Published in final edited form as:

Curr Opin Chem Biol. 2018 October ; 46: 196–202. doi:10.1016/j.cbpa.2018.08.009.

Expansion of the Genetic Code via Expansion of the Genetic Alphabet

Vivian T. Dien[#], Sydney E. Morris[#], Rebekah J. Karadeema[#], and Floyd E. Romesberg
Department of Chemistry, The Scripps Research Institute, La Jolla, CA 92037 USA

[#] These authors contributed equally to this work.

Abstract

Current methods to expand the genetic code enable site-specific incorporation of non-canonical amino acids (ncAAs) into proteins in eukaryotic and prokaryotic cells. However, current methods are limited by the number of codons possible, their orthogonality, and possibly their effects on protein synthesis and folding. An alternative approach relies on unnatural base pairs to create a virtually unlimited number of genuinely new codons that are efficiently translated and highly orthogonal because they direct ncAA incorporation using forces other than the complementary hydrogen bonds employed by their natural counterparts. This review outlines progress and achievements made towards developing a functional unnatural base pair and its use to generate semi-synthetic organisms with an expanded genetic alphabet that serves as the basis of an expanded genetic code.

Graphical Abstract

Introduction.

Biological diversity allows life to adapt to different environments, and over time, evolve new forms and functions. The source of this diversity is the variation within protein sequences provided by the twenty natural amino acids, variation that is encoded in an organism's genome by the four natural DNA nucleotides. While the functional diversity provided by the natural amino acids may be high, the vastness of sequence space dramatically limits what might actually be explored, and moreover, some functionality is simply not available. Nature's use of cofactors for hydride transfer, redox activity, and electrophilic bond

Corresponding author: Romesberg, Floyd E. (floyd@scripps.edu).

Conflict of interest statement

Patent applications have been filed by Synthorx and The Scripps Research Institute covering the UBPs and their use to produce proteins containing ncAAs. F.E.R. also has shares in Synthorx, Inc., a company that has commercial interests in the UBP.

formation, etc. attests to these limitations. Furthermore, with the increasing focus on developing proteins as therapeutics [1], these limitations are problematic, as the physiochemical diversity of the natural amino acids is dramatically restricted compared to that of the small molecule drugs designed by chemists. In principle, it should be possible to circumvent these limitations by expanding the genetic code to include additional, non-canonical amino acids (ncAAs) with desired physiochemical properties.

Almost 20 years ago, Peter Schultz increased the diversity available to living organisms by expanding the genetic code using the amber stop codon (UAG) to encode ncAAs in *Escherichia coli* [•2,•3]. This landmark accomplishment was achieved using a tRNA–amino acid tRNA synthetase (aaRS) pair from *Methanococcus jannaschii*, in which the tRNA was recoded to suppress the stop codon and the aaRS was evolved to charge the tRNA with an ncAA. This method of codon suppression has since been expanded to the other stop codons [4] and even quadruplet codons [5], as well as to the use of several other orthogonal tRNA–aaRS pairs (most notably the Pyl tRNA–synthetase pair from *Methanosarcina barkeri/mazei* [6,7,•8,9]), broadening the scope of ncAAs that may be incorporated into proteins. These methods have already begun to revolutionize both chemical biology [10–12] and protein therapeutics [13].

Though these methods enable incorporation of up to two, different ncAAs in both prokaryotic [14] and eukaryotic [15] cells, the heterologous recoded tRNAs must compete with endogenous release factors (RFs), or in the case of quadruplet codons, normal decoding [16], which limits the efficiency and fidelity of ncAA incorporation. To eliminate competition with RF1, which recognizes the amber stop codon and terminates translation, efforts have been directed toward removal of many or all instances of the amber stop codon in the host genome [17,18] or modification of RF2 [19] to allow for the deletion of RF1. However, eukaryotes have only one release factor, and while it may be modified [20] it cannot be deleted, and with prokaryotes, deletion of RF1 results in greater mis-suppression of the amber stop codon by other tRNAs, which reduces the fidelity of ncAA incorporation [21]. Though Herculean efforts to further exploit codon redundancy to liberate natural codons for reassignment to ncAAs are underway [22], codon reassignment may be complicated by pleiotropic effects, as codons are not truly redundant, for example due to their effects on the rate of translation and protein folding [23]. In addition, codon reappropriation is limited by the challenges of large-scale genome engineering, especially for eukaryotes [24].

