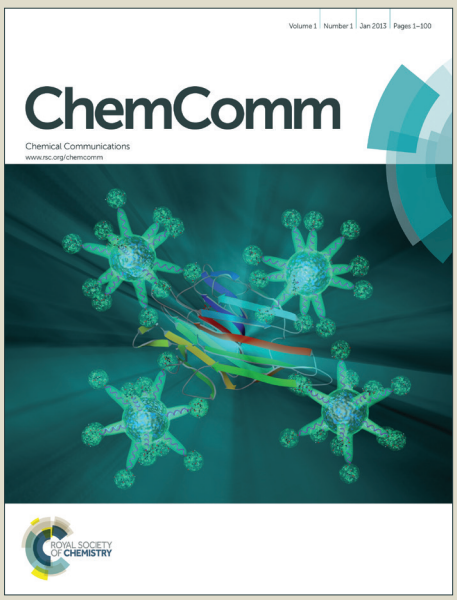
ChemComm

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/chemcomm

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/xxxxxx

ARTICLE TYPE

The genetic incorporation of thirteen novel non-canonical amino acids

Alfred Tuley,[a] Yane-Shih Wang,[a] Xinqiang Fang,[a,¶] Yadagiri Kurra,[a] Yohannes H. Reznom,[a] and Wenshe R. Liu*[a]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX

DOI: 10.1039/b000000x

Thirteen novel non-canonical amino acids were synthesized and tested for suppression of an amber codon using a mutant pyrrolysyl-tRNA synthetase-tRNA$^{Pyl}_{CUA}$ pair. Suppression was observed with varied efficiencies. One non-canonical amino acid in particular contains an azide that can be applied for site-selective protein labeling.

Site-selective installation of non-canonical amino acids (NCAAs) at an amber codon is an efficient approach to synthesize proteins with unique functionalities; applications span from basic studies such as protein cellular localization and protein-protein interaction analysis, to biotechnological applications such as the synthesis of heat stable enzymes and therapeutic protein manufacturing.[1-5] Two aminoacyl-tRNA synthetase-tRNA$_{CUA}$ pairs have been well adapted for the genetic incorporation of NCAAs at amber codons in bacteria. One is the tyrosyl-tRNA synthetase-tRNA$^{Tyr}_{CUA}$ pair that was derived from *Methanocaldococcus jannaschii*.[6-8] The other is the pyrrolysyl-tRNA synthetase (PylRS)-tRNA$^{Pyl}_{CUA}$ pair that naturally occurs in some methanogenic archaea.[9-12] Due to its broad-spectrum orthogonality from bacteria to human cells and the fact that it can be easily engineered to target a large variety of NCAAs, including natural amino acids with posttranslational modifications, the PylRS-tRNA$^{Pyl}_{CUA}$ pair has captivated researchers in the past several years.[13-28] One of our major contributions to the NCAA research field has been the development of PylRS mutants capable of incorporating of a number of phenylalanine derivatives, which are substantially different from the structure of pyrrolysine, the native substrate of PylRS.[29-31] More specifically, we have recently shown that a rationally designed, N346A/C348A mutant of PylRS (PylRS(N346A/C348A)) is capable of incorporating seven *para*- and twelve *meta*-substituted phenylalanine derivatives at amber codons in coordination with tRNA$^{Pyl}_{CUA}$.[30, 31] This broad substrate scope obviates the need to undergo the arduous task of discovering a new mutant for each NCAA. Herein we demonstrate that PylRS(N346A/C348A) has an even broader substrate scope than previously reported.

Our previous studies revealed a large active site pocket in PylRS(N346A/C348A).[30] Removal of the N346 side chain amide dismisses the steric clash that prevents the binding of the aromatic side chain of phenylalanine and the loss of the C348 thiol yields a cavernous pocket capable of binding the *para*- or *meta*-substituted phenylalanine described above. Interestingly, although phenylalanine derivatives with small *para*-substituents have shown to be ineffective substrates for PylRS(N346A/C348A), their isomers with *meta*-substituents act as highly efficient substrates of PylRS(N346A/C348A) for their genetic incorporation at amber codons.[30, 31] In other words, phenylalanine derivatives with *para*-substituents can only be incorporated when they possess large side chains. Upon further inspection, it appears that a majority of the vacancy in the active site pocket of PylRS(N346A/C348A) exists near the *meta* position of phenylalanine. Encouraged by our preliminary work, we reasoned that PylRS(N346A/C348A) could incorporate phenylalanine derivatives with more sterically demanding side chains.

