCCTTAGCTCCTGAAAATCTCGGATTCACGCCCTTCTCTTTTG

SLB2149_vec_fwd: CAAAAGAGAAGGGCGTGAATCCGAGATTTTCAGGAGCTAAGGAAGC

SLB2149_vec_rev: CATTACTGGATCTATCAACAGGAGTCCGCCAAGCTAGCTTGGATTCTCACC

SLB2177_fwd: GGTACTAGTCGGCCGGGCTC

SLB2177_rev: GAGCCCGGCCGACTAGTACCTGGCGCGGAACCCTCG

SLB2192_fwd: ATGAGAGGATCGCATCACCATCACCATCACGGATCCAAACAAAAAAACGGTATCAAAGCG

SLB2192_rev: TAATTTATGGCCTAAGATGTTGCC

SLB2192_LOO7_vec_fwd2: GGCAACATCTTAGGCCATAAATTATAAGGTACCAAGCTTAATTAGCTG

SLB2192_LOO7_vec_rev: GGTGATGCGATCCTCTC

SLB2193_fwd: ATGAGAGGATCGCATCACCATCACCATCACGGATCCGAGGAT GGATCAGTGCAAC

SLB2193_rev: TTTTTGTTTATCGGCTGTGATG

SLB2193_vec_fwd: TCACAGCCGATAAACAAAAATAAGGTACCAAGCTTAATTA GC

SLB2214_ins_fwd: CACACAGAATTCATTAAAGAGGAGAAATTAACTATGGGGG ATATCCAGGTTCAAGTG

SLB2214_ins_rev: AAATGTTTTACAACACTCACGACAACACCAACGATGGGGCT CGAGAGAACCGCCCTC

SLB2214_vec_fwd: ATCGTTGGTGTTGTCGTGAGTGTTGTAAAACATTTTAAGCT TAATTAGCTGAGCTTGGAC

SLB2214_vec_rev: GTTCACTTGAACCTGGATATCCCCCATAGTTAATTTCTCCTC TTTAATGAATTCTGTGTG

SLB2215_ins_rev: GGGCTCGAGAGAACCGCCCTC

SLB2215_vec_fwd: GTGGAAGGACAGTTAGAGGGCGGTTCTCTCGAGCCCDDNT GTTGTDDNDDNTGTTGTDDNTAAGCTTAATTAGCTGAGCTTGGAC

SLB2227_fwd: TATTGTTGTGGGGTTTGTTGTATCTAAGC

SLB2228_3A_rev: AGAACCTCCGGATCCACCTGGACTACCACCTGGGGATCCAC CTGGCTCGAGAGAACCGCC

SLB2243_ins_fwd2: CAATCTAAAGTATATATGAGTAAACTTGGTCTGACAGCGT TTAAGGGCACCAATAACTGC

SLB2243_ins_rev: GGTGGCACTTTTCGGGGAAATGTGCCAACTTTTGGCGAAAA TGAGACGTTG

SLB2243_vec_fwd: CAACGTCTCATTTTCGCCAAAAGTTGGCACATTTCCCCGAA AAGTGC

SLB2243_vec_rev2: GCAGTTATTGGTGCCCTTAAACGCTGTCAGACCAAGTTTA CTCATATATACTTTAGATTG

SLB2322-GA_D1_fwd: GTACGCATGCTGTTACGGTTCCAGTTG

SLB2327_vec_fwd: GGAGGTTCTCCAGGTGGATCCCCAGG

SLB2327_vec_rev: TGATGGTGATGGTGATGCGATCCTCTCATAGTTAATTTCTC CTCTTTAATGAATTCTGTG

SLB2335_GA-fwd: CAACGATCAAGGCGAGTTACATGATCCC

SLB2335_GA-rev: GCACAACATGGGGGATCATGTAACTCG

SLB2363_A-b_rev: TAAGCTTAATTAGCTGAGCTTGGAC

SLB2363_B-b_rev: AGAACCTCCGGATCCAC

SLB2371_ins_fwd: GTTGGAAGGTAGAACAGGCTCCCCAAAAAAACGGGTATGG

SLB2371_ins_rev: GCGTATAATATTTGCCCATGGGGTGATGCCGGCC

SLB2371_vec_fwd: CATCGTGGCCGGCATCACCCCATGGGCAAATATTATACGC

SLB2371_vec_rev: CCATACCCGTTTTTTTGGGGAGCCTGTTCTACCTTCC

SLB2406_fwd: GCTTAATTAGCTGAGCTTGGACTCCTG

SLB2406_rev: GCCTTGAGCCTGTTCTACCTTCC

SLB2429_analysis_fwd: GTATCACGAGGCCCTTTCG

SLB2429_analysis_rev: GCTTAGATACAACAAACCCCACAAC

SLB2436_4SSM-1_fwd: GAGAGGCGCTGCCGCCMNNAGAACCGGTGTCGGCTTG

SLB2436_4SSM-1_rev: CATGCGTTCCCGCGAATTCCTGNNKAAANNKGCTNNKTCTTTCCATACTTCTCAGGAATCCC

SLB2436_4SSM-2_fwd: CACGCGTTACCCCGATACCGTAMNNACCCATCGTCAGGATTTGGTTACGG

SLB2436_4SSM-2_rev: GGCGGCAGCGCCTCTC

SLB2436_GA_vec_rev: CGGTATCGGGGTAACGCGTG

SLB2436_GA_vec_fwd: CAGGAATTCGCGGGAACGCATG

proS_SSM-lib1_fwd: GCCAGCACCTGGAATTCGTGAGAGGCGCTGCCGCCMNNAGAACCGGTGTCGGCTTG

proS_SSM-lib1_rev: AAGGAAAAAAGCGGCCGCGTTTCGGCGTCATGCGTTCCCGCGAATTCCTGNNKAAANNKGCTNNKTCTTTCCATACTTCTCAGGAATCCC

proS_SSM-lib2_fwd: GATTATCCCCGCGGCTACCACACGCGTTACCCCGATACCGTAMNNACCCATCGTCAGGATTTGGTTACGG

proS_SSM-lib2_rev: CTCTCACGAATTCCAGGTGCTGGCGCAGAGCGGTGAAGACGATGTGG

proS_ePCR_fwd3: GGTGAGAATCCAAGCTAGCTCAGCC

proS_ePCR_rev3: CAACTGGAACCGTAACAGCATGCGTAC

proS_GA_vec_rev: GGCTGAGCTAGCTTGGATTCTCACC

5.6.5 g-Block gene fragments

g-Block gene fragments were purchased from Integrated DNA Technologies (IDT).

<u>LOO#</u>:

ATGGCATCAAAGGGAGAAGAGTTGTTTACAGGTGTAGTGCCCATCTTGGTTG
AATTGGACGGCGATGTCAACGGACACAAATTCTCTGTACGCGGAGAGGGGGA

GGGGGACGCTACCATCGGTAAACTTACGCTTAAATTTATTTGTACAACGGGA
AAATTGCCCGTCCCCTGGCCCACATTGGTAACAACCTTGACATATGGTGTGCA
GTGTTTTAGCCGTTACCCAGATCATATGAAACGTCACGATTTCTTTAAGAGCG
CGATGCCCGAGGGTTACGTTCAGGAGCGTACAATCTCGTTTAAGGATGACGG
GAAGTATAAGACCCGCGCAGTCGTAAAGTTCGAGGGGGACACACTTGTCAAT
CGCATTGAATTGAAGGGAACTGACTTTAAAGAGGATGGCAACATCTTAGGCC
ATAAATTAGAGTACAACTTTAACTCACATAACGTCTACATCACAGCCGATAA
ACAAAAAAACGGTATCAAAGCGAACTTTACAGTACGCCACAACGTTGAGGAT
GGATCAGTGCAACTGGCGGATCATTATCAGCAGAACACGCCAATCGGGGACG
GGCCGGTCCTTTTGCCCGACAACCACTATTTAAGCACTCAAACGGTATTGTCA
AAAGATCCTAATGAAAACGCGATCACATGGTACTGCTGGAATTTGTCACTG
CCGCCGGTATTACGCATGGTATGGACGAGCTGTACAAAGGGGGTACTGGTGG
TAGT

E6-GFP7:

ATATGCGGGGTACCCCCATTAAAGAGGAGAAATTAACTATGCGCGGCTCTCA
CCACCACCATCATCATGGCTCTGTGCCGGGTGCGGGCGTTCCGGGAGCTGGT
GTCCCAGGCGAAGGAGTTCCGGGTGCGGGCGTGCCCGGAGCTGGTTCTGGTA
GTGGATCAGGGTCTGAATATAACTTCAACAGTCATAATGTGTATATCACTGCA
GACGGTTCAGGATCGGGGTCCGAATACAATTTCAATTCCCATAATGTCTACAT
CACAGCTGACTGAGCCTGCAGGCATGCAAGCTTGATTACA

E6-GFP8:

ATATGCGGGGTACCCCCATTAAAGAGGAGAAATTAACTATGCGTGGTAGTCA
CCACCATCATCATCATGGTTCAGTTCCAGGTGCGGGTGTACCCGGCGCAGGT
GTTCCAGGCGAAGGTGTGCCGGGGGCGGGGGTTCCCGGCGCGGGCAGTGGTT
CCGGTTCCGGTTCAAATGGGATTAAAGCTAATTTTACAGTACGCCACAATGTC
GGGAGTGGTAGCGGCTCGAACGGCATTAAGGCAAATTTTACAGTCCGTCATA
ATGTCTAAGCCTGCAGGCATGCAAGCTTGATTACA

DHR14:

ATGAGAGGATCGCATCACCATCACCATCACATGGATAGCGAGGAAGTGAATG
AACGTGTAAAACAACTGGCAGAGAAAGCAAAGGAAGCAACCGATAAAGAAG
AGGTGATTGAAATTGTAAAAGAGTTAGCCGAATTGGCCAAGCAGAGTACTGA
CTCCGAATTAGTCAACGAGATTGTGAAGCAATTAGCAGAAGTGGCGAAGGAG
GCGACAGATAAGGAGTTGGTAATTTACATCGTGAAGATCTTGGCTGAATTAG
CGAAGCAAAGCACCGACAGCGAACTGGTAAACGAGATTGTAAAACAGCTTG
CGGAGGTGGCTAAAGAGGCAACGGACAAAGAGTTAGTGATTTATATCGTAAA
GATCCTGGCTGAGTTAGCCAAGCAATCCACTGATTCAGAGTTGGTTAACGAA
ATTGTAAAGCAGCTTGAAGAGGTCGCCAAAGAGGCTACTGACAAAGAACTG
GTAGAACACATTGAAAAGATCCTGGAGGAGCTTAAAAAGCAGTCCACAGAC
GGTTGGCTGGAGGGAGGTTCTCCAGGTGGATCCCCA

5.6.6 Plasmids and cloning

pBAD33_mWasabi: the mWasabi gene was amplified from plasmid pKPY515 (Ref. 50) and installed in the pBAD33 backbone using KpnI and HindIII cut sites.

