Title: Engineering post-translational proofreading to discriminate non-standard amino acids

Authors: Aditya M. Kunjapur[1]*, Devon A. Stork[1], Erkin Kuru[1], Oscar Vargas-Rodriguez[2], Matthieu Landon[1], Dieter Söll[2], and George M. Church[1]*

Affiliations:

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA.

[2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA.

*Correspondence to: AMK (kunjapur@alum.mit.edu) or GMC (gchurch@genetics.med.harvard.edu).

Abstract: Progress in genetic code expansion requires accurate, selective, and high-throughput detection of non-standard amino acid (NSAA) incorporation into proteins. Here, we discover how the N-end rule pathway of protein degradation applies to commonly used NSAAs. We show that several NSAAs are N-end stabilizing and demonstrate that other NSAAs can be made stabilizing by rationally engineering the N-end rule adaptor protein ClpS. We use these insights to engineer synthetic quality control, termed "Post-Translational Proofreading" (PTP). By implementing PTP, false positive proteins resulting from misincorporation of structurally similar standard amino acids or undesired NSAAs rapidly degrade, enabling high-accuracy discrimination of desired NSAA incorporation. We illustrate the utility of PTP during evolution of the biphenylalanine orthogonal translation system used for synthetic biocontainment. Our new OTS is more selective and confers lower escape frequencies and greater fitness in all tested biocontained strains. Our work presents a new paradigm for molecular recognition of amino acids in target proteins.

One Sentence Summary: We modify a natural cellular quality control process to accurately evaluate and evolve systems that incorporate non-standard amino acids into proteins.

Main Text:

Genetic code expansion broadens the structural and functional diversity of proteins and can enable synthetic biocontainment, which will be required for the 57-codon *E. coli* strain that we are constructing due to its anticipated multi-virus resistance (*1*). Genetic code expansion relies on engineered orthogonal translation systems (OTSs) that use aminoacyl-tRNA synthetases (AARSs) to catalyze esterification of tRNAs to non-standard amino acids for subsequent incorporation into proteins (*2*). Four primary OTS families have been developed for NSAA incorporation by suppression of amber (UAG) stop codons in targeted sequences (*3*, *4*) (**Fig. S1A)**. However, promiscuity of engineered OTSs for standard amino acids (SAAs) and for undesired NSAAs is a major barrier to genetic code expansion. The low fidelity of several OTSs is documented, revealing that even after multiple rounds of negative selection they misacylate tRNA with SAAs that their parental variants acted upon, such as tyrosine and tryptophan for engineered variants of the Tyr and Trp OTSs, respectively (*5*–*9*). Furthermore, members of Tyr/Trp/Pyl OTS families exhibit overlap of substrate ranges, which limits their simultaneous use for potential applications involving multiple types of NSAAs (*10*–*12*). OTS cross-talk with SAAs lowers the effectiveness of previously demonstrated synthetic biocontainment because promiscuity of the biphenylalanine (BipA) OTS, which is an evolved derivative of the *Methanococcus jannaschii* Tyr OTS, can promote escape (*13*). Finally, many other NSAA applications such as protein double labelling, FRET, and antibody conjugation, require high fidelity incorporation to avoid heterogeneous protein production.

Currently, the identity of an incorporated amino acid is best determined using low-throughput protein purification and mass spectrometry. We were interested in evaluating the promiscuity of the BipA OTS and first performed higher-throughput SAA spiking experiments in minimal media lacking BipA, which suggested Y and L misincorporation (**Fig. S1B**). We then used mass spectrometry to reveal that target peptides produced upon expression of the BipA OTS but in the absence of BipA contained 90%+ Y/L/F at the target site, with Q also present due to expected near-cognate suppression (*14*, *15*) (**Fig. S1C**). These experiments further demonstrate that SAA misincorporation is a barrier to genetic code expansion and synthetic biocontainment. To address this problem for the BipA OTS and OTSs more generally, we sought to develop a new incorporation detection system with the following design criteria: (i) the ability to controllably mask and unmask misincorporation *in vivo*; (ii) compatibility with different reporter proteins; (iii) customizability for most commonly used NSAAs. We were especially interested in associating NSAA incorporation with protein stability to accomplish these aims.

