Next-generation genetic code expansion
Pol Arranz-Gibert\textsuperscript{1,2,3}, Koen Vanderschuren\textsuperscript{1,2,3} and Farren J Isaacs\textsuperscript{1,2}

Engineering of the translation apparatus has permitted the site-specific incorporation of nonstandard amino acids (nsAAs) into proteins, thereby expanding the genetic code of organisms. Conventional approaches have focused on porting tRNAs and aminoacyl-tRNA synthetases (aaRSs) from archaea into bacterial and eukaryotic systems where they have been engineered to site-specifically encode nsAAs. More recent work in genome engineering has opened up the possibilities of whole genome recoding, in which organisms with alternative genetic codes have been constructed whereby codons removed from the genetic code can be repurposed as new sense codons dedicated for incorporation of nsAAs. These advances, together with the advent of engineered ribosomes and new molecular evolution methods, enable multisite incorporation of nsAAs and nonstandard monomers (nsMs) paving the way for the template-directed production of functionalized proteins, new classes of polymers, and genetically encoded materials.

Addresses
\textsuperscript{1} Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, USA
\textsuperscript{2} Systems Biology Institute, Yale University, West Haven, CT 06516, USA
\textsuperscript{3} Equal contribution

Corresponding author: Isaacs, Farren J. (farren.isaacs@yale.edu)

Introduction
Proteins have evolved to assume diverse cellular functions, such as structural, catalytic or regulatory roles essential for cells. Enzymatic-driven post-translational modifications increase the chemical diversity of proteins well beyond the twenty standard amino acids (AA), thereby considerably enhancing their functions. Research in the field of genetic code expansion – that is, the addition of nonstandard amino acids (nsAAs) to the genetic code – has the potential to greatly expand the chemistries available for protein and peptide synthesis. More than 150 different nsAAs have been incorporated to date for a variety of scientific and biotechnological applications, such as handles for bioconjugation, photo-cross-linking, biophysical probes and genetically encoded “post-translational” modifications \cite{1,2}. Additionally, there is increasing interest in genetically encoded monomers with backbones and chemistries that go beyond the L-\textalpha- AA configuration (i.e. non-standard monomers (nsMs)), setting the stage for synthesis of entirely new classes of biopolymers and materials.

Expansion of the genetic code with nsAAs requires introduction of orthogonal translation systems to enhance the chemical repertoire of proteins while also preserving and not cross-reacting with native protein synthesis. Analogous to standard AAs, the first step is aminoacylation of the tRNA with its cognate AA by an aminoacyl-tRNA synthetase (aaRSs) (Figure 1). Next, the charged tRNA must form a ternary complex with EF-Tu-GTP, where both the tRNA body and the AA contribute to binding \cite{3}. These ternary complexes sample the open codon by partially occupying the ribosomal A site (aminoacyl-tRNA binding site). If the mRNA codon and tRNA anticodon match, the tRNA is released from the ternary complex and the acceptor stem enters the ribosomal peptidyl transferase-center (PTC). Here, the polypeptide is transferred from the P site (peptidyl-tRNA site) to the A site tRNA. Next, translocation of the tRNAs releases the empty tRNA from the E site (exit site), and the next codon enters the A site. When a stop codon is presented in the A site, translation is terminated and the polypeptide is released \cite{4}. In prokaryotes, the stop codons UAG and UAA are recognized by the Release Factor 1 (RF1), and UAA and UGA by RF2.

Engineering of the translation apparatus
There are two main strategies that have been employed to genetically encode nsAAs. The first re-purposes an existing aaRS for the incorporation of an nsAA. This strategy involves the expression of the protein of interest (POI) in an auxotrophic expression host, that is, a strain unable to biosynthesize the AA being replaced by the target nsAA. Substitution of the standard AA with a close nsAA analog in the growth media will cause global suppression whereby the nsAA is incorporated in all positions throughout the proteome and POI instead of the standard AA \cite{5,6}. In this strategy, all components are already in place; however, it does not allow the addition of an AA to the genetic code but rather the replacement of a standard one with a nsAA. Furthermore, the substitution of the
standard AA with the nsAA is oftentimes cytotoxic, preventing stable expression with the nsAA.