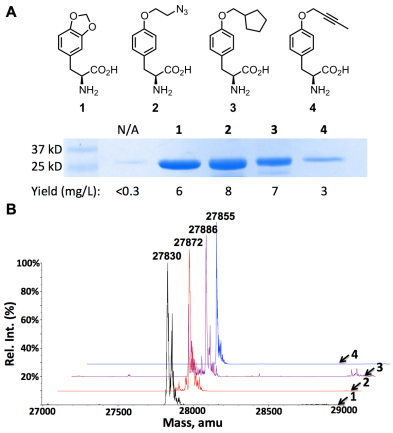


Figure 1. (**A**) Structures of **1**-**4** and their site-specific incorporation into sfGFP at its S2 position. (**B**) Deconvoluted ESI-MS spectra of sfGFP variants incorporated with **1**-**4**. Their theoretical values are 27,832 Da for **1**, 27,873 Da for **2**, 27,886 Da for **3**, and 27,856 Da for **4**. Satellite signals are largely due to metal ion adducts (i.e. Li, Na, K).

Our investigation began with the synthesis and genetic incorporation of four different *para*-substituted phenylalanine derivatives (**1**-**4** in **Figure 1A**), each with a unique functionality and steric requirement. Synthesis of these derivatives followed the same strategy presented in one of our previous reports of the N346A/C348A mutant,[30] with the exception of NCAA **1**, which was synthesized using a different approach (see the supporting information). These four NCAAs were then tested for their tolerability by PylRS(N346A/C348A). An *E. coli* BL21(DE3) cell that harbours two plasmids, pEVOL-pylT-PylRSN346A/C348A and pET-pylT-sfGFP2TAG, was employed for the investigation. pEVOL-pylT-PylRSN346A/C348A contains genes coding PylRS(N346A/C348A) and tRNA$^{Pyl}_{CUA}$; pET-pylT-sfGFP2TAG carries a tRNA$^{Pyl}_{CUA}$ coding gene and a non-sequence-optimized superfolder green fluorescent protein (sfGFP) gene with an amber mutation in position S2 (sfGFP2TAG). The same cells were used in the initial test of the recognition of *para*-substituted phenylalanine derivatives by PylRS(N346A/C348A).[30] Growth in minimal media supplemented with 1 mM IPTG and 0.2% arabinose without

NCAA afforded a minimal expression level of full-length sfGFP (<0.3 mg/L). Providing any of **1-4** at 2 mM into the medium all promoted full-length sfGFP expression (**Figure 1A**). The expression levels for **1-3** are comparable to that for *para*-propargyloxy-phenylalanine (7.8 mg/L),[31] and the electrospray ionization mass spectrometry analysis of four purified sfGFP variants displayed molecular weights that agreed well with their theoretical values corresponding to full-length proteins with the first methionine (**Figure 1B**).

Our data for **1-4** demonstrates that PylRS(N346A/C348A) tolerates phenylalanine derivative substrates with rigid and bulky substituents at the *para* position. However, ESI-MS data for compound **3** also shows a small side peak corresponding to phenylalanine incorporation, a result we have observed previously. The remaining satellite peaks for these compounds are common metal adducts in ESI-MS. Additionally, results for **1** demonstrated that both *meta* and *para* positions can be occupied at no detriment to expression levels. These results, coupled with our previous endeavours, led us to wonder if phenylalanine derivatives with long-chain *meta*-substituents could serve as substrates of PylRS(N346A/C348A) for genetic incorporation as well. To investigate this hypothesis, a series of *meta*-alkoxy and *meta*-acyl phenylalanines with substituent chain lengths of up to six carbons were synthesized. We chose these specific derivatives because the parent NCAAs *meta*-methoxy-phenylalanine and *meta*-acetyl-phenylalanine act as efficient substrates for PylRS(N346A/C348A). The synthesis of *meta*-alkoxy-phenylalanines was straightforward, starting with a published route to obtain protected *meta*-tyorsine, at which point the intermediate was subjected to various alkyl halides to afford different derivatives. Acidic deprotection then yielded free amino acids as racemic chloride salts. The synthesis of *meta*-acyl-phenylalanines was more divergent. Alkyl grignards were added to a solution of *meta*-tolunitrile, which afforded acylbenzenes upon acidic workup. Radical bromination and then displacement with diethylacetamidomalonate afforded protected *meta*-acyl-phenylalanines that were deprotected in 6 M HCl to obtain free amino acids. More detailed synthetic routes can be found in the supporting information.