pQE80PI-proS was previously described.[19]

pQE80_11-E-11_proS: To facilitate site-saturation mutagenesis library creation, cut sites NotI and SacII were installed into the *proS* gene of pQE80β11(Eβ11)-proS.[19] These cut sites flank the library residues of interest, but introduce silent mutations into the *proS* coding sequence. We used an overlap extension PCR approach: PCR fragment 1 (amplified with primers proS_NheI_fwd & proS_SacII_rev) was stitched together with PCR fragment 2 (proS_SacII_fwd & proS_KpnI_rev), then ligated into the vector backbone using a restriction enzyme cloning approach (NheI and KpnI cut sites) to install the SacII cut site. PCR fragment 3 (proS_PstI_fwd & proS_NotI_rev) was stitched together with PCR fragment 4 (proS_NotI_fwd & proS_SacI_rev) and ligated into the digested vector backbone (PstI & SacI cut sites) to install the NotI cut site.

ptac_GFP(1-10) was previously described.[19]

pQE80_proS: The 11-E-11 gene was excised from the plasmid pQE80_11-E-11_proS using cut sites XhoI and EcoRI; this excised sequence was replaced with a short DNA fragment resulting from annealing the oligos del_11E11_fwd and del_11E11_rev.

pBAD33_11-E-11: The 11-E-11 gene was amplified from plasmid pQE80_11-E-11_proS using primers 11E11_KpnI_fwd & 11E11_PstI_rev and installed into the pBAD33 backbone using restriction enzyme cloning (KpnI & PstI cut sites). The RBS was then installed with a one-piece Gibson Assembly approach with the DNA fragment resulting

from PCR amplification of the vector backbone using primers 11E11_RBS_fwd & 11E11_RBS_rev.

pQE80_GFP1-10: GFP1-10 was amplified from ptac_GFP(1-10) with primers SLB2122_fwd & SLB2121_rev and installed into the pQE-80L backbone using a restriction enzyme approach (HindIII and SphI cut sites).

pQE80_MBP-GFP1-10: The MBP gene was first installed at the N-terminus of GFP1-10 in vector ptac_GFP(1-10): MBP was amplified from plasmid pET His6 MBP TEV LIC (a gift from Scott Gradia, Addgene plasmid #29656) with primers SLB2110_fwd & SLB2110_rev, then installed at the BamHI cut site in ptac_GFP(1-10). The MBP-GFP1-10 fusion protein gene was transferred to the pQE-80L backbone using the HindIII and SphI cut sites after amplification with SLB2128_fwd & SLB2121-b_rev. A similar approach was used to install NusA, TrxA, and SUMO at the N-terminus of GFP1-10.

pQE80_MBP-GFP1-10_proS: We used a Gibson Assembly approach to install the *proS* gene into the pQE80_MBP-GFP1-10 backbone, with primers SLB2149_ins_fwd & SLB2149_ins_rev (*proS* amplification), and SLB2149_vec_fwd & SLB2149_vec_rev (backbone amplification).

pBAD33_X144-GFP11 was cloned using a Gibson Assembly approach. The XTEN144 gene was amplified from pX_X144 (a gift from Peter B. Rapp) with primers SLB2139_ins_fwd & SLB2139_ins_rev, and replaced the N-terminal GFP11-elastin fragment in the pBAD33_11-E-11 vector (primers SLB2139_vec_fwd & SLB2139_vec_rev).

pBAD33_X72-GFP11: The C-terminal half of the XTEN144 gene in pBAD33_X144-GFP11 was removed with a one-piece Gibson Assembly approach, after amplification with primers SLB2177_fwd & SLB2177_rev.

ptac_sn1-10: β-strands 1-10 from super-negative GFP[32] were amplified from pET-6xHis-(-30)GFP (a gift from David Liu, Addgene plasmid #62936) with primers SLB2111_fwd & SLB2111_rev. The amplicon was installed in place of GFP1-10 in vector ptac_GFP(1-10) using restriction enzyme cloning (cut sites BamHI & KpnI).

ptac_sp1-10: β-strands 1-10 from super-positive GFP[32] were amplified from pET-6xHis-(pos36)GFP (a gift from David Liu, Addgene plasmid #62937) with primers SLB2115_ins_fwd & SLB2115-b_ins_rev. The amplicon was installed in place of GFP1-10 in vector ptac_GFP(1-10) using restriction enzyme cloning (cut sites BamHI & KpnI).

pQE80_LOO7_proS: A g-Block gene fragment (LOO#) corresponding to the sequence of the circularly permuted GFP was circularized by ligation. It was amplified by primers SLB2192_fwd & SLB2192_rev to give the large LOO7 GFP fragment. This GFP fragment replaced the MBP-GFP1-10 gene in vector pQE80_MBP-GFP1-10 by Gibson Assembly: the backbone fragment was amplified with primers SLB2192_LOO7_vec_fwd2 & SLB2192_LOO7_vec_rev.

pQE80_LOO8_proS: A g-Block gene fragment (LOO#) corresponding to the sequence of the circularly permuted GFP was circularized by ligation. It was amplified by primers SLB2193_fwd & SLB2193_rev to give the large LOO8 GFP fragment. This GFP fragment replaced the MBP-GFP1-10 gene in vector pQE80_MBP-GFP1-10 by Gibson Assembly:

the backbone fragment was amplified with primers SLB2193_vec_fwd & SLB2192_LOO7_vec_rev.

pBAD33_E-7: A g-Block gene fragment (E6-GFP7) was designed to contain two copies of the GFP7 strand downstream of a six-mer elastin fragment. It was installed in the pBAD33 vector using restriction enzyme cloning (KpnI and HindIII cut sites).

pBAD33_E-8: A g-Block gene fragment (E6-GFP8) was designed to contain two copies of the GFP8 strand downstream of a six-mer elastin fragment. It was installed in the pBAD33 vector using restriction enzyme cloning (KpnI and HindIII cut sites).

pQE80_Top7-P-TC_proS: The initial design of the Top7-TC gene included one proline codon between the Top7 gene and the TC motif HRWCCRECCKTF.[51] The Top7 gene was amplified with primers SLB2214_ins_fwd & SLB2214_ins_rev from plasmid Top7 K39E K40E V48V K55E, a gift from David Baker (Addgene plasmid # 12464). It replaced the 11-E-11 gene in pQE80_11-E-11_proS by Gibson Assembly (SLB2214_vec_fwd & SLB2214_vec_rev).

pQE80_Top7-TC3_proS: We installed a longer proline-containing linker between Top7 and the TC motif TC3 to increase screening stringency. The sorted TC library variant pQE80_Top7-P-TC3_proS was amplified with primers SLB2227_fwd & SLB2228_3A_rev. The linear PCR product was phosphorylated, then ligated via blunt end ligation.

pQE80-c_Top7-TC3_proS: A Gibson Assembly approach was used to replace the ampR cassette with that of CAT from pBAD33. pQE80_Top7-TC3_proS was amplified with SLB2243_vec_fwd & SLB2243_vec_rev2; the CAT gene from pBAD33 was amplified

with SLB2243_ins_fwd2 & SLB2243_ins_rev. This plasmid was used for Top7-TC3 expression in strain SLB2160, which retained ampicillin resistance after lambda red recombination.

pQE80_DHR14-TC3_proS: Gibson Assembly was used to replace the Top7 gene with DHR14. pQE80_Top7-TC3_proS was amplified with primers SLB2327_vec_fwd & SLB2327_vec_rev, and was assembled with the codon-optimized DHR14 g-Block. A similar approach was used to generate plasmids corresponding to the other proline-free genes screened for toxicity and FlAsH labeling.

pQE80_DHR14-TC3: The DHR14-TC3 gene (from plasmid pQE80_DHR14TC3_proS) was installed into the pQE-80L backbone by restriction enzyme cloning (XhoI & HindIII cut sites).

pQE80_DHR14-3P_proS: The TC3 motif was removed from the DHR14 gene by a blunt-end ligation approach after PCR amplification of pQE80_DHR14-TC3_proS with primers SLB2363_A-b_rev & SLB2363_B-b_rev.

pQE80_DHR14-TC3_proS-ara: We cloned an arabinose-inducible version of the *proS* gene through a three-part Gibson Assembly approach. The AraC fragment was amplified from pBAD33 with primers SLB2371_ins_fwd & SLB2371_ins_rev. pQE80_DHR14-TC3_proS was amplified as two fragments using the following primer sets: SLB2371_vec_fwd / AmpR-GA_rev and AmpR-GA_fwd / SLB2371_vec_rev.

M157Q & C443G point mutations: A blunt-end ligation approach was used to install the desired point mutations into the *proS* gene. Template plasmids were amplified with

AL01005_fwd & AL01005_rev2 (M157Q), or AL01004_fwd & AL01004_rev (C443G). The linear PCR product was phosphorylated, then ligated via blunt end ligation.

5.6.7 Library generation

<u>TC motif library</u>: The 8-mer randomized amino acid sequence XCCXXCCX was designed, in which DDN (D = G/A/T; N = A/T/G/C) codons for each X residue were used. We installed this TC motif library at the C-terminus of the Top7 gene using a Gibson Assembly approach: the Top7 gene was amplified with primers SLB2214_ins_fwd & SLB2215_ins_rev from plasmid Top7 K39E K40E V48V K55E, (a gift from David Baker, Addgene plasmid # 12464) and pQE80_11-E-11_proS was amplified with SLB2215_vec_fwd & SLB2214_vec_rev.

<u>5-site saturation *proS* mutagenesis</u>: The *proS* gene was amplified from plasmid pQE80_11-E-11_proS in two fragments. Insert 1 contains the NNK-randomized positions 157, 159, 161, and 202 (K = G/T; N = A/T/G/C), and was amplified with primers proS_SSM-lib1_fwd & proS_SSM-lib1_rev. Insert 2 contains randomization at position 443, and was amplified with primers proS_SSM-lib2_fwd & proS_SSM-lib2_rev. Insert 1 and insert 2 were stitched together using overlap extension PCR. The resulting PCR product was installed into vector pQE80_Top7-TC3_proS using a restriction enzyme approach (SacII & NotI cut sites). We obtained approximately $2x10^7$ transformants, and NNK randomization was verified by Sanger sequencing.

<u>4-site saturation *proS* mutagenesis</u>: We constructed our 4SSM library through two approaches. In the original restriction enzyme approach, the *proS* gene was amplified from plasmid pQE80_DHR14-TC3_proS with primers proS_SSM-lib1_fwd & proS_SSM-

lib2_rev, leading to NNK-randomization at positions 157, 159, 161, and 443. The resulting PCR product was installed in the pQE80_DHR14-TC3_proS backbone using restriction enzyme cloning (SacII and NotI cut sites). We obtained approximately 4×10^5 transformants.