Here, we discover how the N-end rule pathway of protein degradation, a natural protein regulatory and quality control pathway conserved across prokaryotes and eukaryotes (*16*–*18*), applies to commonly used NSAAs. We hypothesized that NSAAs may be N-end stabilizing, whereas their SAA analogs (Y/F/W/L/K/R) are known to be N-end destabilizing residues in *E. coli*, which result in protein half-lives on the timescale of minutes (*17*). We tested the effect of incorporation of commonly used NSAAs at the N-terminus and use our findings to develop "Post-Translational Proofreading" (PTP), which enables high-accuracy discrimination of NSAA incorporation *in vivo*. PTP is a remarkably modular, generalizable, and tunable system for specific protein recognition based on the identity of a single amino acid at the N-terminus, which is a position increasingly targeted for applications in chemical biology (*19*). We subsequently show that the ability to optionally degrade proteins containing SAA misincorporation events dramatically facilitates directed evolution for selective OTSs.

To begin investigation into how the N-end rule applies to NSAAs, we constructed a reporter consisting of a cleavable ubiquitin domain (Ub) followed by one UAG codon, a conditionally strong N-degron (*20, 21*), and a super-folder green fluorescent protein (sfGFP) with a C-terminal His6x-tag (**Fig. 1A**). We genomically integrated this reporter into a recoded *E. coli* strain devoid of UAG codons and associated release factor (C321.ΔA) (*22*). The use of only one UAG codon increases assay sensitivity for promiscuity compared to the use of multi-UAG codon reporters (*23*), and genomic integration of the reporter increases reproducibility by eliminating plasmid copy number effects (*24*). We obtained several commonly used NSAAs and began testing with BipA (**Fig. 1B**). Experiments with and without BipA (BipA+ or BipA-) revealed that expression of the orthogonal $\mathrm{tRNA}_{CUA}^{Tyr}$ alone was responsible for a moderate amount of GFP accumulation in cells (FL/OD signal), but that expression of the BipARS together with $\mathrm{tRNA}_{CUA}^{Tyr}$ resulted in nearly equivalent signal in BipA+ or BipA- cases (**Fig. 1C**). Expression of an N-terminally truncated yeast Ub cleavase protein (UBP1) (*25*, *26*) to expose the target residue at the N-terminus caused significant (~4-fold) reduction of the BipA- signal but no significant change in the BipA+ signal. The decrease in only BipA- signal supported our hypothesis that BipA would be N-end stabilizing. BipA- signal decreased further upon overexpression of ClpS, the adaptor protein that binds peptides containing primary N-end substrates (Y/F/W/L) and delivers them to the ClpAP AAA+ protease complex for unfolding and degradation (*27*), suggesting that the rate of N-end discrimination was previously limiting. ClpS overexpression may decrease substrate competition for proteins targeted for PTP because ClpS is known to inhibit other ClpA substrates such as SsrA-tagged proteins (*28*).