An alternative approach to natural codon reassignment is the creation of entirely new codons that are free of any natural function or constraint, and whose recognition at the ribosome is inherently more orthogonal. This may be accomplished through the creation of organisms that harbor a fifth and sixth nucleotide that form an unnatural base pair (UBP). Such semi-synthetic organisms (SSOs) would need to faithfully replicate DNA containing the UBP, efficiently transcribe it into mRNA and tRNA containing the unnatural nucleotides, and then efficiently decode unnatural codons with cognate unnatural anticodons. Such SSOs would have a virtually unlimited number of new codons to encode ncAAs.

Development of a functional UBP.

The first challenge in developing an expanded genetic alphabet is the identification of unnatural nucleotides that selectively pair in duplex DNA and during replication by a DNA polymerase. While the Benner group (who has contributed to this special issue) has approached this challenge with synthetic nucleotides that pair via hydrogen bonding (H-bonding) patterns that are orthogonal to the natural base pairs [25], our group, as well as the Hirao group (who has also contributed to this special issue), took inspiration from the demonstration by the Kool group that hydrogen bonds are not required for DNA polymerase-mediated insertion of a triphosphate [26,27]. While the Hirao group has pursued shape analogs of the natural nucleobases [28], we have pursued the *de novo* development of a UBP, starting from simple aromatic nucleobase analogs with the ability to pair via hydrophobic and packing forces. We have taken a medicinal chemistry-like approach to guide the exploration of nearly 200 analogs, with design and analysis proceeding iteratively via the generation of structure-activity relationships (SARs).

These medicinal chemistry-like efforts have been reviewed [•29] and they ultimately culminated in the discovery of a family of UBPs that were well retained during PCR amplification [30,31]. Several key and interesting SARs are of note, but perhaps the most important is related to the nucleobase substituent *ortho* to the glycosidic bond. The data clearly demonstrate that efficient unnatural triphosphate insertion and continued extension of the growing strand of DNA require an *ortho* substituent that is capable of both hydrophobic packing and accepting an H-bond. This apparent physiochemical contradiction appears to have been resolved by the thioamide and methoxy substituents, as found in the dNaM-d5SICS and dNaM-dTPT3 UBPs (Fig. 1a). The polarizability of the sulfur increases its ability to pack, relative to an oxygen, without ablating its ability to accept an H-bond, and the methoxy substituent, via simple bond rotation, can present either a hydrophobic methyl group or a more hydrophilic ether oxygen.

In addition to the nature of the *ortho* substituent, SAR data revealed that nucleobase aromatic surface area extending into the developing major groove favors both triphosphate insertion and extension, presumably through optimized packing interactions (Fig. 2). However, careful optimization was required to favor packing between the unnatural nucleobase and the nucleobase of the primer terminus, as opposed to packing with the unnatural and natural nucleobases in the template strand via cross-strand intercalation [32,33]. In free duplex DNA, the UBPs adopt a cross-strand intercalated structure, but remarkably, in the polymerase active site when the unnatural triphosphate is paired opposite its cognate unnatural nucleotide in the template, a Watson-Crick-like structure is adopted [34,35]. Thus, DNA containing the UBP is replicated with a mutually induced-fit mechanism, wherein formation of the UBP drives the same conformational change required to form the closed polymerase complex that is induced by the formation of a natural base pair, and the tightly packed environment of the closed polymerase drives the UBP to adopt the required (de-intercalated) structure. However, after the polymerase translocates to position the next templating nucleotide in the active site, the nascent UBP again cross-stand intercalates, which distorts the primer terminus and mandates de-intercalation for efficient continued primer elongation. Using this structural data as a guide, we contracted and

derivatized the d5SICS nucleobase with a sulfur atom to yield dTPT3, which appears to favor de-intercalation and intrastrand packing and which thus proved to be a more optimal partner for dNaM. Although it was never the focus of the SAR studies, we also examined the transcription of DNA containing the UBPs to ensure that optimization for replication did not compromise transcription into RNA. Indeed, the ultimately identified family of UBPs is efficiently transcribed *in vitro* by T7 RNA polymerase [33,36].