In the second approach, we performed a four-part Gibson Assembly with fragments resulting from the PCR amplifications of pQE80_DHR14-TC3_proS using the following primer sets: (1) SLB2436_4SSM-1_fwd & SLB2436_4SSM-1_rev; (2) SLB2436_4SSM-2_fwd & SLB2436_4SSM-2_rev; (3) SLB2335_GA-fwd & SLB2436_GA_vec_rev; and (4) SLB2436_GA_vec_fwd & SLB2335_GA-rev. In an initial test used to measure plasmid stability, we obtained approximately 2×10^5 transformants.

Error-prone PCR: We constructed our ePCR library through two approaches. In the original restriction enzyme approach, the *proS* gene from plasmid pQE80_DHR14-TC3_proS was amplified with primers proS_ePCR_fwd3 & proS_ePCR_rev3 with the Taq DNA polymerase (NEB) and in the presence of 0-400 µM $MnCl_2$. The PCR product was installed into pQE80_DHR14-TC3_proS using restriction enzyme cloning (NheI & SphI cut sites) and an extended 16 hour ligation reaction at 16°C with T4 ligase. Increasing $MnCl_2$ concentrations yielded increasing *proS* mutation rates, as expected (Table 5.S1). We chose to sort the library generated under 200 µM $MnCl_2$ conditions (2.13 ± 1.87 mutations per *proS* gene), and obtained 1×10^6 transformants.

In the second approach, we performed a three-part Gibson Assembly. Fragments corresponding to the plasmid backbone were amplified with Q5 DNA polymerase and the following primer sets: (1) SLB2335_GA-fwd & ProS_GA_vec_rev; and (2) SLB2322-

GA_D1_fwd & SLB2335_GA-rev. The *proS* gene was amplified with Taq polymerase in the presence of 200 µM MnCl$_2$ with primers proS_ePCR_fwd3 & proS_ePCR_rev3. In an initial test used to measure plasmid stability, we obtained approximately 3x10^4 transformants.

5.6.8 Nucleotide and amino acid sequences

mWasabi:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCATGGTGAGCAAGGGCG
AGGAGACCACAATGGGCGTAATCAAGCCCGACATGAAGATCAAGCTGAAGA
TGGAGGGCAACGTGAATGGCCACGCCTTCGTGATCGAGGGCGAGGGCGAGG
GCAAGCCCTACGACGGCACCAACACCATCAACCTGGAGGTGAAGGAGGGAG
CCCCCCTGCCCTTCTCCTACGACATTCTGACCACCGCGTTCAGTTACGGCAAC
AGGGCCTTCACCAAGTACCCCGACGACATCCCCAACTACTTCAAGCAGTCCT
TCCCCGAGGGCTACTCTTGGGAGCGCACCATGACCTTCGAGGACAAGGGCAT
CGTGAAGGTGAAGTCCGACATCTCCATGGAGGAGGACTCCTTCATCTACGAG
ATACACCTCAAGGGCGAGAACTTCCCCCCCAACGGCCCCGTGATGCAGAAGG
AGACCACCGGCTGGGACGCCTCCACCGAGAGGATGTACGTGCGCGACGGCGT
GCTGAAGGGCGACGTCAAGATGAAGCTGCTGCTGGAGGGCGGCGGCCACCA
CCGCGTTGACTTCAAGACCATCTACAGGGCCAAGAAGGCGGTGAAGCTGCCC
GACTATCACTTTGTGGACCACCGCATCGAGATCCTGAACCACGACAAGGACT
ACAACAAGGTGACCGTTTACGAGATCGCCGTGGCCCGCAACTCCACCGACGG
CATGGACGAGCTGTACAAGTAA

MRGSHHHHHHGSMVSKGEETTMGVIKPDMKIKLKMEGNVNGHAFVIEGEGEG
KPYDGTNTINLEVKEGAPLPFSYDILTTAFSYGNRAFTKYPDDIPNYFKQSFPEGY
SWERTMTFEDKGIVKVKSDISMEEDSFIYEIHLKGENFPPNGPVMQKETTGWDAS
TERMYVRDGVLKGDVKMKLLLEGGGHHRVDFKTIYRAKKAVKLPDYHFVDHR
IEILNHDKDYNKVTVYEIAVARNSTDGMDELYK

GFP11-Elastin-GFP11:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCCGGCCGGGCTCTGGTT
CTGGTTCAGGCTCTCGTGACCATATGGTTCTTCATGAGTACGTAAATGCTGCT
GGCATCACAGGGTCGACTGTGCCGGGTGCGGGCGTTCCGGGAGCTGGTGTCC
CAGGCGAAGGAGTTCCGGGTGCGGGCGTGCCCGGAGCTGGTACTAGTCGGCC
GGGCTCTGGTTCTGGTTCAGGCTCTCGTGACCATATGGTTCTTCATGAGTACG
TAAATGCTGCTGGCATCACAGGTTCAGGCTCTATGCATTAA

MRGSHHHHHHGSRPGSGSGSGSRDHMVLHEYVNAAGITGSTVPGAGVPGAGVP
GEGVPGAGVPGAGTSRPGSGSGSGSRDHMVLHEYVNAAGITGSGSMH

GFP1-10:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCATGTCCAAAGGAGAAG
AACTGTTTACCGGCGTTGTGCCAATTTTGGTTGAACTCGATGGTGATGTCAAC
GGACATAAGTTCTCAGTGAGAGGCGAAGGAGAAGGTGACGCCACCATTGGA
AAATTGACTCTTAAATTCATCTGTACTACTGGTAAACTTCCTGTACCATGGCC
GACTCTCGTAACAACGCTTACGTACGGAGTTCAGTGCTTTTCGAGATACCCAG
ACCATATGAAAAGACATGACTTTTTTAAGTCGGCTATGCCTGAAGGTTACGTG
CAAGAAAGAACAATTTCGTTCAAAGATGATGGAAAATATAAAACTAGAGCA
GTTGTTAAATTTGAAGGAGATACTTTGGTTAACCGCATTGAACTGAAAGGAA
CAGATTTTAAAGAAGATGGTAATATTCTTGGACACAAACTCGAATACAATTTT
AATAGTCATAACGTATACATCACTGCTGATAAGCAAAAGAACGGAATTAAAG
CGAATTTCACAGTACGCCATAATGTAGAAGATGGCAGTGTTCAACTTGCCGA
CCATTACCAACAAAACACCCCTATTGGAGACGGTCCGGTACTTCTTCCTGATA
ATCACTACCTCTCAACACAAACAGTCCTGAGCAAAGATCCAAATGAAAAATA
A

MRGSHHHHHHGSMSKGEELFTGVVPILVELDGDVNGHKFSVRGEGEGDATIGK
LTLKFICTTGKLPVPWPTLVTTLTYGVQCFSRYPDHMKRHDFFKSAMPEGYVQE
RTISFKDDGKYKTRAVVKFEGDTLVNRIELKGTDFKEDGNILGHKLEYNFNSHN
VYITADKQKNGIKANFTVRHNVEDGSVQLADHYQQNTPIGDGPVLLPDNHYLST
QTVLSKDPNEK

XTEN144-GFP11:

ATGAGAGGATCGCATCACCATCACCATCACGGATCCGGCTCCCCGGCTGGCA
GCCCGACCAGCACTGAAGAGGGCACGAGCGAGTCGGCGACCCCGGAGTCTG
GTCCGGGCACCTCCACCGAACCGTCTGAGGGCAGCGCACCGGGTAGCCCGGC
CGGTAGCCCTACCAGCACCGAAGAGGGTACCAGCACGGAACCGAGCGAAGG
CTCGGCACCGGGTACGAGCACCGAGCCGTCCGAGGGTTCCGCGCCAGGTACC
AGCGAGAGCGCAACGCCGGAGTCCGGTCCGGGCAGCGAACCAGCGACCAGC
GGCAGCGAAACGCCGGGGTTCAGAGCCGGCGACGAGCGGTAGCGAGACTCCG
GGCAGCCCGGCTGGTAGCCCGACGTCCACCGAAGAAGGCACCAGCGAAAGC
GCCACCCCGGAGAGCGGTCCTGGTACGTCTACCGAGCCATCTGAGGGTAGCG
CGCCGGGTACTAGTCGGCCGGGCTCTGGTTCTGGTTCAGGCTCTCGTGACCAT
ATGGTTCTTCATGAGTACGTAAATGCTGCTGGCATCACAGGTTCAGGCTCTAT
GCATTAA

MRGSHHHHHHGSGSPAGSPTSTEEGTSESATPESGPGTSTEPSEGSAPGSPAGSPT
STEEGTSTEPSEGSAPGTSTEPSEGSAPGTSESATPESGPGSEPATSGSETPGSEPAT
SGSETPGSPAGSPTSTEEGTSESATPESGPGTSTEPSEGSAPGTSRPGSGSGSGSRD
HMVLHEYVNAAGITGSGSMH

Prolyl-tRNA synthetase (ProRS):

ATGCGTACTAGCCAATACCTGCTCTCCACTCTCAAGGAGACACCTGCCGACG
CCGAGGTGATCAGCCATCAGCTGATGCTGCGCGCCGGGATGATCCGCAAGCT
GGCCTCCGGGTTATATACCTGGCTGCCGACCGGCGTGCGCGTTCTGAAAAAA
GTCGAAAACATCGTGCGTGAAGAGATGAACAACGCCGGTGCGATCGAGGTGT
CGATGCCGGTGGTTCAGCCAGCCGATTTGTGGCAAGAGAGTGGTCGTTGGGA
ACAGTACGGTCCGGAACTGCTGCGTTTTGTTGACCGTGGCGAGCGTCCGTTCG
TACTCGGCCCAACTCATGAAGAAGTTATCACTGACCTGATTCGTAACGAGCTC
AGCTCTTACAAACAGCTGCCGCTGAACTTCTATCAGATCCAGACCAAGTTCC
GCGACGAAGTGCGGCCGCGTTTCGGCGTCATGCGTTCCCGCGAATTCCTGAT
GAAAGATGCTTACTCTTTCCATACTTCTCAGGAATCCCTGCAGGAAACCTACG
ATGCAATGTATGCGGCCTACAGCAAAATCTTCAGCCGCATGGGGCTGGATTT
CCGCGCCGTACAAGCCGACACCGGTTCTATCGGCGGCAGCGCCTCTCACGAA