The second approach permits the site-specific incorporation of an nsAA alongside the full complement of the 20 standard AAs. This strategy involves the development of orthogonal translation systems (OTS) whereby engineered tRNA-aaRS pairs function alongside the native translation machinery. In this scheme, the orthogonal tRNA is not aminoacylated by any of the native aaRSs, and similarly, none of the native tRNAs are charged by the orthogonal aaRS. Two of the most widely used OTSs in bacteria are the natural amber suppressor pyrrolysyl-aaRS (PylRS)-tRNA_{Pyl} pair derived from *Methanosarcina barkeri* [7], and the tyrosyl-RS (TyrRS)-tRNA<sup>Tyr</sup> from *Methanothermobacter jannaschii*. In eukaryotes, both the PylOTS and *E. coli* TyrRS-tRNA<sup>Tyr</sup> can be used to genetically encode nsAAs [7–9]. Although OTSs are generally employed to encode nsAAs, they may also be used to encode standard AA and thus functionally replace a native tRNA/aaRS pair. In turn, this native tRNA/aaRS pair can be repurposed for genetic code expansion [10,11]. For example, the endogenous tryptophanyl-RS and tRNA pair from *E. coli* was functionally replaced with its *Saccharomyces cerevisiae* counterpart. Subsequently, the *E. coli* pair was reintroduced as either a UAG or UGA suppressor to

---

**Figure 1**

Translation system and sites of engineering for nsAA incorporation: tRNA, charged by its cognate aaRS with the corresponding AA, is recognized by EF-Tu-GTP; after the initiation of translation, the ternary complex aa-tRNA-EF-Tu-GTP binds the A site; later, the EF-Tu hydrolyzes GTP, changes its conformation and leaves the ribosome; tRNAs at A and P sites are translocated to P and E sites, respectively; the empty tRNA located at the E site is released; at the end of translation, a stop codon is recognized by a release factor that also interacts with acceptor stem of the P site tRNA and triggers hydrolysis of peptidyl-tRNA at the PTC. Red dotted circles show the identity elements for EF-Tu and aaRSs to recognize their cognate AAs and tRNAs. For aaRS, they often localize in the anticodon loop and acceptor stem; however, the variable arm (VA) and D-loop can also be recognized. EF-Tu virtually binds all aa-tRNAs through direct interaction with both AA and tRNA.
expand the genetic code with several tryptophan analogs, such as 5-hydroxytryptophan [10*].

As goals for genetic code expansion aim to go beyond the incorporation of a single nsAA into proteins, the enhancement of activity and orthogonality of OTSs that can operate simultaneously has emerged as a key challenge. Many of the current OTSs are polyspecific and permit charging of multiple nsAAs by the same orthogonal aaRS. To enable site-specific incorporation of multiple genetically encoded nsAAs, it will be imperative to ensure the compatibility and selectivity of the OTSs.

To address these challenges, powerful engineering strategies are needed to optimize properties of OTSs. Advances in molecular evolution and genome engineering have presented new methods to evolve the performance of OTSs in cells. For example, in a recent study by Bryson et al. phage-assisted continuous evolution (PACE) was used to generate 45-fold more active PyRS variants and MJRS variants with improved specificity [12*]. Another study by Amiram et al. combined the use of multiplex automated genome engineering (MAGE) in a genomically recoded organism (GRO; see next section) to produce 25-fold more active MJRS enzymes capable of introducing 30 instances of an nsAA in a single protein [13*]. The MAGE-OTS platform in this study also isolated aaRS enzymes with tunable specificities to >200 nsAAs, including one exclusive for para-azido-L-phenylalanine (pAzF).

In general, engineering the aaRS binding pocket to facilitate the accommodation of the desired nsAA may be sufficient; however, generating well-performing OTSs may require evolution of multiple components in parallel. For example, the performance of several aaRS/rRNA pairs was greatly enhanced by engineering the AA binding pocket of the aaRS in conjunction with the binding interface between the tRNA anticodon loop and the aaRS [13*]. In other cases, it became apparent that poor binding characteristics between the tRNA<sup>nsAA</sup> and EF-Tu greatly limited the incorporation of the nsAA O-phosphoserine [14]. Indeed, it was necessary to modify the EF-Tu [14] or the tRNA [15] to tune their binding kinetics for improved incorporation. A similar strategy improved the incorporation efficiency of selenocysteine and bulky residues (e.g. dl-2-anthraquinonylalanine) [16,17]. We anticipate future genetic code expansion efforts will also need to conduct EF-Tu engineering.

Attempts to incorporate bulky or charged nsAAs or even altered backbone monomers (nsM) will likely confront challenges dispersed across multiple steps in the translation process. Such efforts will likely require engineering of the ribosome. Despite the fact that the ribosome limits the incorporation of exotic backbones, in vitro work has shown that N<sup>α</sup>-methylated [18,19], D-α- [20–22] and β-amino acids [23*] and N-substituted glycines [24] can be successfully incorporated in peptides. Notably, for several of these monomers, the efficiency was improved by mutating the ribosome in the PTC [18–23*,24]. Recently, the incorporation of β-amino acids has even been reported in vivo using an engineered ribosome [25].