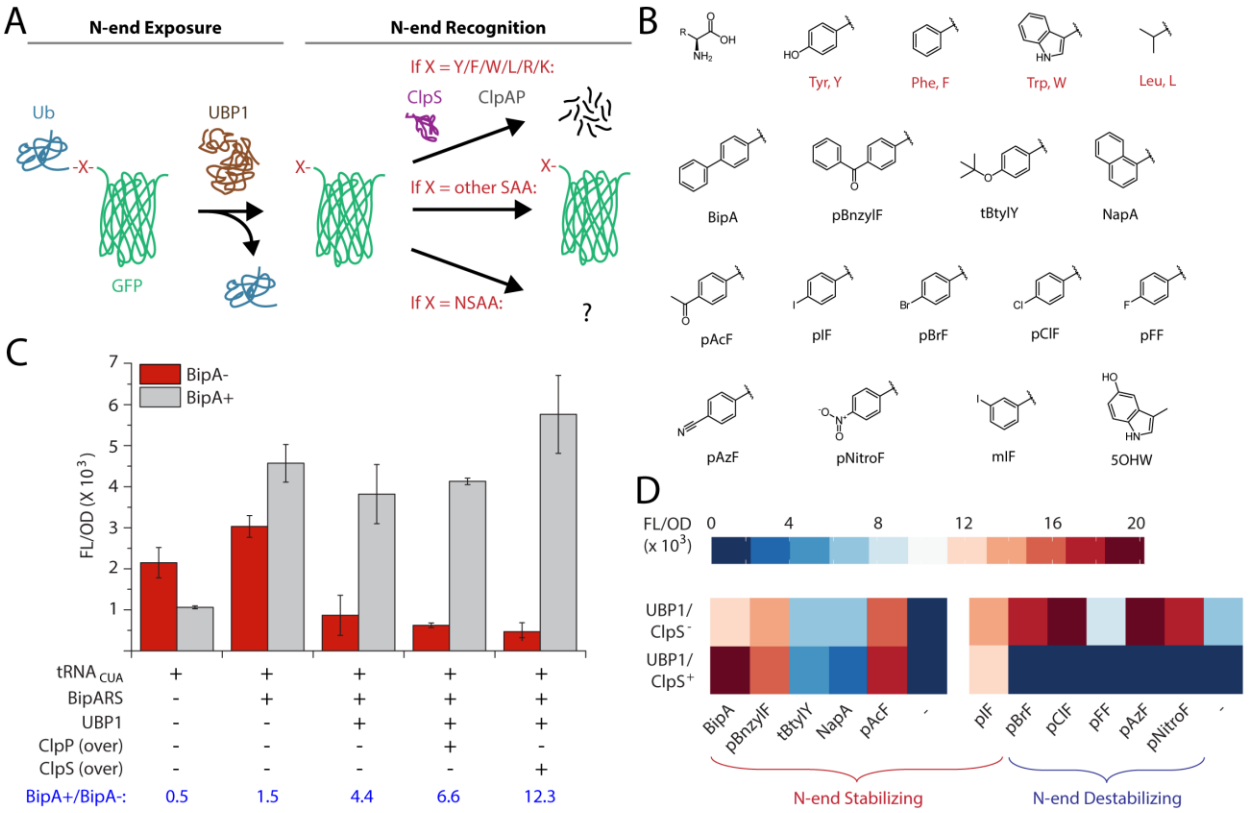

Fig. 1. Post-Translational Proofreading (PTP) proof of concept. (A) Scheme for PTP consisting of N-end exposure and recognition steps applied to synthetic substrates. Ubiquitin (Ub) is cleaved by ubiquitin cleavase UBP1 to expose the target site as N-terminal. ClpS is the native N-recognin in *E. coli* and ClpAP forms a AAA+ protease complex for N-end rule degradation. (B) NSAAs used in this study (full chemical names in SI). (C) Incorporation assay showing fluorescence resulting from GFP expression normalized by optical density (FL/OD) in the absence/presence of BipA and expression of various OTS or N-end rule components. "Over" indicates overexpression of natively expressed components. Error bars represent SD, N=3. (D) Heatmap of FL/OD signals obtained from an NSAA panel arranged roughly in descending size from left to right without PTP in top row and with PTP in bottom row. Left panel reflects activity of BipyA OTS and right reflects pAcF OTS. Heatmap values here and elsewhere are average of N=3.

To examine how the N-end rule applies to a larger set of NSAAs, we used the bipyridinylalanine (BipyA) OTS to screen 11 NSAAs because of its low NSAA- signal (**Fig. S2**). However, this OTS resulted in observable NSAA+ signal for only 5 out of 11 tested phenyl-NSAAs, with preference for large hydrophobic side chains at the para position of phenylalanine (**Fig. 1D**). Notably, NSAA+ signal for these 5 NSAAs was unaffected by PTP. The p-Acetyl-phenylalanine (pAcF) OTS was used to test incorporation of the 6 remaining NSAAs and appeared to broadly increase signal for these 6 NSAAs with PTP "Off" (ie., no expression of UBP1/ClpS). We observed marked differences in signal between PTP "Off" and "On" states based roughly on NSAA size. For p-Iodo-phenylalanine (pIF) and larger NSAAs, signal did not significantly change, and therefore pIF and larger NSAAs are N-end stabilizing. However, for p-Bromo-phenylalanine (pBrF) and other smaller or polar phenyl-NSAAs such as pAzF, signal was significantly diminished when PTP was "On" relative to when it was "Off". The data suggest that smaller deviations from Y/F are tolerated by the ClpS binding pocket, making smaller NSAAs such as pBrF and pAzF N-end destabilizing.