TTCCAGGTGCTGGCGCAGAGCGGTGAAGACGATGTGGTCTTCTCCGACACCT
CTGACTATGCAGCGAACATTGAACTGGCAGAAGCTATCGCGCCGAAAGAACC
GCGCGCTGCTGCTACCCAGGAAATGACGCTGGTTGATACGCCGAACGCGAAA
ACCATCGCGGAACTGGTTGAACAGTTCAATCTGCCGATTGAGAAAACGGTTA
AGACTCTGCTGGTTAAAGCGGTTGAAGGCAGCAGCTTCCCGCAGGTTGCGCT
GCTGGTGCGCGGTGATCACGAGCTGAACGAAGTTAAAGCAGAAAAACTGCC
GCAGGTTGCAAGCCCGCTGACTTTCGCGACCGAAGAAGAAATTCGTGCCGTG
GTTAAAGCCGGTCCGGGTTCACTGGGTCCGGTAAACATGCCGATTCCGGTGG
TGATTGACCGTACCGTTGCGGCGATGAGTGATTTCGCTGCTGGTGCTAACATC
GATGGTAAACACTACTTCGGCATCAACTGGGATCGCGATGTCGCTACCCCGG
AAGTTGCAGATATCCGTAACGTGGTGGCTGGCGATCCAAGCCCGGATGGCCA
GGGTAGGCTGCTGATCAAACGTGGTATCGAAGTTGGTCACATCTTCCAGCTG
GGTACCAAGTACTCCGAAGCACTGAAAGCCTCCGTACAGGGTGAAGATGGCC
GTAACCAAATCCTGACGATGGGTTGCTACGGTATCGGGGTAACGCGTGTGGT
AGCCGCGGCGATTGAGCAGAACTACGACGAACGAGGCATCGTATGGCCTGA
CGCTATCGCGCCGTTCCAGGTGGCGATTCTGCCGATGAACATGCACAAATCC
TTCCGCGTACAAGAGCTTGCTGAGAAACTGTACAGCGAACTGCGTGCACAAG
GTATCGAAGTGCTGCTGGATGACCGCAAAGAGCGTCCGGGCGTGATGTTTGC
TGATATGGAACTGATCGGTATTCCGCACACTATTGTGCTGGGCGACCGTAACC
TCGACAACGACGATATCGAATATAAATATCGTCGCAACGGCGAGAAACAGTT
AATTAAGACTGGTGACATCGTCGAATATCTGGTGAAACAGATTAAAGGCTGA

MRTSQYLLSTLKETPADAEVISHQLMLRAGMIRKLASGLYTWLPTGVRVLKKVE
NIVREEMNNAGAIEVSMPVVQPADLWQESGRWEQYGPELLRFVDRGERPFVLG
PTHEEVITDLIRNELSSYKQLPLNFYQIQTKFRDEVRPRFGVMRSREFLMKDAYSF
HTSQESLQETYDAMYAAYSKIFSRMGLDFRAVQADTGSIGGSASHEFQVLAQSG
EDDVVFSDTSDYAANIELAEAIAPKEPRAAATQEMTLVDTPNAKTIAELVEQFNL
PIEKTVKTLLVKAVEGSSFPQVALLVRGDHELNEVKAEKLPQVASPLTFATEEEIR
AVVKAGPGSLGPVNMPIPVVIDRTVAAMSDFAAGANIDGKHYFGINWDRDVAT
PEVADIRNVVAGDPSPDGQGRLLIKRGIEVGHIFQLGTKYSEALKASVQGEDGRN
QILTMGCYGIGVTRVVAAAIEQNYDERGIVWPDAIAPFQVAILPMNMHKSFRVQ

ELAEKLYSELRAQGIEVLLDDRKERPGVMFADMELIGIPHTIVLGDRNLDNDDIE
YKYRRNGEKQLIKTGDIVEYLVKQIKG

Top7-TC3:

ATGGGGGATATCCAGGTTCAAGTGAACATTGACGATAATGGTAAAAACTTCG
ATTATACTTATACCGTGACCACAGAATCGGAATTACAGAAGGTTTTGAACGA
ACTGAAAGACTACATCGAAGAACAGGGGGCTAAACGCGCGCGCATCTCGATT
ACCGCTCGTACGGAAAAGAAGCAGAAAATTTGCCGCGATTTTGATTAAAG
TTTTCGCCGAACTGGGTTATAATGATATTAACGTTACGTGGGACGGCGATACG
GTGACAGTGGAAGGACAGTTAGAGGGCGGTTCTCTCGAGCCCCATCGTTGGT
GTTGTCGTGAGTGTTGTAAACATTTTAA

MGDIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKDYIEEQGAKRARISITAR
TEKEAEKFAAILIKVFAELGYNDINVTWDGDTVTVEGQLEGGSLEPHRWCCREC
CKTF

DHR14-TC3:

ATGAGAGGATCGCATCACCATCACCATCACATGGATAGCGAGGAAGTGAATG
AACGTGTAAAACAACTGGCAGAGAAAGCAAAGGAAGCAACCGATAAAGAAG
AGGTGATTGAAATTGTAAAAGAGTTAGCCGAATTGGCCAAGCAGAGTACTGA
CTCCGAATTAGTCAACGAGATTGTGAAGCAATTAGCAGAAGTGGCGAAGGAG
GCGACAGATAAGGAGTTGGTAATTTACATCGTGAAGATCTTGGCTGAATTAG
CGAAGCAAAGCACCGACAGCGAACTGGTAAACGAGATTGTAAACAGCTTG
CGGAGGTGGCTAAAGAGGCAACGGACAAAGAGTTAGTGATTTATATCGTAAA
GATCCTGGCTGAGTTAGCCAAGCAATCCACTGATTCAGAGTTGGTTAACGAA
ATTGTAAAGCAGCTTGAAGAGGTCGCCAAAGAGGCTACTGACAAAGAACTG
GTAGAACACATTGAAAAGATCCTGGAGGAGCTTAAAAAGCAGTCCACAGAC
GGTTGGCTGGAGGGAGGTTCTCCAGGTGGATCCCCAGGTGGTAGTCCAGGTG
GATCCGGAGGTTCTTATTGTTGTGGGGTTTGTTGTATCTAA

MRGSHHHHHHMDSEEVNERVKQLAEKAKEATDKEEVIEIVKELAELAKQSTDSE
LVNEIVKQLAEVAKEATDKELVIYIVKILAELAKQSTDSELVNEIVKQLAEVAKE

ATDKELVIYIVKILAELAKQSTDSELVNEIVKQLEEVAKEATDKELVEHIEKILEEL
KKQSTDGWLEGGSPGGSPGGSPGGSGGSYCCGVCCI

5.6.9 Non-canonical proline incorporation and protein expression

A general protocol for ncPro incorporation is described here; exact conditions for some experiments may differ. A proline auxotroph strain of *E. coli* (here, usually strain CAG18515) carrying a plasmid encoding a protein of interest (POI) and overexpressing the *Ec*ProRS (or a point mutant thereof) was grown from a single colony to stationary phase overnight in rich medium supplemented with ampicillin (or the appropriate alternative antibiotic). The culture was diluted 1:100 into 1x 20aa M9 medium of the following composition: 8.5 mM NaCl, 18.7 mM NH_4Cl, 22 mM KH_2PO_4, 47.8 mM Na_2HPO_4, 0.1 mM $CaCl_2$, 1 mM $MgSO_4$, 3 mg L^{-1} $FeSO_4$, 1 µg L^{-1} trace metals [Cu^{2+}, Mn^{2+}, Zn2+, MoO_4^{2-}], 35 mg L^{-1} thiamine HCl, 10 mg L^{-1} biotin, 20 mM D-glucose, 100 mg L^{-1} ampicillin, 50 mg L^{-1} of each of the twenty canonical L-amino acids. Cultures were grown at 37°C with shaking until they reached late-exponential phase (OD_{600} ~0.8), at which point they were subjected to a medium shift: cells were pelleted via centrifugation (4,000 g for 6 min, 4°C), and washed twice with ice-cold 0.9% NaCl. Washed cells were re-suspended in 1.25x 19aa M9 medium (1.25x-concentrated M9 medium, without L-proline). After cells were incubated for 30 min at 37°C to deplete L-proline, 0.3 M NaCl and 0.5 mM ncPro were added to each culture. Cells were incubated for 30 min at 37°C to allow ncPro uptake, and expression of the POI was induced with IPTG.

For staggered expression of GFP1-10/11 fragments, expression of the GFP1-10 fragment was induced with IPTG at mid-log phase (OD_{600} ~0.4) at 25°C before the medium shift

into 1.25x 19aa M9 medium; at this stage, glycerol replaced glucose as the carbon source for araBAD promoter use. After proline depletion and ncPro uptake steps, expression of the GFP11 strand was induced with 0.1% arabinose.

5.6.10 Quantification of ncPro incorporation

ncPro residues were incorporated into proinsulin under the general conditions described above using plasmid pQE8PI-proS. Proinsulin expression was induced with 1 mM IPTG for 2.5 h, after which cells were harvested by centrifugation and stored at -80°C until further processing.

Cell pellets were thawed and lysed with B-PER complete (Thermo Scientific) for 1 h at room temperature. Lysates were cleared by centrifugation (10,000 g, 10 min), and the supernatant was discarded. The insoluble inclusion body fraction was washed once with wash buffer (2 M urea, 20 mM tris, 1% triton X-100, pH 8.0) and twice with ddH$_2$O. The inclusion body was resuspended in solubilization buffer (8 M urea, 300 mM NaCl, 50 mM NaH$_2$PO$_4$, pH 8.0) and incubated for 1 h at RT, or overnight at 4°C. Insoluble debris was removed by centrifugation (10,000 g, 10 min) and discarded. The supernatant (now containing solubilized proinsulin) was analyzed by SDS-PAGE and MALDI-TOF mass spectrometry. The presence of a distinct proinsulin band by SDS-PAGE usually corresponded to high incorporation efficiencies by mass spectrometry.

To quantify ncPro incorporation, we digested proinsulin with the peptidase Glu-C and analyzed the peptide fragment ^{33}RGFFYT**P**KTRRE by MALDI-TOF. Crude solubilized proinsulin was reduced (5 mM DTT, 55°C, 20 min) and alkylated (15 mM iodoacetamide, RT, 15 min) prior to 10-fold dilution in to 100 mM NH$_4$HCO$_3$, pH 8.0 (100 µM total

volume). 0.6 µL of 0.5 µg µL^{-1} Glu-C (Promega) was added, and samples were digested for 2.5 h at 37°C. Digestion was halted by the addition of 10 µL of 5% TFA, and peptides were desalted using ZipTip with C$_{18}$ resin (MilliporeSigma) according to the manufacturer's protocols. Desalted peptides (in 50% acetonitrile, 0.1% TFA) were diluted 3:1 into α-CN matrix solution (α-cyanohydroxycinnamic acid in 50% ACN, 0.1% TFA) and analyzed by MALDI-TOF MS. Analog incorporation was calculated by comparing total area under the curve (AUC) of the ion corresponding to the ncPro form of the peptide, with total AUC of that of the canonical proline peptide (m/z = 1558).

5.6.11 Western blot analysis of protein solubility

1 mL samples were centrifuged (14 kg, 1 min) then stored at -80°C. Cells were lysed with B-PER Complete (Thermo Fisher, 5 mL/g cell pellet) for 1 h at room temperature, then centrifuged (14 kg, 10 min). The supernatant was removed as the soluble fraction. The inclusion body fraction was washed once with Triton X wash buffer and twice with ddH$_2$O. The washed inclusion body was resuspended in Ni-NTA binding buffer for 1 h at room temperature, then centrifuged to remove insoluble debris. The supernatant was collected as the inclusion body fraction. 9 µL of solubilized inclusion body was mixed with 3 µL SDS-PAGE loading buffer and heated to 95°C for 5-10 min, then loaded onto a Nu-PAGE bis-tris 4-12% pre-cast gel. Proteins were transferred onto a nitrocellulose membrane by the iBlot 2 Dry Blotting System (Thermo Fisher). The blot was blocked with a 5% powdered milk solution in PBST and stained with an Alexa488-conjugated anti-hexahistidine antibody (4°C, overnight). The western blot was washed three times with PBST, then imaged on a Typhoon Trio Variable Mode Imager (GE Healthcare) with a 488 nm laser and 520 nm band-pass filter.