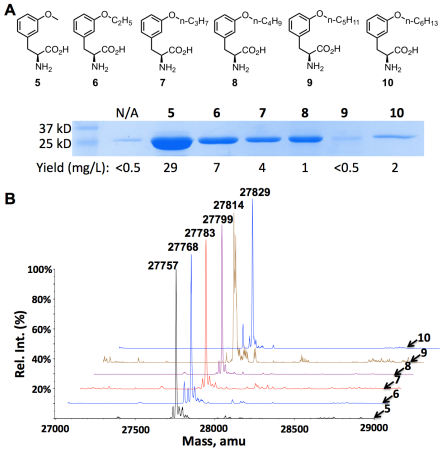


Figure 2. (**A**) Structures of **5-10** and their site-specific incorporation into sfGFP at its S2 position. (**B**) Deconvoluted ESI-MS spectra of sfGFP variants incorporated with **5-10**. Their theoretical values are 27,758 Da for **5**, 27,772 Da for **6**, 27,786 Da for **7**, 27,800 Da for **8**, 27,814 Da for **9**, and 27,828 Da for **10**.

With the desired NCAAs in hand, we thenceforth tested their incorporation efficacies at amber codons using the PylRS(N346A/C48A)- $tRNA^{Pyl}_{CUA}$ pair. The *E. coli* cells used for these compounds harboured two plasmids, pEVOL-pylT-PylRSN346A/C348A and pET-pylT-sfGFPS2TAG'. pET-pylT-sfGFPS2TAG' contains a sequence-optimized sfGFP with an amber mutation at its S2 position (sfGFPS2TAG'). In comparison to the sfGFP2TAG gene in pET-pylT-sfGFP2TAG, sfGFPS2TAG' has one more alanine residue in front of the amber mutation. Growing this cell in minimal media without NCAA yielded a minimal expression level of full-length sfGFP. However, all ether NCAAs **6-10** (2 mM) in the medium promoted the synthesis of sfGFP with a designated NCAA incorporated (**Figure 2A**). In comparison to phenylalanine derivatives with small *meta*-substituents such as **5**, **6-10** apparently have low incorporation levels. Molecular weights of purified sfGFP variants determined by ESI-MS agreed well with their theoretical values corresponding to a designated NCAA at the S2 position and the first methionine hydrolysed (**Figure 2B**). The removal of the first methionine is due to the insertion of alanine after it. A number of smaller signals can be observed, but they are largely common metal adducts; the expected masses were always the major signal. Compounds **8**, **9** and **10** have low solubility; when provided in the medium at 2 mM, compound **10** was observed to precipitate after 12 h expression. The low sfGFP expression levels for **8**, **9** and **10** may be partially due to the toxicity of the compounds; indeed, smaller pellet sizes are observed for **8** and **9**. Although the sfGFP expression level for **9** was very low, the purified sfGFP displayed an ESI-MS molecular weight that still matched the theoretical value of sfGFP with **9** incorporated at S2, indicating a low concentration of **9** was still sufficient to observe incorporation of **9** at the amber mutation site.

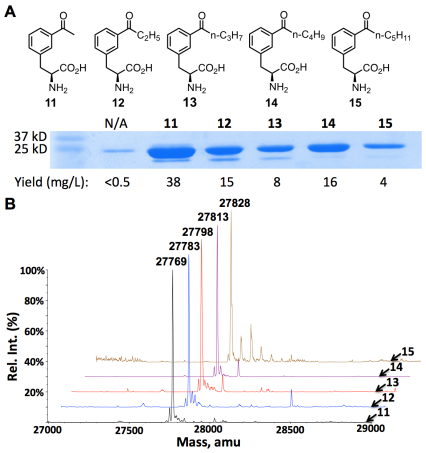


Figure 3. (**A**) Structures of **11-15** and their site-specific incorporation into sfGFP at its S2 position. (**B**) Deconvoluted ESI-MS spectra of sfGFP variants incorporated with **11-15**. Their theoretical values are 27,770 Da for **11**, 27,784 Da for **12**, 27,798 Da for **13**, 27,812 Da for **14**, and 27,826 Da for **15**. Compounds **13** and **14** show small signals corresponding to an N-terminal methionine on sfGFP. Compound **15** has several small signals attributed to sodium and potassium adducts.