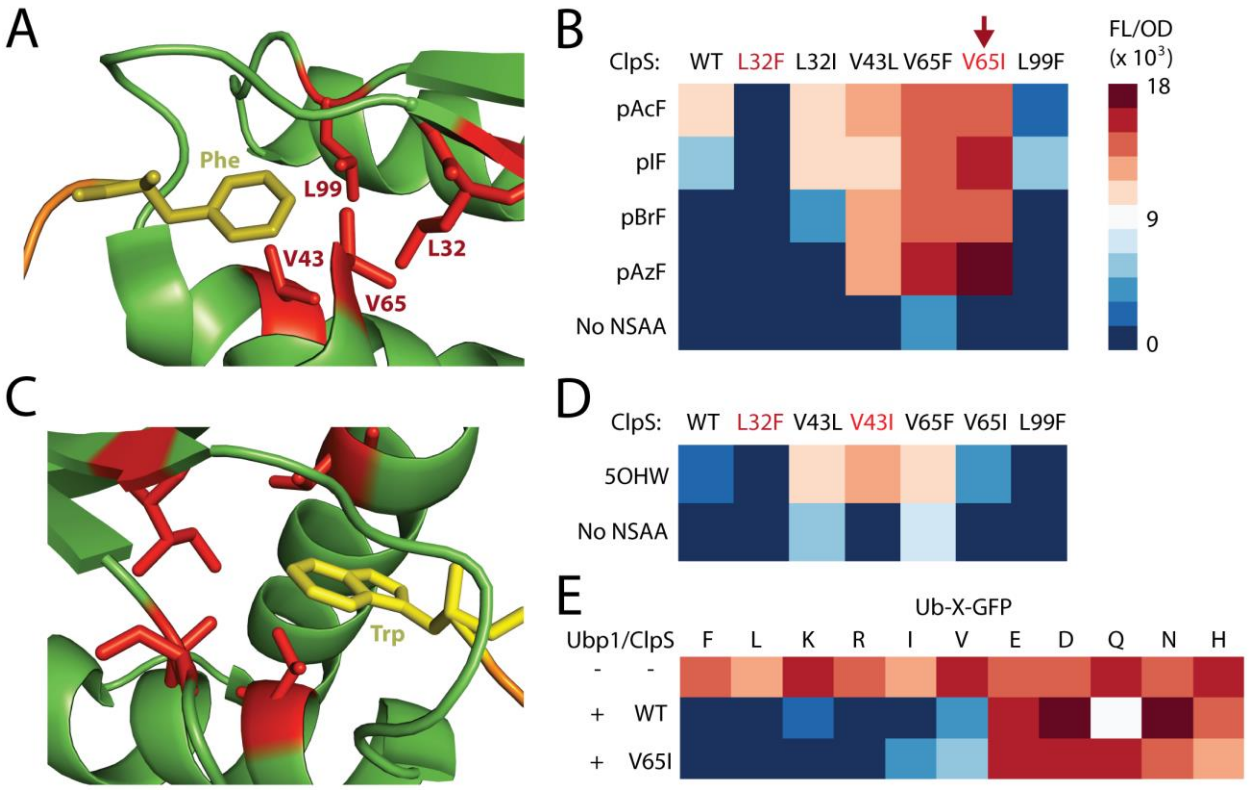

Fig. 2. PTP tunability achieved through rational ClpS engineering. (A) Cartoon generated from crystal structure of *E. coli* ClpS binding N-end Phe peptide (PDB ID: 3O2B) showing four hydrophobic ClpS residues subjected to mutation. (B) Heatmap of FL/OD signals obtained using a ClpS- host expressing UBP1, the pAcF OTS, and variants of ClpS in the presence of different NSAAs. (C) Cartoon generated from crystal structure of *C. crescentus* ClpS binding N-end Trp peptide (PDB ID: 3GQ1). (D) FL/OD heatmap resulting from expression of UBP1, the 5OHW OTS, and ClpS variants in the presence/absence of 5OHW. Scale as in panel B. (E) FL/OD heatmap resulting from expression of UBP1/ClpS in strains with Ub-X-GFP reporter genes expressing SAAs in place of X.