5.6.12 FlAsH labeling (optimized conditions)

TC-tagged proteins were expressed as the POI under the general ncPro incorporation conditions described above, except for the following changes. 100 μM FlAsH-EDT$_2$ and 200 μM EDT were added (as 500x stocks dissolved in DMSO) after the medium shift and resuspension in 1.25x 19aa M9 medium. Samples were kept in the dark after the addition of FlAsH. After proline depletion and ncPro uptake steps, Top7-TC3 or DHR14-TC3 expression was induced with 0.5 mM IPTG for 2.5 h at 37°C, then prepared for flow cytometry (see below).

5.6.13 Flow cytometry and FACS

In preparation for analysis by flow cytometry, samples were centrifuged, washed twice with PBS, and filtered through 40 μm nylon strainers (Corning). For cells labeled with FlAsH, we found that washing once with PBSDE (33 mM EDT, 3% v/v DMSO in PBS), once with PBSD (3% v/v DMSO in PBS), then resuspending samples in PBS before filtering samples modestly reduced background FlAsH labeling and improved resolution between –Proline and +Proline controls. Cells were analyzed on a MoFlo XDP cell sorter (Beckman Coulter) using a 488 nm laser and 530/40 nm bandpass filter.

Events with the desired fluorescence characteristics were sorted: for libraries, these were the brightest ~1%; doublets were excluded from sorting by gating on FL1-Hight vs FL1-Area. Cells were sorted into rich medium and rescued for one hour at 37°C, then diluted into selective medium and grown overnight at 37°C. Samples were either diluted the next day into M9 medium from this overnight culture for subsequent rounds of flow cytometry analysis and FACS, or stored at -80°C in 25% glycerol stocks until further analysis. Rescue

efficiency was determined by plating cells immediately after the 1 h rescue step onto selective agar plates, and was defined as CFUs / events sorted.

Alternatively, cells were sorted into empty tubes, then centrifuged (20,000 g, 10 min). The supernatant was carefully removed, and plasmid DNA was isolated using a ZymoPURE plasmid miniprep kit. The manufacturer's protocols were generally followed, though buffer volumes were scaled by a factor of 2/7, and Zymo-Spin I columns (rather than Zymo-Spin II columns) were used. Eluted plasmid DNA was then used to transform new stocks of *E. coli*.

5.6.14 Replica plating

E. coli strain CAG18515 harboring plasmid pQE80PI-proS, pQE80PI-proS-M157Q, or pQE80PI were plated onto LB agar plates containing ampicillin and grown until large colonies formed. This master plate was imaged (EPSON scanner), then colonies were transferred to sterile velvet. This was used to stamp colonies onto M9 agar plates containing 1, 5, or 10 µM proline, 0.3 M NaCl, and 0, 25, 50, 100, 250, and 500 µM 4S-NH$_2$. Plates were stamped in order of increasing concentrations of 4S-NH$_2$, then finally onto an LB agar plate lacking 4S-NH$_2$. After incubation at 37°C, scanned images of replica plates were obtained and processed with ImageJ to measure growth inhibition.

5.6.15 Identification of proline-free carrier proteins

We reasoned that the Protein Data Bank may be a good resource to identify soluble proteins that express well in *E. coli*. We identified 2057 unique PDBid's corresponding to protein entities that do not contain proline residues. We limited this list to proteins that were expressed in *E. coli*, and those that existed as monomers. We confirmed that proteins were

purified from the soluble fraction, and avoided membrane proteins, enzymes, DNA binding proteins, and those that bound cofactors. One publication[44] designed many proteins that fit our criteria; in this case, two proteins from this manuscript were chosen. Codon optimization occasionally failed; in these cases, proteins were either omitted (4H3L 3), or truncated (Tako8). The final list of proline-free proteins tested here is displayed in Table 5.S2.

5.6.16 Microscopy

Images were obtained at the Biological Imaging Facility at Caltech on a Zeiss LSM800 confocal microscope. A 488 nm laser was used for fluorescence images.