Overall, providing ketone derivatives **12-15** at 2 mM in the medium promoted high sfGFP expression yields, and longer alkyl lengths had less of an impact on protein yields in comparison to the ether series **6-10**, though the sfGFP expression levels for **12-15** are lower than that for **11** (**Figure 3A**). This series of NCAAs are also readily soluble, with no precipitation observed in the medium after overnight incubation. ESI-MS analysis of the purified sfGFP variants confirmed high incorporation fidelities of **12-15** at the S2 site.

Among all of the novel NCAAs that can be taken by PylRS(N346A/C348A), **2** has an active azide functionality for click reaction with an alkyne[32] and **12-15** contain a ketone group that potentially reacts with a hydroxylamine. Both functionalities

can be applied for site-selective labeling of proteins incorporated with **2** and **12-15**. Since labeling of sfGFP incorporated with **11** with a hydroxylamine dye was demonstrated previously,[31] we chose to demonstrate the selective labeling of **2** using a diarylcyclooctyne dye **D1** in this study. **D1** contains a strained alkyne that undergoes spontaneous reaction with an azide.[33] Incubating sfGFP incorporated with **2** with **D1** overnight led to an intensely fluorescently labeled protein; however, the same reaction with sfGFP incorporated with **3** did not yield any fluorescently labeled final product. This result indicates that genetically incorporated **2** can be applied to site-specifically introduce biophysical and biochemical probes to proteins for a large variety of studies.

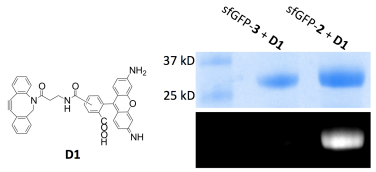


Figure 4. (**A**) Labeling of sfGFP incorporated with **2** (sfGFP-**2**) and sfGFP incorporated with **3** (sfGFP-**3**) with a dye **D1**. The top panel shows the Coomassie blue stained SDS-PAGE gel and the bottom panel shows the fluorescent image of the same gel under UV irradiation before the gel was stained with Coomassie blue.

In summary, we showed thirteen novel NCAAs were genetically incorporated into protein at amber codon in *E. coli* using the PylRS(N346A/C348A)-tRNA$^{Pyl}_{CUA}$ pair. This result, coupled with our previous findings, shows a surprisingly broad substrate scope for PylRS(N346A/C348A). Investigations are underway to determine aspects of the active site pocket of PylRS(N346A/C348A) that lead to this broad substrate spectrum. The current study has great implications in understanding amino acid structure tolerance of the protein translation system. The expanded genetically encoded NCAA pool can also be applied to generate phage and *E. coli* displayed peptide libraries with expanded chemical moieties for drug discovery, a direction we are actively pursuing at the current stage.

Notes and references

[a] Department of Chemistry, Texas A&M University, College Station, TX 77843; wliu@chem.tamu.edu.
¶ Current address: Fujian Institute of Research on the Structure of Matter, Chinese Academy of Sciences, Fuzhou, Fujian 350002, China
Electronic Supplementary Information (ESI) available: Synthesis, protein expression, protein labeling, and mass spectrometry analysis. See DOI: 10.1039/c000000x/