We hypothesized that we could engineer ClpS to alter N-end rule classification of these smaller NSAAs. We targeted four hydrophobic residues in the ClpS binding pocket for single point mutagenesis covering F/L/I/V (**Fig. 3A**). Sequence alignments of ClpS homologs across prokaryotes and eukaryotes showed conservation of these residues among related hydrophobic amino acids (**Fig. S3**). By screening the resulting 12 single mutants in a ClpS-deficient version of our reporter strain with select NSAAs and the pAcF OTS, we identified a variant ($ClpS^{V65I}$) that resulted in stabilization of all screened N-end phenyl NSAAs while still degrading SAAs (**Fig. 3B**). In addition, we identified a variant ($ClpS^{L32F}$) that resulted in complete degradation of all but the two largest screened N-end phenyl NSAAs (**Fig. S4**). We also attempted to distinguish tryptophanyl analogs from tryptophan using the 5OHW OTS (**Fig. 3C**). Although 5OHW is N-end destabilizing with ClpS, we observed that $ClpS^{V43I}$ and $ClpS^{V65I}$ improved discrimination of 5OHW from W (**Fig. 3D**). Given the desirable properties of $ClpS^{V65I}$, we wanted to examine whether it alters N-end rule classification for SAAs. We substituted the UAG codon in our GFP reporter for codons encoding a representative panel of SAAs and found that $ClpS^{V65I}$ affects stability of these N-end SAAs no differently than ClpS (**Fig. 3E**). Rational designs from our small library can precisely distinguish small modifications on a variety of chemical templates, such as NSAAs with phenyl as well as indole sidechains, showcasing the remarkably tunability of PTP. Interestingly, overexpression of either ClpS or $ClpS^{V65I}$ lead to degradation of N-end I/V, residues that are previously shown to be only weakly N-end destabilizing *in vitro* (*29*).

The ability of PTP to discriminate incorporation of intended NSAAs from related SAAs is useful for high-throughput screening of OTS libraries. To demonstrate this, we integrated the *UBP1-clpS^{V65I}* expression cassette into our ClpS-deficient reporter strain and used this strain to improve the selectivity of the parental ("WT") BipA OTS. Previous efforts to engineer MjTyrRS variants like BipARS focused on site-directed mutagenesis on positions near the amino acid binding pocket (*6, 30*). To generate a novel BipARS library, we used error-prone PCR to introduce 2-4 mutations throughout the *bipARS* gene. These libraries were transformed into C321.Nend and screened with three rounds of FACS sorting: (i) positive sort for GFP+ cells in BipA+; (ii) negative sort for GFP- cells in BipA-; (iii) final positive sort for GFP+ cells in BipA+ (**Fig. 4A**). To decrease promiscuity against other NSAAs, we altered the negative screening stringency by varying addition of undesired NSAAs (pAcF, pAzF, tBtylY, NapA, and/or pBnzylF), which changed the profile of isolated variants (**Fig. S5**). Retransformation of the 11 most enriched variants into C321.Ub-UAG-sfGFP (no PTP) showed that most of our variants increased BipA+ signal and decreased NSAA- signal compared to the WT OTS (**Fig. 4B** and **Table S1**, Variants 1-6). Supplementation with undesired NSAAs enriched for mutants with even greater selectivity against SAAs and undesired NSAAs (Variants 4, 9-11) but also gave rise to an extremely promiscuous variant (Variant 8), suggesting that these conditions may be nearly too harsh. One mutant only isolated in higher stringencies, Variant 10, exhibited high activity on BipA and no observable activity on any other NSAAs except tBtylY, whose structure is very similar to BipA and contains the inert tert-Butyl protecting group (**Fig. 4B**). SDS-PAGE of Ub-X-GFP resulting from the Variant 10 OTS after expression and affinity purification showed no observable BipA- protein production in contrast to WT BipA OTS, which shows a distinct BipA- band (**Fig. S6A**). Furthermore, mass spectrometry confirmed site-specific BipA incorporation (**Fig. S6B-D**).