5.7 Supplementary figures and tables
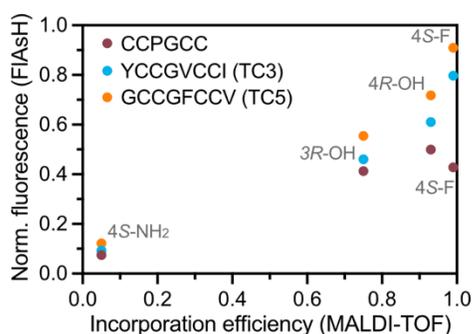
Figure 5.S1. Proline-containing TC motifs do not accurately measure proline analog incorporation. *E. coli* strain CAG18515 overexpressing the ProRS was treated with the indicated ncPro after a medium shift and FlAsH labeling. Expression of Top7, which was translationally fused to a proline-containing linker and the indicated TC tag, was then induced, and fluorescence was determined by flow cytometry. Normalized fluorescence is relative to the –Proline and +Proline controls for each strain. Compared to proline-free TC tags TC3 and TC5, normalized fluorescence for ncPro-treated samples with the canonical CCPGCC TC tag does not correlate well with known incorporation efficiencies as determined by MALDI-TOF.
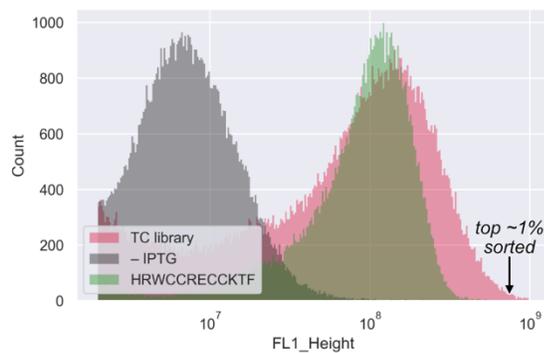
Figure 5.S2. Development of a proline-free TC motif. A library of proline-free TC motifs was fused to the C-terminus of Top7 and a proline-containing linker. Shown here is the fluorescence histogram of TC library against the previously reported proline-free TC motif used in mammalian cells: HRWCCRECCKTF.[51] We sorted the brightest ~1% of events once, which yielded TC3 (YCCGVCCI) and TC5 (GCCGFCCV) as the top-performing TC tags (data not shown).
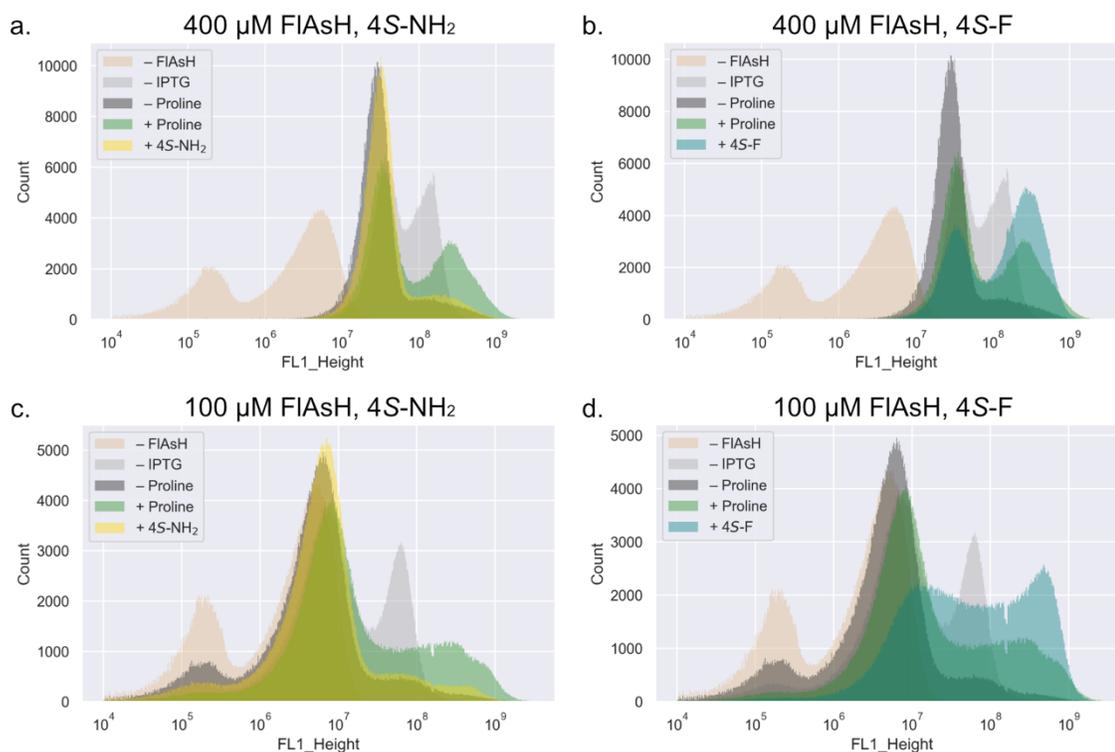
Figure 5.S3. FlAsH labeling after Top7 expression leads to heterogeneous fluorescence. Cultures were treated with the indicated concentrations of FlAsH, and four-fold concentrations of EDT, for 30 min after Top7-TC3 expression under ncPro incorporation conditions. Similar histograms were obtained for a broad range of ncPro residues tested (data not shown); in these cases, the brightest ncPro population corresponded to its known incorporation efficiency. We note the two populations across all samples, and conclude that under these conditions, only a small subset of cells is sufficiently labeled with FlAsH. We hypothesize that the limiting factor is FlAsH entry into cells, since extended incubation of FlAsH with these strains leads to a higher proportion of bright cells (i.e., Figure 5.14f).
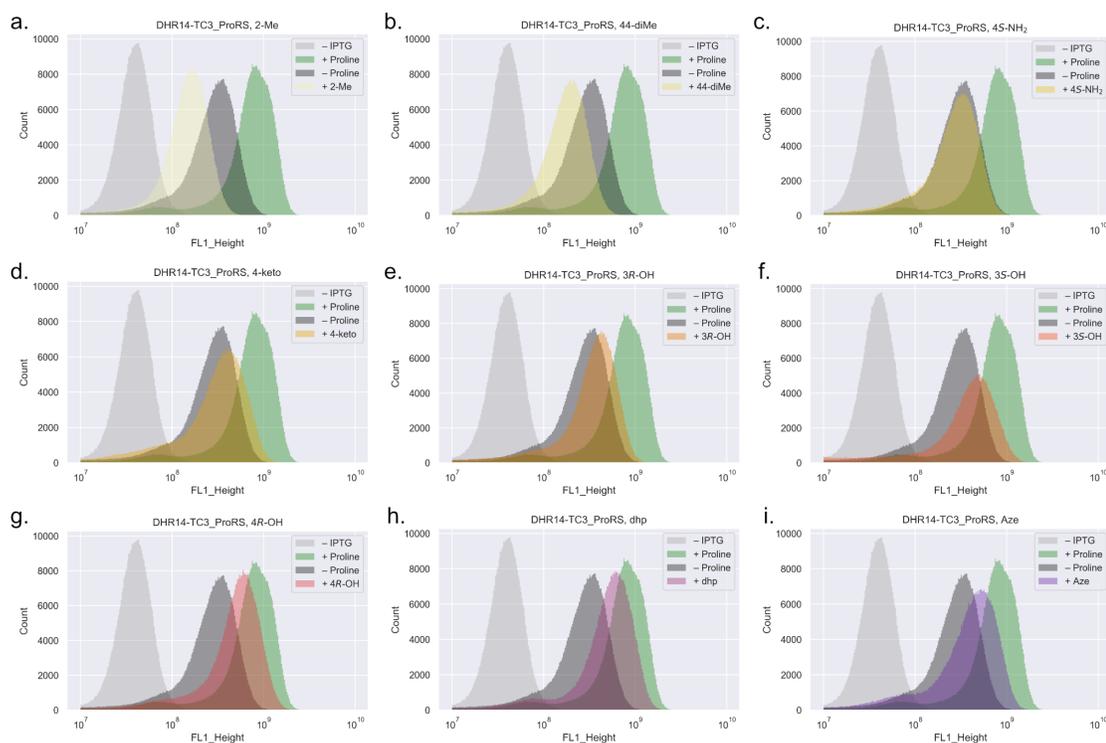
Figure 5.S4. **FlAsH labeling after DHR14-TC3 expression can measure proline analog incorporation**. Interestingly, the fluorescence data here suggest that ncPro residues 2-methylproline (2-Me) and 4,4-dimethylproline (44-diMe) inhibit residual proline uptake or incorporation, since samples treated with these proline analogs were less fluorescent than the – Proline negative control. These results are consistent with SDS-PAGE and MALDI-TOF data (Figure 2.S1, Table 2.S2). A summary of these flow cytometry histograms is displayed in Figure 5.17a.
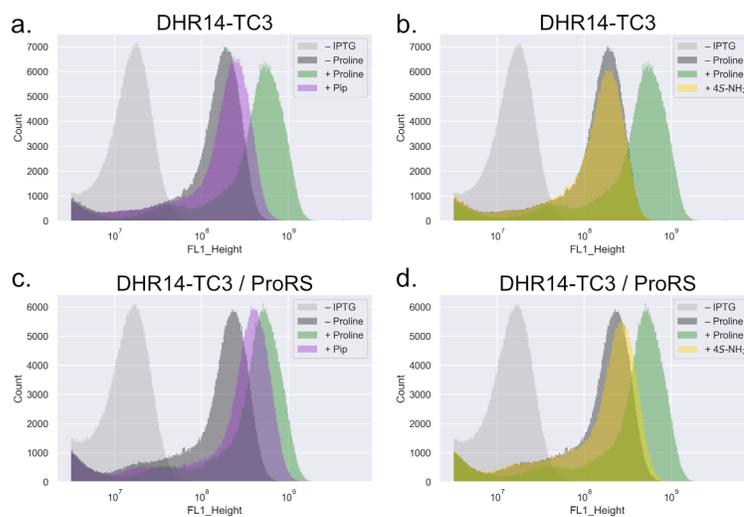
Figure 5.S5. ProRS overexpression increases proline analog incorporation. Strains harboring a plasmid either without (a-b) or with (c-d) *proS* overexpression were treated with the ncPro residues Pip (a,c) or 4*S*-NH$_2$ (b,d). A summary of these data is displayed in Figure 5.17b.
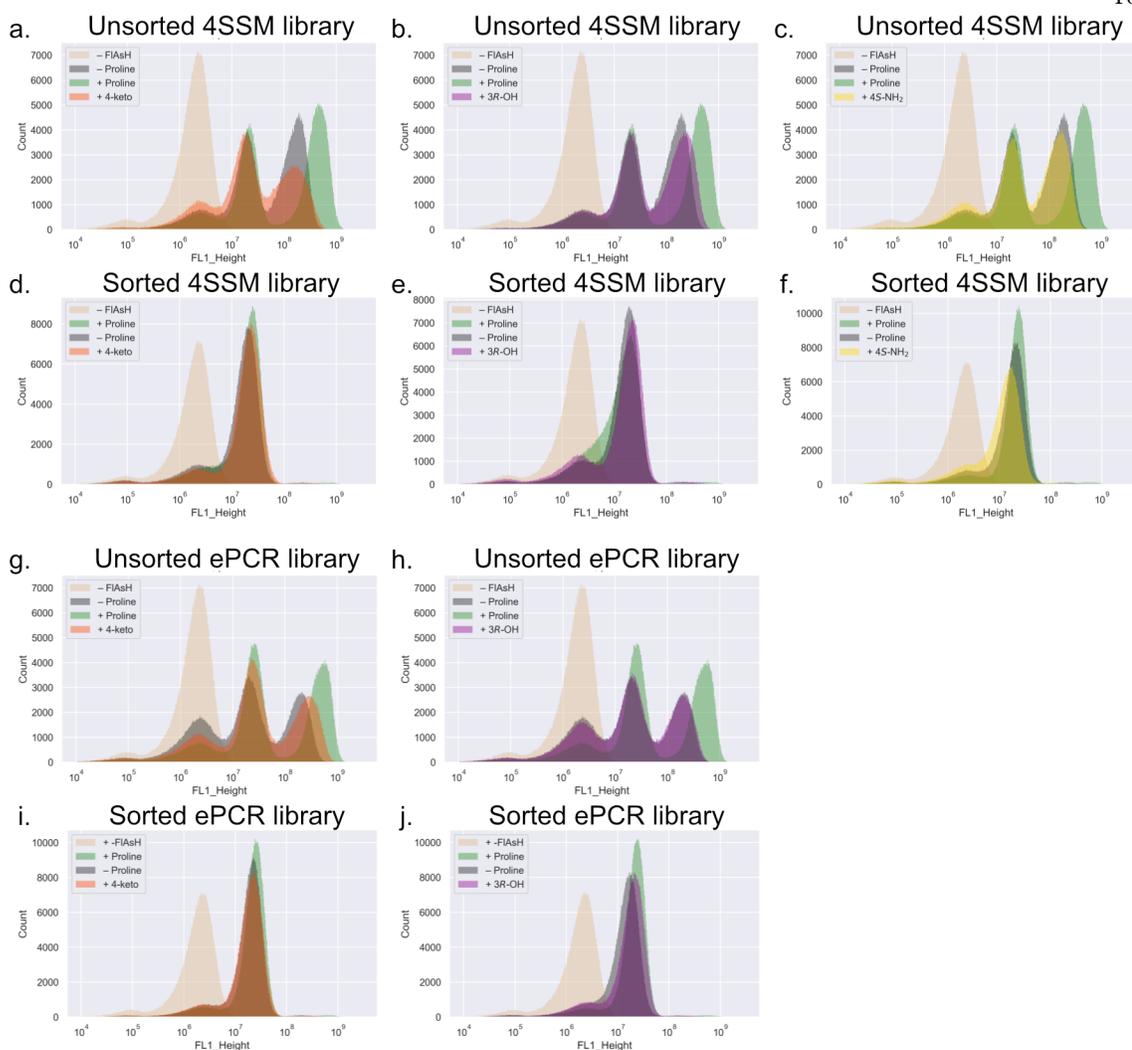
Figure 5.S6. Cheaters overwhelm library samples after sorting. 4-codon site saturation mutagenesis (4SSM, a-c) and error-prone PCR (ePCR, g-h) libraries were constructed by a restriction enzyme approach and treated with the proline analogs 4-keto (a,g), 3R-OH (b,h), and 4S-NH$_2$ (c). The brightest ~1% of each of these ncPro samples was sorted. Plasmid DNA was purified from sorted cells, and used to transform fresh stocks of CAG18515 for a second round of analysis by flow cytometry (d-f, i-j). As with the ePCR library treated with 4S-NH$_2$ shown in Figure 5.18a-b, each of the sorted samples contains only a small (~1%) population of bright cells.
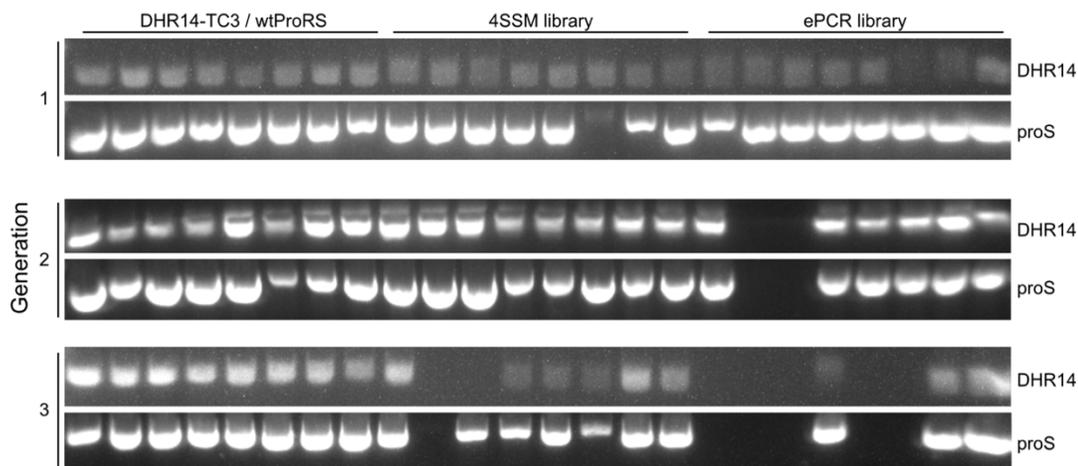
Figure 5.S7. Plasmid stability over rounds of transformation and growth. The indicated plasmid DNA was used to transform strain CAG18515, a portion of which were plated onto selective (ampicillin) agar plates. Colony PCR was used to interrogate the presence of the plasmid-borne DHR14-TC3 gene, or overexpressed *proS* gene, using primers SLB2429_analysis_fwd / SLB2429_analysis_rev, and SLB2406_fwd / SLB2406_rev, respectively. The remaining culture after transformation was grown to stationary phase in rich medium, after which plasmids were purified and used to transform subsequent generations. As opposed to previous experiments in which a restriction enzyme approach was used, the libraries here was constructed with Gibson Assembly. A summary of these data is displayed in Figure 5.818d; a colony was considered "PCR-positive" if a band was observed for both DHR14-TC3 and *proS* genes.
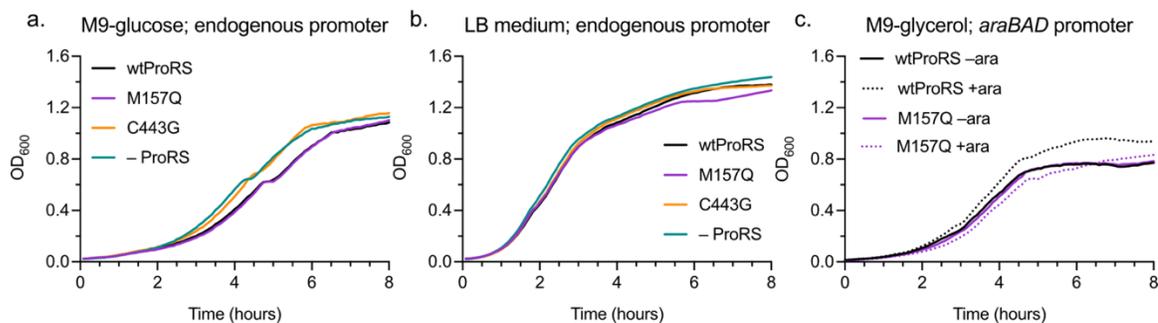
Figure 5.S8. Growth depends upon ProRS variant and promoter. **a-b**. *E. coli* strain CAG18515 harboring plasmid pQE80_DHR14-TC3_proS (or pQE80_DHR14-TC3) were grown in a plate reader at 37°C with shaking in 20aa M9 medium with glucose as a carbon source (a), or in LB medium (b); growth was monitored over time by tracking absorbance at 600 nm. **c**. CAG18515 harboring plasmid pQE80_DHR14-TC3_proS-araC, or the M157Q point mutant of *proS*, were grown similarly in M9 medium with glycerol as the carbon source; *proS* expression was induced by the addition of arabinose. Shown here is the average of three technical replicates.