While chemical optimization of the UBP continues, the identification of a family of UBPs that when incorporated into DNA are well replicated and transcribed represents a landmark in our efforts to expand the genetic alphabet and code. It clearly demonstrates that hydrophobic and packing forces can be optimized to bestow a single pair of nucleotides with the ability to participate in replication and transcription, at least *in vitro*. From a practical perspective, we concluded that the identified family of UBPs was sufficiently promising to initiate efforts toward using them as the foundation of an SSO with an expanded genetic alphabet.

An SSO with an expanded genetic alphabet.

E. coli was an obvious target for the creation of the first SSO, being an extensively studied organism with vast resources and tools for genetic manipulation. To overcome the crucial obstacle of making unnatural triphosphates available for replication within the SSO, we turned to the plasmids of *Phaeodactylum tricornutum* that use nucleoside triphosphate transporters (NTTs) to import triphosphates from the cytosol. Expression of the transporter PNTT2 enables uptake of a wide variety of natural and modified deoxy- and ribonucleotide triphosphates into the SSO [•37,38], including dNaMTP, d5SICSTP, and dTPT3TP. Thus, the first SSO (Fig. 3a) was generated by transformation of *E. coli* expressing PNTT2 with a plasmid harboring a single UBP. This SSO retained the corresponding UBP over many cell divisions during growth in media supplemented with dNaMTP and d5SICSTP [•37].

While this was the first demonstration of life with an additional base pair, the expression of PNTT2 resulted in slow growth. In its native host, the 65 N-terminal amino acids of PNTT2 are likely absent in the mature protein [39], and we suspected that they are not removed in *E. coli*. Expression of a variant lacking the N-terminal sequence eliminated the majority of toxicity without compromising activity, and we integrated this variant into the SSO's chromosome under the control of a constitutive promoter [•40]. These modifications dramatically improved the SSO's growth and increased retention of the dNaM-d5SICS UBP, likely due to the elimination of plasmid copy number fluctuations that would occasionally force the original SSO to replicate the UBP with insufficient levels of the unnatural triphosphates. However, unlike with the *in vitro* studies described above, we observed that the fidelity of dNaM-d5SICS replication is sequence dependent *in vivo*, with some sequences showing no UBP retention, even after only limited growth. To address this problem, we examined the replication of the UBP dNaM-dTPT3, which replicates with higher fidelity than dNaM-d5SICS *in vitro* [30]. The combination of our genetically optimized SSO with a more chemically optimized UBP resulted in dramatically improved retention with reduced sequence context bias; for example, five sequences that showed no retention with dNaM-d5SICS showed virtually no loss with dNaM-dTPT3. However, UBP

retention still remained imperfect in some sequence contexts, and complete loss was still observed in both the **CXG** and **CXA** (**X**=**dNaM**) sequence contexts.

Because the *in vitro* environment used to evaluate and optimize the UBPs during their development is significantly different from the *in vivo* environment of the SSO, we examined the *in vivo* performance of UBPs formed by 135 analogs that had previously only been evaluated *in vitro* [41]. Interestingly, while no derivatives of **dTPT3** were found to be more optimal, we found four **dNaM** analogs that were significantly better partners for **dTPT3** than **dNaM** itself, despite being less optimal partners *in vitro*. In particular, the **dCNMO-dTPT3** UBP (Fig. 1b) showed virtually no loss in the sequences that only moderately retained **dNaM-dTPT3**, and was even moderately well retained in the particularly problematic **CXG** sequence context, in which virtually no retention of **dNaM-dTPT3** is observed. These results clearly demonstrate the utility of exploring UBP replication in the *in vivo* environment where they must ultimately function, and suggest that further optimization of the UBP is likely still possible.