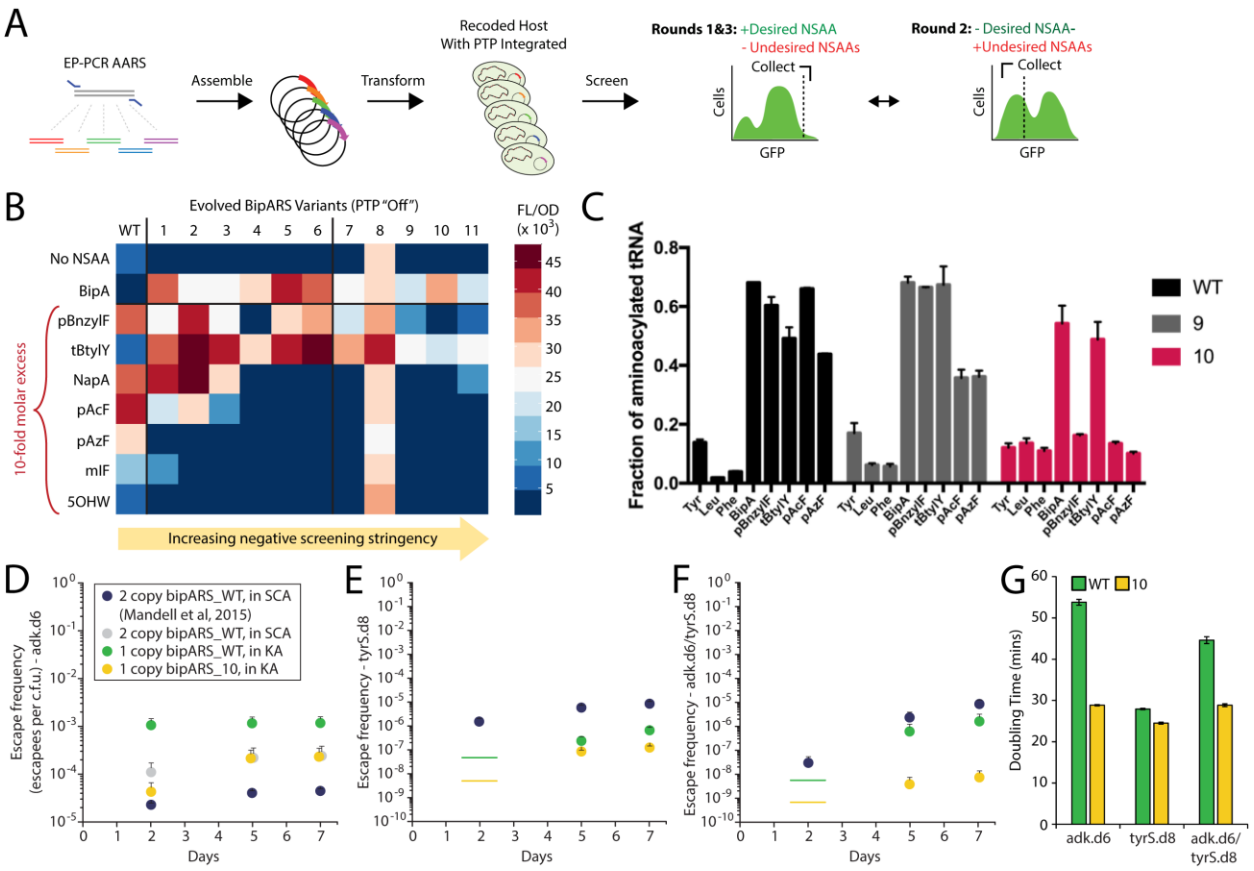

Fig. 3. Selective BipA OTS evolution using PTP. (A) FACS evolution scheme with EP-PCR AARS libraries transformed into hosts with PTP (using $ClpS^{V65I}$) genomically integrated before 3 sorting rounds. (B) Evaluation of most enriched evolved BipARS variants in clean backgrounds on a panel of NSAAs ([BipA] = 100 μM, [rest] = 1 mM, which are their standard concentrations). (C) *In vitro* amino acid substrate specificity profile of BipA OTS variants. (D) Escape frequencies over time for *adk.d6* strains transformed with constructs indicated in legend. Navy circles represent previously published data. Gray circles for adk.d6 represent repeat of previously published data. Green and yellow circles are the most relevant constructs to compare for this study. KA: Kanamycin+Arabinose. SCA: SDS+Chloramphenicol+Arabinose. Error bars in D-F represent SEM, N=3. (E) Escape frequencies over time for *tyrS.d8* strains. Lines represent assay detection limit in cases where no colonies were observed. (F) Escape frequencies over time for *adk.d6/tyrS.d8* strains. (G) Doubling time for biocontained strains with WT or Variant 10 OTS. Error bars = SD, N=3.

We discovered spontaneous tRNA mutations in our most selective variants, such as 4, 9, and 10, perhaps because of our use of a MutS-deficient host (**Fig. S7A**). When we reverted these tRNA mutations, each corresponding BipA OTS became more promiscuous (**Fig. S7B**), suggesting that observed tRNA mutations increase selectivity. The G51 position (G50 in *E. coli* nomenclature) mutated in tRNA Variant 10 is the most significant base pair in determining acylated tRNA binding affinity to elongation factor Tu (EF-Tu), which influences incorporation selectivity downstream of the AARS (*31*, *32*). To more rigorously assess OTS selectivity, we purified AARS and tRNA for the WT, Variant 9, and Variant 10 OTSs. The observed *in vitro* substrate specificity as determined by tRNA aminoacylation is in excellent agreement with our *in vivo* assays (**Fig. 4C**), and the data suggests that AARS and tRNA variants each contribute to selectivity improvements (**Fig. S7C-D**). The Variant 10 OTS exhibited the highest selectivity for BipA and was chosen for subsequent applications.