Table 5.S1. **Mutations rates for error-prone PCR library generation**

[MnCl$_2$] (µM)	Mutations per *proS* gene
0	0.67 ± 1.50
200	2.13 ± 1.87
300	3.63 ± 2.12
400	7.37 ± 4.40

Table 5.S2. **Amino acid sequences of proline-free proteins**

Name	Amino acid sequence	MW (kDa)
Tako8	mrgshhhhhhMGQSLRTLQGHQSAVTSLQFNDNIVVSGSDD STVKVWDIKTGQSLRTLQGHQSAVTSLQFNDNIVVSGSD DSTVKVWDIKTGGSggspggspggspggsggsyccgvcci	12.4
Ika4	mrgshhhhhhMGQELVSLEGHQSAITALAFSKNIVVSGAAD GTIKVWDILTGQLLRDHDGHQSEVTALQFKDNIVVSGA KDGTVKVWYIGTGQELVSLEGHQSAITALAFSKNIVVSG AADGTIKVWDILTGQLLRDHDGHQSEVTALQFKDNIVV SGAKDGTVKVWYIGTGGSggspggspggspggsggsyccgvcci	20.7
DHR10	mrgshhhhhhMSSEKEELRERLVKIVVENAKRKGDDTEEAR EAAREAFELVREAAERAGIDSSEVLELAIRLIKEVVENAQ REGYDISEAARAAAEAFKRVAEAAKRAGITSSEVLELAI RLIKEVVENAQREGYDISEAARAAAEAFKRVAEAAKRA GITSSETLKRAIEEIRKRVEEAQREGNDISEAARQAAEEFR KKAEELKRRGDGWLEggspggspggspggsggsyccgvcci	26.2
DHR14	mrgshhhhhhMDSEEVNERVKQLAEKAKEATDKEEVIEIVK ELAELAKQSTDSELVNEIVKQLAEVAKEATDKELVIYIV KILAELAKQSTDSELVNEIVKQLAEVAKEATDKELVIYIV KILAELAKQSTDSELVNEIVKQLEEVAKEATDKELVEHIE KILEELKKQSTDGWLEggspggspggspggsggsyccgvcci	22.1
3H5L_2	mrgshhhhhhNEDDMKKLYKQMVQELEKARDRMEKLYKE MVELIQKAIELMRKIFQEVKQEVEKAIEEMKKLYDEAKK KIEQMIQQIKQGGDKQKMEELLKRAKEEMKKVKDKME KLLEKLKQIMQEAKQKMEKLLKQLKEEMKKMKEKMEK LLKEMKQRMEEVKKKMDGDDELLEKIKKNIDDLKKIAE DLIKKAEENIKEAKKIAEQLVKRAKQLIEKAKQVAEELIK KILQLIEKAKEIAEKVLKGLEggspggspggspggsggsyccgvcci	32.4
Blo t 21	mrgshhhhhhNTATQRFHEIEKFLLHITHEVDDLEKTGNKDE KARLLRELTVSEAFIEGSRGYFQRELKRTDLDLLEKFNFE AALATGDLLLKDLKALQKRVQDSEggspggsp	14.6
Pkd2	mrgshhhhhhMGSTAIGINDTYSEVKSDLAQQKAEMELSDLI RKGYHKALVKLKLKKNTVDDISESLRQGGGKLNFDELR QDLKGKGHTDAEIEAIFTKYDQDGDQELTEHEHQQMRD DLEKEREDLDLDHSSLggspggspggspggsggsyccgvcci	17.5
Utr-SR1	mrgshhhhhhDMDLDSYQIALEEVLTWLLSAEDTFQEQDDI SDDVEDVKEQFATHETFMMELSAHQSSVGSVLQAGNQL MTQGTLSDEEEFEIQEQMTLLNARWEALRVESMERQSR LHDALMELQKKQLQQLggspggspggspggsggsyccgvcci	17.7
Top7	MGDIQVQVNIDDNGKNFDYTYTVTTESELQKVLNELKD YIEEQGAKRARISITARTEKEAEKFAAILIKVFAELGYNDI NVTWDGDTVTVEGQLEGGSLEpggspggspggsggsyccgvcci	13.1

*The N-terminal 6xHis tag and C-terminal proline linker + TC3 sequences are in lowercase; the sequences corresponding to the proline-free protein are in UPPERCASE.

5.8 References

(1)   Ling, J.; Reynolds, N.; Ibba, M. Aminoacyl-tRNA Synthesis and Translational Quality Control. *Annu Rev Microbiol* **2009**, *63*, 61–78. https://doi.org/10.1146/annurev.micro.091208.073210.

(2)   Chin, J. W. Expanding and Reprogramming the Genetic Code. *Nature* **2017**, *550*, 53–60. https://doi.org/10.1038/nature24031.

(3)   Yoo, T. H.; Tirrell, D. A. High-Throughput Screening for Methionyl-tRNA Synthetases That Enable Residue-Specific Incorporation of Noncanonical Amino Acids into Recombinant Proteins in Bacterial Cells. *Angew Chem Int Ed* **2007**, *46*, 5340–5343. https://doi.org/10.1002/anie.200700779.

(4)   Ngo, J. T.; Champion, J. A.; Mahdavi, A.; Tanrikulu, I. C.; Beatty, K. E.; Connor, R. E.; Yoo, T. H.; Dieterich, D. C.; Schuman, E. M.; Tirrell, D. A. Cell-Selective Metabolic Labeling of Proteins. *Nat Chem Biol* **2009**, *5* (10), 715–717. https://doi.org/10.1038/nchembio.200.

(5)   Renner, C.; Alefelder, S.; Bae, J. H.; Budisa, N.; Huber, R.; Moroder, L. Fluoroprolines as Tools for Protein Design and Engineering. *Angew Chem Int Ed* **2001**, *40* (5), 923–925. https://doi.org/10.1002/1521-3773(20010302)40:5<923::AID-ANIE923>3.0.CO;2-#.

(6)   Breunig, S. L.; Tirrell, D. A. Incorporation of Proline Analogs into Recombinant Proteins Expressed in *Escherichia coli. Methods Enzymol* **2021**, *656*, 545–571. https://doi.org/10.1016/BS.MIE.2021.05.008.

(7)   Kim, W.; George, A.; Evans, M.; Conticello, V. P. Cotranslational Incorporation of a Structurally Diverse Series of Proline Analogues in an *Escherichia coli* Expression System. *ChemBioChem* **2004**, *5* (7), 928–936. https://doi.org/10.1002/cbic.200400052.

(8)   Packer, M. S.; Liu, D. R. Methods for the Directed Evolution of Proteins. *Nat Rev Genet* **2015**, *16*, 379–394. https://doi.org/10.1038/nrg3927.

(9)   Krahn, N.; Tharp, J. M.; Crnković, A.; Söll, D. Engineering Aminoacyl-tRNA Synthetases for Use in Synthetic Biology. *Enzymes* **2020**, *48*, 351. https://doi.org/10.1016/BS.ENZ.2020.06.004.

(10)   Chatterjee, A.; Xiao, H.; Schultz, P. G. Evolution of Multiple, Mutually Orthogonal Prolyl-tRNA Synthetase/tRNA Pairs for Unnatural Amino Acid Mutagenesis in *Escherichia coli. Proc Natl Acad Sci* **2012**, *109* (37), 14841–14846. https://doi.org/10.1073/pnas.1212454109.

(11)   Carpenter, H. E. Biomaterials Design: Creation of a Novel Orthogonal Translational System in *E. coli* for the Site-Specific Incorporation of Proline Analogues, Emory University, 2002.

(12)   Johnson, J. A.; Lu, Y. Y.; Van Deventer, J. A.; Tirrell, D. A. Residue-Specific Incorporation of Non-Canonical Amino Acids into Proteins: Recent Developments and Applications. *Curr Opin Chem Biol* **2010**, *14*, 774–780. https://doi.org/10.1016/j.cbpa.2010.09.013.

(13) Steiner, T.; Hess, P.; Bae, J. H.; Wiltschi, B.; Moroder, L.; Budisa, N. Synthetic Biology of Proteins: Tuning GFPs Folding and Stability with Fluoroproline. *PLoS One* **2008**, *3* (2), e1680. https://doi.org/10.1371/journal.pone.0001680.

(14) Deepankumar, K.; Nadarajan, S. P.; Ayyadurai, N.; Yun, H. Enhancing the Biophysical Properties of mRFP1 through Incorporation of Fluoroproline. *Biochem Biophys Res Commun* **2013**, *440* (4), 509–514. https://doi.org/10.1016/j.bbrc.2013.09.062.

(15) To, T. M. T.; Kubyshkin, V.; Schmitt, F. J.; Budisa, N.; Friedrich, T. Residue-Specific Exchange of Proline by Proline Analogs in Fluorescent Proteins: How "Molecular Surgery" of the Backbone Affects Folding and Stability. *J Vis Exp* **2022**, *2022* (180). https://doi.org/10.3791/63320.

(16) Shoulders, M. D.; Raines, R. T. Collagen Structure and Stability. *Annu Rev Biochem* **2009**, *78*, 929–958. https://doi.org/10.1146/annurev.biochem.77.032207.120833.

(17) Ganguly, H. K.; Basu, G. Conformational Landscape of Substituted Prolines. *Biophys Rev* **2020**, *12*, 25–39. https://doi.org/10.1007/s12551-020-00621-8.