Along with the continued chemical optimization of the UBP, we have also continued to explore how the SSO itself might be further optimized. These efforts have resulted in the development of two strategies: one based on error elimination and one based on error avoidance. The error elimination strategy relies on the use of Cas9 from *Streptococcus pyogenes* programmed with guide RNAs that target and cleave DNA that has lost the UBP [40]. SSOs benefiting from this error elimination mechanism retained the **dNaM-dTPT3** UBP with virtually no loss in fourteen out of the sixteen possible sequence contexts, and even retained two UBPs with virtually no loss over a demanding growth regimen that included more than 120 doublings during liquid growth. Despite *SpCas9*'s propensity for poor single nucleotide mismatch discrimination with natural nucleotides [42], the success of our error elimination strategy suggests that DNA containing the UBP is not cleaved (or cleaved inefficiently) by virtue of only a single mismatch between an unnatural nucleotide and a natural nucleotide in the guide, which attests to the advantage of its orthogonal mechanism of pairing. Nevertheless, two problematic sequences (**CXG** and **CXA**) remained, which suggests that UBP loss likely occurs rapidly upon attempted replication in these contexts and is thus difficult to rescue via error elimination alone.

The error avoidance strategy is based on studies of how DNA containing the **dNaM-dTPT3** UBP is replicated and maintained by the SSO [43]. We found that while nucleotide excision repair (NER) does not recognize the UBP, methyl-directed mismatch repair (MMR) actually increases its retention, likely because the orthogonality of the hydrophobic unnatural nucleobases precludes recognition and excision, whereas the natural nucleotides of natural-unnatural mispairs are still recognized and excised. Importantly, we found that the major route to loss of the UBP was RecA-mediated recombinational repair of stalled replication forks, and that while the normal replicative polymerase, polymerase (Pol) III, commonly mediates the replication of DNA containing the UBP, it is aided at problematic sequences by the enigmatic Pol II. In fact, the most significant effect associated with the deletion of the gene encoding Pol II identified to date is reduced UBP retention in these problematic sequences. Based on this data, we constructed an SSO in which the RecA gene is deleted (as is common in strains used for cloning) and Pol II is released from its normal repression by

eliminating the upstream LexA binding site. Compared to an SSO lacking these modifications, the optimized SSO demonstrated improved dNaM-dTPT3 retention in all sequences examined, including CXG, as well as improved retention when the UBP was integrated into its chromosome, where it was retained through a hundred generations of growth with a fidelity per doubling of 99.6%. This error avoidance mechanism is likely to be particularly advantageous for some applications, such as the propagation of libraries of DNA containing the UBP in different sequence contexts, because it does not require the use or optimization of the guide RNAs that are required with the Cas9 error correction mechanism.

An SSO with an expanded genetic code.

Having developed SSOs with an expanded genetic alphabet, we turned our attention to using the expanded alphabet as the basis of an expanded code (Fig. 3b). In addition, in order to explore whether the UBP on its own could productively participate in every step of information storage and retrieval, despite completely lacking hydrogen bonds, our initial efforts employed an SSO that did not benefit from either the error elimination or avoidance mechanisms described above [••44]. We incorporated the UBP into the genes encoding superfolder green fluorescent protein (sfGFP) and an *E. coli* seryl-tRNA, such that the mRNA would be transcribed with NaM at the second position of codon 151 (*i.e.* sfGFP(AXC)¹⁵¹; X=NaM) and the tRNA would be transcribed with TPT3 at the corresponding position of its anticodon, with both genes under the control of inducible T7 RNAP promoters. With addition of both deoxy- and ribo-unnatural triphosphates to the media ((d)NaMTP and (d)TPT3TP), induction of the mRNA and tRNA resulted in cellular fluorescence that was similar to that of control cells expressing natural sfGFP. In contrast, expression of the mRNA without the seryl-tRNA or with co-expression of near cognate seryl-tRNAs recoded with GNT (N=A, T, C, or G) anticodons resulted in little to no fluorescence, suggesting that only tRNAs with the unnatural anticodon are capable of efficiently decoding the unnatural codon. After protein isolation, analysis by western blot revealed that only full-length protein was produced, and analysis by LC-MS/MS revealed that 98.5% of the protein was expressed with the Ser residue encoded by AXC. This data demonstrated that the SSO is capable of decoding the AXC codon with a fidelity and efficiency approaching that of a natural codon.