To demonstrate the utility of a more selective OTS for synthetic biocontainment, we substituted the WT BipA OTS construct previously used in three biocontained strains that exhibit observable escape frequencies with plasmids containing either the WT or Variant 10 OTS. These three biocontained strains (*adk.d6*, *tyrS.d8*, and *adk.d6/tyrS.d8*) harbor computational redesigns of two essential genes (*adk* and *tyrS*) to make their stability dependent on BipA (*33*). We monitored escape frequencies on non-permissive media for seven days and observed lower escape frequencies for strains containing the Variant 10 OTS at all measured time-points (**Fig. 4D-F**, **Fig. S8**). The difference in escape frequency was most apparent for the *adk.d6/tyrS.d8* strain, which exhibited a 7-day escape frequency of 7.4×10^{-9}, a value more than two orders of magnitude lower than observed for any C321.ΔA-derived strain containing only two altered genes. Furthermore, the fitness of all three strains improved with the Variant 10 OTS, with doubling time decreasing by nearly 2-fold (**Fig. 4G**). Finally, Variant 10 also delayed onset of growth of *adk.d6/tyrS.d8* on non-cognate NSAAs (**Table S2**). We expect these benefits to carry over to all strains which employ Variant 10 over WT OTS.

In addition to providing a new paradigm for OTS evaluation and evolution, PTP can be transformative for applications in which the identity of a single amino acid is critical, such as screening of natural synthetases for NSAA acceptance, sense codon reassignment, post-translational modifications, and for industrial uses where purity is extremely important, such as NSAA-containing biologics. PTP may also find use in translational regulation and as an orthogonal biocontainment strategy. Given that all 20 SAAs are known to be N-end destabilizing under certain contexts (*34*), conditionally expressed components could be transferred across organisms to dramatically alter the set of N-end destabilizing SAAs for a particular application.