(18) Ong, W. J. H.; Alvarez, S.; Leroux, I. E.; Shahid, R. S.; Samma, A. A.; Peshkepija, P.; Morgan, A. L.; Mulcahy, S.; Zimmer, M. Function and Structure of GFP-like Proteins in the Protein Data Bank. *Mol Biosyst* **2011**, *7* (4), 984–992. https://doi.org/10.1039/C1MB05012E.

(19) Fang, K. Y. Modulating Biophysical Properties of Insulin with Non-Canonical Mutagenesis at Position B28, California Institute of Technology, 2017. https://doi.org/10.7907/Z9V40S6J.

(20) Cabantous, S.; Terwilliger, T. C.; Waldo, G. S. Protein Tagging and Detection with Engineered Self-Assembling Fragments of Green Fluorescent Protein. *Nat Biotechnol* **2005**, *23* (1), 102–107. https://doi.org/10.1038/nbt1044.

(21) de Boer, H. A.; Comstock, L. J.; Vasser, M. The Tac Promoter: A Functional Hybrid Derived from the Trp and Lac Promoters. *Proc Natl Acad Sci* **1983**, *80* (1), 21–25. https://doi.org/10.1073/pnas.80.1.21.

(22) Deng, A.; Boxer, S. G. Structural Insight into the Photochemistry of Split Green Fluorescent Proteins: A Unique Role for a His-Tag. *J Am Chem Soc* **2018**, *140*, 375–381. https://doi.org/10.1021/jacs.7b10680.

(23) Guzman, L.-M.; Belin, D.; Carson, M. J.; Beckwith, J. Tight Regulation, Modulation, and High-Level Expression by Vectors Containing the Arabinose PBAD Promoter. *J Bacteriol* **1995**, *177* (14), 4121–4130. https://doi.org/10.1128/jb.177.14.4121-4130.1995.

(24) Kamiyama, D.; Sekine, S.; Barsi-Rhyne, B.; Hu, J.; Chen, B.; Gilbert, L. A.; Ishikawa, H.; Leonetti, M. D.; Marshall, W. F.; Weissman, J. S.; Huang, B. Versatile Protein Tagging in Cells with Split Fluorescent Protein. *Nat Commun* **2016**, *7*, 11046. https://doi.org/10.1038/ncomms11046.

(25)   Cabantous, S.; Waldo, G. S. In Vivo and in Vitro Protein Solubility Assays Using Split GFP. *Nat Methods* **2006**, *3* (10), 845–854. https://doi.org/10.1038/nmeth932.

(26)   Feinberg, E. H.; VanHoven, M. K.; Bendesky, A.; Wang, G.; Fetter, R. D.; Shen, K.; Bargmann, C. I. GFP Reconstitution Across Synaptic Partners (GRASP) Defines Cell Contacts and Synapses in Living Nervous Systems. *Neuron* **2008**, *57* (3), 353–363. https://doi.org/10.1016/j.neuron.2007.11.030.

(27)   Pinaud, F.; Dahan, M. Targeting and Imaging Single Biomolecules in Living Cells by Complementation-Activated Light Microscopy with Split-Fluorescent Proteins. *Proc Natl Acad Sci* **2011**, *108* (24), E201–E210. https://doi.org/10.1073/pnas.1101929108.

(28)   Hyun, S. I.; Maruri-Avidal, L.; Moss, B. Topology of Endoplasmic Reticulum-Associated Cellular and Viral Proteins Determined with Split-GFP. *Traffic* **2015**, *16* (7), 787–795. https://doi.org/10.1111/tra.12281.

(29)   van Engelenburg, S. B.; Palmer, A. E. Imaging Type-III Secretion Reveals Dynamics and Spatial Segregation of Salmonella Effectors. *Nat Methods 2010 7:4* **2010**, *7* (4), 325–330. https://doi.org/10.1038/nmeth.1437.

(30)   Kaddoum, L.; Magdeleine, E.; Waldo, G. S.; Joly, E.; Cabantous, S. One-Step Split GFP Staining for Sensitive Protein Detection and Localization in Mammalian Cells. *Biotechniques* **2010**, *49* (4), 727–736. https://doi.org/10.2144/000113512.

(31)   Kim, Y. E.; Kim, Y. N.; Kim, J. A.; Kim, H. M.; Jung, Y. Green Fluorescent Protein Nanopolygons as Monodisperse Supramolecular Assemblies of Functional Proteins with Defined Valency. *Nat Commun 2015 6:1* **2015**, *6* (1), 1–9. https://doi.org/10.1038/ncomms8134.

(32)   Lawrence, M. S.; Phillips, K. J.; Liu, D. R. Supercharging Proteins Can Impart Unusual Resilience. *J Am Chem Soc* **2007**, *129*, 10110–10112. https://doi.org/10.1021/ja071641y.

(33)   Huang, Y.-M.; Nayak, S.; Bystroff, C. Quantitative in Vivo Solubility and Reconstitution of Truncated Circular Permutants of Green Fluorescent Protein. *Protein Sci* **2011**, *20*, 1775–1780. https://doi.org/10.1002/pro.735.

(34)   Minks, C.; Alefelder, S.; Moroder, L.; Huber, R.; Budisa, N. Towards New Protein Engineering: In Vivo Building and Folding of Protein Shuttles for Drug Delivery and Targeting by the Selective Pressure Incorporation (SPI) Method. *Tetrahedron* **2000**, *56* (48), 9431–9442. https://doi.org/10.1016/S0040-4020(00)00827-9.

(35)   Schellenberger, V.; Wang, C.-W.; Geething, N. C.; Spink, B. J.; Campbell, A.; To, W.; Scholle, M. D.; Yin, Y.; Yao, Y.; Bogin, O.; Cleland, J. L.; Silverman, J.; Stemmer, W. P. C. A Recombinant Polypeptide Extends the in Vivo Half-Life of Peptides and Proteins in a Tunable Manner. *Nat Biotechnol* **2009**, *27* (12), 1186–1190. https://doi.org/10.1038/nbt.1588.

(36) Wittrup, K. D.; Mann, M. B.; Fenton, D. M.; Tsai, L. B.; Bailey, J. E. Single-Cell Light Scatter as a Probe of Refractile Body Formation in Recombinant *Escherichia coli. Bio/Technology.* 1988, pp 423–426. https://doi.org/10.1038/nbt0488-423.

(37) Lavis, L. D.; Raines, R. T. Bright Ideas for Chemical Biology. *ACS Chem Biol* **2008**, *3* (3), 142–155. https://doi.org/10.1021/cb700248m.

(38) Griffin, B. A.; Adams, S. R.; Tsien, R. Y. Specific Covalent Labeling of Recombinant Protein Molecules Inside Live Cells. *Science (1979)* **1998**, *281*, 269–272. https://doi.org/10.1126/science.281.5374.269.

(39) Landgraf, D.; Okumus, B.; Chien, P.; Baker, T. A.; Paulsson, J. Segregation of Molecules at Cell Division Reveals Native Protein Localization. *Nat Methods 2012 9:5* **2012**, *9* (5), 480–482. https://doi.org/10.1038/nmeth.1955.

(40) Stroffekova, K.; Proenza, C.; Beam, K. G. The Protein-Labeling Reagent FlAsH-EDT$_2$ Binds Not Only to CCXXCC Motifs but Also Non-Specifically to Endogenous Cysteine-Rich Proteins. *Pflugers Arch* **2001**, *442* (6), 859–866. https://doi.org/10.1007/s004240100619.

(41) Dantas, G.; Watters, A. L.; Lunde, B. M.; Eletr, Z. M.; Isern, N. G.; Roseman, T.; Lipfert, J.; Doniach, S.; Tompa, M.; Kuhlman, B.; Stoddard, B. L.; Varani, G.; Baker, D. Mis-Translation of a Computationally Designed Protein Yields an Exceptionally Stable Homodimer: Implications for Protein Engineering and Evolution. *J Mol Biol* **2006**, *362* (5), 1004–1024. https://doi.org/10.1016/j.jmb.2006.07.092.

(42) Adams, S. R.; Campbell, R. E.; Gross, L. A.; Martin, B. R.; Walkup, G. K.; Yao, Y.; Llopis, J.; Tsien, R. Y. New Biarsenical Ligands and Tetracysteine Motifs for Protein Labeling in Vitro and in Vivo: Synthesis and Biological Applications. *J Am Chem Soc* **2002**, *124* (21), 6063–6076. https://doi.org/10.1021/ja017687n.

(43) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; Lepore, R.; Schwede, T. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res* **2018**, *46* (W1), W296–W303. https://doi.org/10.1093/nar/gky427.

(44) Brunette, T.; Parmeggiani, F.; Huang, P.-S.; Bhabha, G.; Ekiert, D. C.; Tsutakawa, S. E.; Hura, G. L.; Tainer, J. A.; Baker, D. Exploring the Repeat Protein Universe through Computational Protein Design. *Nature* **2015**, *528*, 580–584. https://doi.org/10.1038/nature16162.

(45) Navaratna, T.; Atangcho, L.; Mahajan, M.; Subramanian, V.; Case, M.; Min, A.; Tresnak, D.; Thurber, G. M. Directed Evolution Using Stabilized Bacterial Peptide Display. *J Am Chem Soc* **2020**, *142*, 1882–1894. https://doi.org/10.1021/jacs.9b10716.

(46) Ramesh, B.; Frei, C. S.; Cirino, P. C.; Varadarajan, N. Functional Enrichment by Direct Plasmid Recovery after Fluorescence Activated Cell Sorting. *Biotechniques* **2015**, *59* (3), 157–161. https://doi.org/10.2144/000114329.

(47)   Tam, L. M.; Price, N. E.; Wang, Y. Molecular Mechanisms of Arsenic-Induced Disruption of DNA Repair. *Chem Res Toxicol* **2020**, *33* (3), 709–726. https://doi.org/10.1021/acs.chemrestox.9b00464.

(48)   Kanagawa, T. Bias and Artifacts in Multitemplate Polymerase Chain Reactions (PCR). *J Biosci Bioeng* **2003**, *96* (4), 317–323. https://doi.org/10.1016/s1389-1723(03)90130-7.

(49)   Datsenko, K. A.; Wanner, B. L. One-Step Inactivation of Chromosomal Genes in *Escherichia coli* K-12 Using PCR Products. *Proc Natl Acad Sci* **2000**, *97* (12), 6640–6645. https://doi.org/10.1073/pnas.120163297.

(50)   Yuet, K. P. Tools For Spatiotemporally Specific Proteomic Analysis In Multicellular Organisms By, California Institute of Technology, 2016. https://doi.org/10.7907/z9vd6wdh.

(51)   Martin, B. R.; G Giepmans, B. N.; Adams, S. R.; Tsien, R. Y. Mammalian Cell–Based Optimization of the Biarsenical- Binding Tetracysteine Motif for Improved Fluorescence and Affinity. *Nat Biotechnol* **2005**, *23* (10), 1308–1314. https://doi.org/10.1038/nbt1136.