To demonstrate the use of the expanded genetic alphabet to encode an ncAA, we transcribed sfGFP(AXC)¹⁵¹ or sfGFP(GXC)¹⁵¹ and their cognate unnatural *M. mazei* tRNAs, and expressed the *M. barkeri* PylRS that charges each tRNA with the ncAA propargyl-lysine (PrK). Without the addition of PrK to the media, induction resulted in little to no fluorescence, but with its addition, high level fluorescence was observed. Remarkably, the level of fluorescence observed per cell was higher than that observed when the incorporation of the same ncAA was directed via amber suppression. Analysis by LC-MS/MS again demonstrated high protein purity (up to 98% having PrK). We then explored the use of the UBP to encode the ncAA *p*-azido-phenylalanine (*p*AzF) with an SSO expressing sfGFP(AXC)¹⁵¹ and an evolved *M. jannaschii* tRNA^{Tyr}/TyrRS pair from Schultz and co-workers, with the tRNA recoded with a cognate unnatural codon. Although mass spectrometric analysis of the protein was not possible due to instability of the azide moiety, western blot analysis in combination with azide-reactive dye conjugation revealed that 93%

of the protein contained *pAzF*, which was comparable to the 95% observed with amber suppression. It is of particular note that with each demonstration of UBP decoding, the majority of the reduced fidelity observed at the protein level, as well as the identity of misincorporated natural amino acids, was consistent with UBP loss at the DNA level. While further analysis is clearly required, the data suggest that the unnatural information made available by the **dNaM-dTPT3** UBP can be retrieved via transcription and translation with efficiencies and fidelities approaching that of a natural base pair. This in turn suggests that SSOs benefiting from the developed error avoidance or error elimination systems, in which UBP loss at the DNA level is reduced or effectively eliminated, respectively, should produce unnatural proteins with remarkable efficiency and purity.

Conclusions and future perspectives.

Amber suppression represented a milestone in the biological sciences, but it is not without its limitations, and efforts to overcome these limitations are challenged by fidelity, generality, cost, and possibly pleiotropic effects. The approach based on the development of UBPs requires a significant amount of chemical optimization, but it should be more portable to other host organisms, as it does not require significant genome modification, and it may even prove more optimal due to increased orthogonality, both in terms of mispairing with natural nucleotides and in terms of recognition by other proteins and effects on other cellular processes. While we have demonstrated efficient replication of DNA in a wide variety of sequences, for reasons that are not yet clear, several, in particular **CXG** and **CXA** (**X=NaM**) remain challenging. Similarly, while we have demonstrated efficient decoding at the ribosome with several unnatural codon-anticodon pairs, the majority of the now available unnatural codon-anticodon pairs remain to be characterized, and it is likely that differences will be observed. Finally, the successful use of the already identified UBPs as the basis of SSOs with an expanded genetic alphabet and code does not suggest that they cannot be further chemically optimized. In fact, the efficiency and fidelity of *in vivo* replication, transcription, and translation provides unprecedented SARs that will be invaluable for further UBP optimization. The extension of the approach to eukaryotic systems is clearly a future goal, and while this will raise its own set of challenges, it promises to provide an unprecedented platform for both the study of eukaryotic biology and the production of eukaryotic-specific proteins.

In addition to opening a new and possibly less restricted route to the expansion of the genetic code, the results are conceptually intriguing. They demonstrate that despite being completely devoid of the complementary hydrogen bonds that are so central to their natural counterparts, hydrophobic UBPs can participate in every step of what is perhaps the most intimate of life's processes: the storage and retrieval of information. The results also suggest that the "parts" used by nature may not be as unique and privileged as previously assumed. Finally, the results represent a remarkable integration of a man-made part into the biology of a living cell, a central goal of the emerging field of synthetic biology. With continued optimization of both the UBP and the SSO itself, the currently available SSOs may be just the first forms of semi-synthetic life that can access increased diversity via the production of new and different proteins, and perhaps even attain previously unobtainable forms and functions.

Acknowledgements

This work was supported by the National Institutes of Health (GM118178 to F.E.R. and GM128376 to R.J.K.) and the National Science Foundation (Graduate Research Fellowship NSF/DGE-1346837 to S.E.M.)