1. W. R. Liu, Y. S. Wang and W. Wan, *Mol Biosyst*, 2011, 7, 38-47.
2. A. Gautier, A. Deiters and J. W. Chin, *J Am Chem Soc*, 2011, 133, 2124-2127.
3. M. Zhang, S. Lin, X. Song, J. Liu, Y. Fu, X. Ge, X. Fu, Z. Chang and P. R. Chen, *Nat Chem Biol*, 2011, 7, 671-677.
4. Y. Tang, G. Ghirlanda, N. Vaidehi, J. Kua, D. T. Mainz, I. W. Goddard, W. F. DeGrado and D. A. Tirrell, *Biochemistry*, 2001, 40, 2790-2796.
5. C. C. Liu and P. G. Schultz, *Annu Rev Biochem*, 2010, 79, 413-444.
6. L. Wang, A. Brock, B. Herberich and P. G. Schultz, *Science*, 2001, 292, 498-500.
7. J. Xie and P. G. Schultz, *Nat Rev Mol Cell Biol*, 2006, 7, 775-782.
8. J. W. Chin, S. W. Santoro, A. B. Martin, D. S. King, L. Wang and P. G. Schultz, *J Am Chem Soc*, 2002, 124, 9026-9027.
9. G. Srinivasan, C. M. James and J. A. Krzycki, *Science*, 2002, 296, 1459-1462.
10. S. K. Blight, R. C. Larue, A. Mahapatra, D. G. Longstaff, E. Chang, G. Zhao, P. T. Kang, K. B. Green-Church, M. K. Chan and J. A. Krzycki, *Nature*, 2004, 431, 333-335.
11. H. Neumann, S. Y. Peak-Chew and J. W. Chin, *Nat Chem Biol*, 2008, 4, 232-234.
12. W. Wan, Y. Huang, Z. Wang, W. K. Russell, P. J. Pai, D. H. Russell and W. R. Liu, *Angew Chem Int Ed Engl*, 2010, 49, 3211-3214.
13. S. Greiss and J. W. Chin, *J Am Chem Soc*, 2011, 133, 14196-14199.
14. S. M. Hancock, R. Uprety, A. Deiters and J. W. Chin, *J Am Chem Soc*, 2010, 132, 14819-14824.
15. T. Mukai, T. Kobayashi, N. Hino, T. Yanagisawa, K. Sakamoto and S. Yokoyama, *Biochem Biophys Res Commun*, 2008, 371, 818-822.
16. T. Yanagisawa, R. Ishii, R. Fukunaga, T. Kobayashi, K. Sakamoto and S. Yokoyama, *Chem Biol*, 2008, 15, 1187-1197.
17. A. R. Parrish, X. She, Z. Xiang, I. Coin, Z. Shen, S. P. Briggs, A. Dillin and L. Wang, *ACS Chem Biol*, 2012, 7, 1292-1302.
18. P. R. Chen, D. Groff, J. Guo, W. Ou, S. Cellitti, B. H. Geierstanger and P. G. Schultz, *Angew Chem Int Ed Engl*, 2009, 48, 4052-4055.
19. C. J. Chou, R. Uprety, L. Davis, J. W. Chin and A. Deiters, *Chem Sci*, 2011, 2, 480-483.
20. Y. S. Wang, B. Wu, Z. Wang, Y. Huang, W. Wan, W. K. Russell, P. J. Pai, Y. N. Moe, D. H. Russell and W. R. Liu, *Mol Biosyst*, 2010, 6, 1557-1560.
21. Y. J. Lee, B. Wu, J. E. Raymond, Y. Zeng, X. Fang, K. L. Wooley and W. R. Liu, *ACS Chem Biol*, 2013, 8, 1664-1670.
22. T. Fekner, X. Li, M. M. Lee and M. K. Chan, *Angew Chem Int Ed Engl*, 2009, 48, 1633-1635.
23. X. Li, T. Fekner, J. J. Ottesen and M. K. Chan, *Angew Chem Int Ed Engl*, 2009, 48, 9184-9187.
24. T. Umehara, J. Kim, S. Lee, L. T. Guo, D. Soll and H. S. Park, *FEBS Lett*, 2012, 586, 729-733.
25. C. R. Polycarpo, S. Herring, A. Berube, J. L. Wood, D. Soll and A. Ambrogelly, *FEBS Lett*, 2006, 580, 6695-6700.
26. T. Plass, S. Milles, C. Koehler, C. Schultz and E. A. Lemke, *Angew Chem Int Ed Engl*, 2011, 50, 3878-3881.
27. T. Plass, S. Milles, C. Koehler, J. Szymanski, R. Mueller, M. Wiessler, C. Schultz and E. A. Lemke, *Angew Chem Int Ed Engl*, 2012, 51, 4166-4170.
28. D. P. Nguyen, H. Lusic, H. Neumann, P. B. Kapadnis, A. Deiters and J. W. Chin, *J Am Chem Soc*, 2009, 131, 8720-8721.
29. Y. S. Wang, W. K. Russell, Z. Wang, W. Wan, L. E. Dodd, P. J. Pai, D. H. Russell and W. R. Liu, *Mol Biosyst*, 2011, 7, 714-717.
30. Y. S. Wang, X. Fang, A. L. Wallace, B. Wu and W. R. Liu, *J Am Chem Soc*, 2012, 134, 2950-2953.
31. Y. S. Wang, X. Fang, H. Y. Chen, B. Wu, Z. U. Wang, C. Hilty and W. R. Liu, *ACS Chem Biol*, 2013, 8, 405-415.
32. H. C. Kolb, M. G. Finn and K. B. Sharpless, *Angew Chem Int Ed Engl*, 2001, 40, 2004-2021.
33. J. C. Jewett, E. M. Sletten and C. R. Bertozzi, *J Am Chem Soc*, 2010, 132, 3688-3690.