A large challenge to ribosome engineering is that the fidelity and activity of the ribosome is inherently tied to protein synthesis and as a result to cell viability. To relieve this interdependency, it has been shown that in E. coli an additional, orthogonal ribosome (oRibosome) can be introduced, which is uncoupled from cell viability. The oRibosome has an altered anti Shine-Dalgarno (SD) sequence, which reduces its ability to initiate translation of native mRNA transcripts. Next, orthogonal genes can be introduced with a SD sequence that matches the anti-SD of the oRibosome (Figure 2) [26,27]. Despite these efforts, full orthogonality has not been achieved and complete isolation of oRibosome translation remains a challenge worthy of further pursuit.

A second challenge is that the anti-SD is located in the 16S rRNA, and the 23S rRNA can freely exchange between the native and orthogonal 16S rRNA populations. Recent work has shown that physical tethering of the 16S and 23S rRNAs establishes an oRibosome that will not cross-react with native ribosomes and therefore becomes an isolated ribosomal system [28**,29**]. As a result, this tethered oRibosome can be evolved beyond what native ribosomes can tolerate without impeding native translation and becoming lethal. This establishes a path toward fully isolated ribosomes whose PTC can be mutagenized to accommodate modified backbones and drive new catalysis beyond amide bond formation.

Lastly, cell-free protein synthesis (CFPS) is a unique opportunity to study and manipulate translational systems to produce proteins with nsAAs without the constraints of cell viability [30]. For example, amino acids that cannot pass cell membranes can be readily added in these systems [31]. Even more, if tRNA aminoacylation by a cognate aaRS cannot be achieved, chemical charging or flexizymes can be used to overcome this step [32,33]. Additionally, CFPS systems with defined tRNA composition, either reconstituted from purified components [34] or cell extracts lacking native tRNAs [35], facilitate genetic code reassignment [36].

**Genomically recoded organisms (GROs): organisms with alternative genetic codes**

An important challenge in genetic code expansion is that native systems typically do not possess open codons that can be dedicated to a nsAA. In other words, the OTS competes with a native tRNA or release factor for the assigned codon. For example, if the tRNA from the OTS decodes the UAG amber stop codon, then it competes
with RF1. In this scenario, a fraction of the POI will be terminated prematurely at the UAG codon, resulting in truncated POI and reduced yields of full-length protein. This issue is magnified for efforts aimed at multisite incorporation of nsAAs. Additionally, the OTS may also suppress native stop codons, leading to misincorporation of the nsAA at more than 300 UAG codons found in native proteins (Figure 3). This misincorporation may result in non-functional proteins and cytotoxicity, further lowering yields of the POI [37].

Benefiting from recent advances in genome engineering, these issues can be overcome through whole-genome recoding, in which all instances of a codon can be eliminated from the genome. To achieve compatibility of recoding at the translational level, these efforts need to be coordinated with elimination or modification of native translation machinery that decodes the target codon at the ribosome. This was first demonstrated in *E. coli* MG1655 by converting all 321 UAG (amber) codons to UAA (ochre), constructing the first GRO (Figure 3) [38]. For this work, the UAG to UAA mutations were introduced in 32 strains in parallel through MAGE [39] and recoded genome fragments were then combined through five hierarchical steps with conjugative assembly genome engineering (GAGE) [40]. In addition, RF1 (prfA gene) was deleted to eliminate native UAG function. This GRO, named C321.DΔ, exhibited enhanced functions [37], which include blocking of horizontally transferred genetic elements, resistance to multiple viruses [41], biocontainment [42,43], and improved properties for nsAA incorporation [13**].

Using a similar strategy, all 123 essential genes containing AGA and AGG codons (arginine) in *E. coli* were reassigned to CGN (Figure 3) [44]. Also, more aggressive attempts to recode the genome by removing 13 rare codons from 42 highly expressed essential genes have also been pursued using similar methodology [45]. This work has inspired several other ambitious recoding strategies that include sense recoding. One large effort aims to generate a GRO with a 57-codon genome by introducing synthetic DNA fragments (recoding of serine: AGC, AGU; leucine: UUA, UUG; arginine: AGG, AGA; stop: UAG; Figure 3) [46]. Another project has developed SIRCAS (stepwise integration of rolling circle amplified segments) to construct a *Salmonella typhimurium* (*S. typhimurium*) strain where 1557 leucine codons were replaced with synonymous codons across 176 genes (Figure 3) [47]. Lastly, a project dedicated to the design of a synthetic *Saccharomyces cerevisiae* genome (Sc 2.0) is
re-writing synthetic yeast genomes with numerous planned alterations, which include no introns, no IS elements, an inducible evolution system, and UAG to UAA codon reassignment across the genome to harness unique properties of GROs (Figure 3) [48**].