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest

•• of outstanding interest

1. Leader B, Baca QJ, Golan DE: Protein therapeutics: a summary and pharmacological classification. *Nat Rev Drug Discov* 2008, 7:21–39. [PubMed: 18097458]
- 2. Chin JW: Expanding and reprogramming the genetic code. *Nature* 2017, 550:53–60. [PubMed: 28980641] This review provides a recent summary of the amber suppression methodology for precise ncAA incorporation in order to probe, image, and control protein structure and function.
- 3. Young DD, Schultz PG: Playing with the molecules of life. *ACS Chem Biol* 2018, 13:854–870. [PubMed: 29345901] This review provides a recent summary of the amber suppression methodology, describing improved orthogonal translational machinery for applying ncAAs as probes and the development of new therapeutics.
4. Italia JS, Addy PS, Wrobel CJ, Crawford LA, Lajoie MJ, Zheng Y, Chatterjee A: An orthogonalized platform for genetic code expansion in both bacteria and eukaryotes. *Nat Chem Biol* 2017, 13:446–450. [PubMed: 28192410]
5. Neumann H, Wang K, Davis L, Garcia-Alai M, Chin JW: Encoding multiple unnatural amino acids via evolution of a quadruplet-decoding ribosome. *Nature* 2010, 464:441–444. [PubMed: 20154731]
6. Srinivasan G, James CM, Krzycki JA: Pyrrolysine encoded by UAG in Archaea: charging of a UAG-decoding specialized tRNA. *Science* 2002, 296:1459–1462. [PubMed: 12029131]
7. Wan W, Tharp JM, Liu WR: Pyrrolysyl-tRNA synthetase: an ordinary enzyme but an outstanding genetic code expansion tool. *Biochim Biophys Acta* 2014, 1844:1059–1070. [PubMed: 24631543]
- 8. Willis JCW, Chin JW: Mutually orthogonal pyrrolysyl-tRNA synthetase/tRNA pairs. *Nat Chem* 2018, 10:831–837. [PubMed: 29807989] This paper reports the development of mutually orthogonal Pyl tRNA-synthetase pairs.
9. Guo LT, Wang YS, Nakamura A, Eiler D, Kavran JM, Wong M, Kiessling LL, Steitz TA, O'Donoghue P, Soll D: Polyspecific pyrrolysyl-tRNA synthetases from directed evolution. *Proc Natl Acad Sci USA* 2014, 111:16724–16729. [PubMed: 25385624]
10. Coin I: Application of non-canonical crosslinking amino acids to study protein-protein interactions in live cells. *Curr Opin Chem Biol* [this issue]
11. Kelemen RE, Erickson SB, Chatterjee A: Synthesis at the interface of virology and genetic code expansion. *Curr Opin Chem Biol* 2018 [this issue]
12. Courtney T, Dieters A: Recent advances in the optical control of protein function through genetic code expansion. *Curr Opin Chem Biol* 2018 [this issue]
13. Kang M, Lu Y, Tian F: Harnessing the power of an expanded genetic code toward next-generation biopharmaceuticals *Curr Opin Chem Biol* [this issue]
14. Lee TC, Kang M, Kim CH, Schultz PG, Chapman E, Deniz AA: Dual unnatural amino acid incorporation and click-chemistry labeling to enable single-molecule FRET studies of p97 folding. *Chembiochem* 2016, 17:981–984. [PubMed: 27115850]
15. Zheng Y, Addy PS, Mukherjee R, Chatterjee A: Defining the current scope and limitations of dual noncanonical amino acid mutagenesis in mammalian cells. *Chem Sci* 2017, 8:7211–7217. [PubMed: 29081953]
16. Chatterjee A, Lajoie MJ, Xiao H, Church GM, Schultz PG: A bacterial strain with a unique quadruplet codon specifying non-native amino acids. *Chembiochem* 2014, 15:1782–1786. [PubMed: 24867343]