References and Notes:

1. N. Ostrov *et al.*, *Science* **353**, 819 (2016).
2. M. Ibba, D. Söll, *Genes Dev.* **18**, 731 (2004).
3. J. W. Chin, *Annu. Rev. Biochem.* **83**, 379 (2014).
4. A. Dumas *et al.*, *Chem. Sci.* **6**, 50 (2015).
5. K. Oki *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 13298 (2008).
6. T. S. Young *et al.*, *J. Mol. Biol.* **395**, 361 (2010).
7. A. K. Antonczak *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1320–5 (2011).
8. S. Nehring *et al.*, *PLoS One*. **7**, e31992 (2012).
9. J. W. Monk *et al.*, *ACS Synth. Biol.* **6**, 45 (2016).
10. C. Fan *et al.*, *ChemBioChem*. **15**, 1805 (2014).
11. L.-T. Guo *et al.*, *Proc. Natl. Acad. Sci. U. S. A.* **111**, 16724 (2014).
12. Y.-S. Wang *et al.*, *Mol. Biosyst.* **7**, 714 (2011).
13. J. Xie *et al.*, *Angew. Chemie*. **119**, 9399 (2007).
14. D. B. F. Johnson *et al.*, *Nat Chem Biol*. **7**, 779 (2011).
15. P. O'Donoghue *et al.*, *FEBS Lett.* **586**, 3931 (2012).
16. A. Bachmair *et al.*, *Science* **234**, 179 (1986).
17. J. W. Tobias *et al.*, *Science* **254**, 1374 (1991).
18. T. Tasaki *et al.*, *Annu. Rev. Biochem.* **81**, 261 (2012).
19. C. B. Rosen, M. B. Francis, *Nat. Chem. Biol.* **13**, 697 (2017).
20. K. H. Wang *et al.*, *Genes Dev.* **21**, 403 (2007).
21. K. H. Wang *et al.*, *J. Biol. Chem.* **283**, 24600 (2008).
22. M. J. Lajoie *et al.*, *Science* **342**, 357 (2013).
23. M. Amiram *et al.*, *Nat Biotech*. **33**, 1272 (2015).
24. S. Million-Weaver *et al.*, in *Methods in molecular biology* (2012, vol. 834, pp. 33–48.)
25. J. W. Tobias, A. Varshavsky, *J. Biol. Chem.* **266**, 12021 (1991).
26. A. Wojtowicz *et al.*, *Microb. Cell Fact.* **4**, 1 (2005).
27. G. Román-Hernández *et al.*, *Mol. Cell*. **43**, 217 (2011).
28. D. A. Dougan *et al.*, *Mol. Cell*. **9**, 673 (2002).
29. K. H. Wang *et al.*, *Mol. Cell*. **32**, 406 (2008).
30. L. Wang *et al.*, *Science* **292**, 498 (2001).

Acknowledgements: We thank Dr. Daniel J. Mandell (Harvard), Dr. Irene M. B. Reizman (Rose-Hulman), and Bernardo Cervantes (MIT) for discussions. We thank Chad Araneo for FACS assistance and Dr. Bogdan Budnik for mass spectrometry assistance. We also thank Dr. Karl Schmitz (MIT) for manuscript comments. This project was graciously funded by the U.S. Dept. of Energy, grant DE-FG02-02ER63445. GMC has related financial interests in ReadCoor, EnEvolv, and GRO Biosciences. AMK and GMC have filed a provisional patent on PTP, and AMK/DS/EK/GMC have filed a provisional patent on evolved BipA OTS variants. For a complete list of GMC's financial interests, please visit http://arep.med.harvard.edu/gmc/tech.html. Author Contributions: AMK conceived the project and designed and performed most experiments. DAS generated EP-PCR AARS libraries, cloned AARS/tRNA combinations, purified reporter protein, and submitted samples for mass spectrometry. AMK, DAS, and EK screened AARS libraries. EK performed minimal media AA spiking experiments. OVR performed AARS and tRNA purification and biochemical

characterization of tRNA aminoacylation. MML generated strains containing N-end selectable markers. DS and GMC supervised experiments. AMK and DAS led manuscript writing and all authors reviewed before submission.

Supplementary Materials:

Materials and Methods

Supplementary Text

Figures S1-S11

Tables S1-S4

References (*31-51*)