17. Mukai T, Hayashi A, Iraha F, Sato A, Ohtake K, Yokoyama S, Sakamoto K: Codon reassignment in the *Escherichia coli* genetic code. *Nucleic Acids Res* 2010, 38:8188–8195. [PubMed: 20702426]
18. Lajoie MJ, Rovner AJ, Goodman DB, Aerni HR, Haimovich AD, Kuznetsov G, Mercer JA, Wang HH, Carr PA, Mosberg JA, Rohland N et al.: Genomically recoded organisms expand biological functions. *Science* 2013, 342:357–360. [PubMed: 24136966]
19. Johnson DB, Xu J, Shen Z, Takimoto JK, Schultz MD, Schmitz RJ, Xiang Z, Ecker JR, Briggs SP, Wang L: RF1 knockout allows ribosomal incorporation of unnatural amino acids at multiple sites. *Nat Chem Biol* 2011, 7:779–786. [PubMed: 21926996]
20. Schmied WH, Elsasser SJ, Uttamapinant C, Chin JW: Efficient multisite unnatural amino acid incorporation in mammalian cells via optimized pyrrolysyl tRNA synthetase/tRNA expression and engineered eRF1. *J Am Chem Soc* 2014, 136:15577–15583. [PubMed: 25350841]
21. Aerni HR, Shifman MA, Rogulina S, O'Donoghue P, Rinehart J: Revealing the amino acid composition of proteins within an expanded genetic code. *Nucleic Acids Res* 2015, 43:e8. [PubMed: 25378305]
22. Ostrov N, Landon M, Guell M, Kuznetsov G, Teramoto J, Cervantes N, Zhou M, Singh K, Napolitano MG, Moosburner M, Shrock E et al.: Design, synthesis, and testing toward a 57-codon genome. *Science* 2016, 353:819–822. [PubMed: 27540174]
23. Plotkin JB, Kudla G: Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 2011, 12:32–42. [PubMed: 21102527]
24. Richardson SM, Mitchell LA, Stracquadanio G, Yang K, Dymond JS, DiCarlo JE, Lee D, Huang CL, Chandrasegaran S, Cai Y, Boeke JD, Bader JS: Design of a synthetic yeast genome. *Science* 2017, 355:1040–1044. [PubMed: 28280199]
25. Karalkar NB, Leal NA, Kim MS, Bradley KM, Benner SA: Synthesis and enzymology of 2'-deoxy-7-deazaisoguanosine triphosphate and its complement: a second generation pair in an artificially expanded genetic information system. *ACS Synth Biol* 2016, 5:672–678. [PubMed: 26914388]
26. Moran S, Ren RX, Kool ET: A thymidine triphosphate shape analog lacking Watson-Crick pairing ability is replicated with high sequence selectivity. *Proc Natl Acad Sci USA* 1997, 94:10506–10511 [PubMed: 9380669]
27. Moran S, Ren RX, Rumney S, Kool ET: Difluorotoluene, a nonpolar isostere for thymine, codes specifically and efficiently for adenine in DNA replication. *J Am Chem Soc* 1997, 119:2056–2057. [PubMed: 20737028]
28. Okamoto I, Miyatake Y, Kimoto M, Hirao I: High fidelity, efficiency and functionalization of Ds-Px unnatural base pairs in PCR amplification for a genetic alphabet expansion system. *ACS Synth Biol* 2016, 5:1220–1230 [PubMed: 26814421]
- 29. Feldman AW, Romesberg FE: Expansion of the genetic alphabet: A chemist's approach to synthetic biology. *Acc Chem Res* 2018, 51:394–403. [PubMed: 29198111] This review provides an extensive recounting of fifteen years of UBP optimization.
30. Li L, Degardin M, Lavergne T, Malyshev DA, Dhimi K, Ordoukhanian P, Romesberg FE: Natural-like replication of an unnatural base pair for the expansion of the genetic alphabet and biotechnology applications. *J Am Chem Soc* 2014, 136:826–829 [PubMed: 24152106]
31. Malyshev DA, Dhimi K, Quach HT, Lavergne T, Ordoukhanian P, Torkamani A, Romesberg FE: Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc Natl Acad Sci USA* 2012, 109:12005–12010. [PubMed: 22773812]
32. Matsuda S, Fillo JD, Henry AA, Rai P, Wilkens SJ, Dwyer TJ, Geierstanger BH, Wemmer DE, Schultz PG, Spraggon G, Romesberg FE: Efforts toward expansion of the genetic alphabet: structure and replication of unnatural base pairs. *J Am Chem Soc* 2007, 129:10466–10473. [PubMed: 17685517]
33. Seo YJ, Hwang GT, Ordoukhanian P, Romesberg FE: Optimization of an unnatural base pair toward natural-like replication. *J Am Chem Soc* 2009, 131:3246–3252. [PubMed: 19256568]
34. Betz K, Malyshev DA, Lavergne T, Welte W, Diederichs K, Dwyer TJ, Ordoukhanian P, Romesberg FE, Marx A: KlenTaq polymerase replicates unnatural base pairs by inducing a Watson-Crick geometry. *Nat Chem Biol* 2012, 8:612–614. [PubMed: 22660438]

35. Betz K, Malyshev DA, Lavergne T, Welte W, Diederichs K, Romesberg FE, Marx A: Structural insights into DNA replication without hydrogen bonds. *J Am Chem Soc* 2013, 135:18637–18643. [PubMed: 24283923]
36. Seo YJ, Romesberg FE: Major groove derivatization of an unnatural base pair. *ChemBiochem* 2009, 10:2394–2400. [PubMed: 19722235]
- 37. Malyshev DA, Dhami K, Lavergne T, Chen T, Dai N, Foster JM, Correa IR, Jr., Romesberg FE: A semi-synthetic organism with an expanded genetic alphabet. *Nature* 2014, 509:385–388 [PubMed: 24805238] This paper reports the use of the d5SICS–dNaM UBP and the PtNTT transporter to create the first SSO with an expanded genetic alphabet.
38. Feldman AW, Fischer EC, Ledbetter MP, Liao JY, Chaput JC, Romesberg FE: A tool for the import of natural and unnatural nucleoside triphosphates into bacteria. *J Am Chem Soc* 2018, 140:1447–1454. [PubMed: 29338214]
39. Ast M, Gruber A, Schmitz-Esser S, Neuhaus HE, Kroth PG, Horn M, Haferkamp I: Diatom plastids depend on nucleotide import from the cytosol. *Proc Natl Acad Sci USA* 2009, 106:3621–3626. [PubMed: 19221027]
- 40. Zhang Y, Lamb BM, Feldman AW, Zhou AX, Lavergne T, Li L, Romesberg FE: A semisynthetic organism engineered for the stable expansion of the genetic alphabet. *Proc Natl Acad Sci USA* 2017, 114:1317–1322 [PubMed: 28115716] This paper reports the use of the dNaM–dTPT3 UBP, a modified transporter, and the Cas9-based error correct system to optimize the SSO.
41. Feldman AW, Romesberg FE: In vivo structure-activity relationships and optimization of an unnatural base pair for replication in a semi-synthetic organism. *J Am Chem Soc* 2017, 139:11427–11433 [PubMed: 28796508]
42. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, Cradick TJ et al.: DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* 2013, 31:827–832. [PubMed: 23873081]
43. Ledbetter MP, Karadeema RJ, Romesberg FE: Reprogramming the replisome of a semisynthetic organism for the expansion of the genetic alphabet. *J Am Chem Soc* 2018, 140:758–765. [PubMed: 29309130]
- 44. Zhang Y, Ptacin JL, Fischer EC, Aerni HR, Caffaro CE, San Jose K, Feldman AW, Turner CR, Romesberg FE: A semi-synthetic organism that stores and retrieves increased genetic information. *Nature* 2017, 551:644–647 [PubMed: 29189780] This paper reports the first fully functional SSO that stores the UBP in its DNA and also retrieves the increased information in the form of proteins containing ncAAs.

Highlights

- An unnatural base pair has been developed and used as the foundation of semi-synthetic organisms
- The semi-synthetic organisms can use the unnatural base pair to efficiently make unnatural proteins
- Hydrophobic and packing interactions can replace hydrogen bonds to control base pairing

Figure 1. Base pairs. **(a)** The dNaM-d5SICS and dNaM-dTPT3 UBPs and a natural dG-dC pair (with complementary H-bonding shown). **(b)** The dCNMO-dTPT3 UBP. Phosphate and sugar moieties omitted for clarity.

Figure 2.

Proposed mechanism of replication of DNA containing a UBP. The steps corresponding to unnatural triphosphate incorporation and extension are indicated. During synthesis within the DNA polymerase active site, the UBP adopts a Watson-Crick-like structure, but after translocation it is prone to intercalate, which hinders extension and requires de-intercalation. dX=dNaM and dY=dTPT3.

Figure 3.
(a) SSO that replicates the UBP. (b) SSO that replicates and transcribes the UBP and decodes an unnatural codon to produce unnatural protein. **X=NaM** and **Y=TPT3**.