Expanding the genetic code: 4/5-base codon and nonstandard nucleobases (nsNB)

Other exciting strategies aim to expand the genetic code by stepping outside the standard 64 codons. The first of these aims to introduce tRNA species that read quadruplet codons, rather than the canonical triplets. Almost fifty years ago, quadruplet suppression was observed in strains of *S. typhimurium* carrying the mutation *suFD* (a CCCC quadruplet anticodon), thereby producing a tRNA^{Gly} containing the frameshift suppressor mutation, a CCCC anticodon [49]. Recently, quadruplet decoding, and thus +1 frameshifting, has been used for nsAA incorporation [50,51]. Much work has focused on UAGN codons, which could theoretically expand the genetic code with four new open codons.

The second strategy is based on the introduction of nonstandard nucleobases (nsNB). Benner and co-workers designed an nsNB pair (P–Z) with a different hydrogen bonding donor/acceptor pattern compared to that found in standard nucleobases [52,53], whereas Hirao and Romesberg’s groups developed hydrophobic nsNB pairs, Ds-Px [54–57] and NaM-TPT3 [58–60], respectively. Impressively, the Romesberg lab recently developed a strain of *E. coli* that imports [61,62] and uses NaM-TPT3 as the third pair of nucleotides: they can be maintained in the genome [63] and transcribed [64**]. Additionally, protein is translated from these mRNAs, enabling the incorporation of nsAA [64**]. Theoretically, the addition of this orthogonal nsNB pair to the genetic code would expand it from 64 to 216 possible codons; however, it is difficult to assess the physical and biological properties of this new nsNB pair, and it is unlikely that all new codons may allow genetic code expansion.

Conclusions

Three major efforts in the field of genetic code expansion are rapidly advancing the possibilities of nsAA incorporation in proteins: (1) advanced molecular evolution strategies to engineer OTSs, (2) efforts to free codons that can be re-assigned for dedicated incorporation of nsAAs, and
6 Synthetic biology

(3) ribosome engineering to expand nsAA incorporation beyond L-α-AAs.

Engineering components of translation requires the use of in vitro or in vivo approaches that diversify and select for evolved components (e.g., aaRS, tRNA, and EF-Tu). However, the combinatorial possibilities to evolve more than one component at a time are vast and empirically unattainable, yet evolution of single components alone has limitations. Integration of combined efforts to evolve all of these individual components to work in a selective and efficient manner with the native translation apparatus will allow faithful expansion of the genetic code. In this regard, two main features need to be addressed using advanced molecular evolution strategies to enable efficient incorporation of multiple nsAAs, namely OTS efficiency (aaRS activity is typically 10^2–10^3-fold below their native counterparts) [65,66] and specificity, as many orthogonal aaRSs exhibit polyspecificity towards several nsAAs [67].

Historically, OTS performance limited the nsAA incorporation to only one or a few instances per protein, but this limitation is also due to the lack of open codons dedicated to the incorporation of an nsAA. The construction of a GRO has enabled the incorporation of multiple instances of an nsAA into a single protein. Further recording efforts, advances in quadruplet suppression and introduction of nsNBs are likely to expand the number of codons available for genetic code expansion in the near future. This will pave the way for efficient genetic code expansion with multiple nsAAs.

Finally, the work on orthogonal and tethered ribosomes provides the cell with a complementary translation apparatus that can be fully dedicated to the synthesis of biotechnological products, breaking the link between cell survival and translational synthesis.

We envision GROs with multiple open codons combined with multiple OTSs and orthogonal, tethered ribosomes integrated in the genome, providing a biological framework capable of establishing a stable and expanded genetic code (Figure 4). Accurate, specific, and multi-site incorporation of diverse nsAAs is a major goal to be met soon and enable the production of functionalized proteins

Figure 4

Expansion of the genetic code in GROs containing multiple open coding channels. Native translation (left) is preserved and isolated from orthogonal translation systems (OTS, right) containing o-tRNA, o-aaRS, oRibO. The OTS is designed to minimize cross-reacting or interfering with the native translational machinery. The open coding channels in the GRO permit efficient, multi-site incorporation of nonstandard amino acids (nsAAs) into proteins. Future efforts aimed at engineering all components of OTSs may enable the use of nonstandard monomers (nsM) possessing chemistries and modified backbones distinct from standard AA or nsAAs. Such capabilities could permit multi-site incorporation of multiple combinations of AA, nsAAs or nsMs for templated-directed production of genetically encoded materials (GEMs).
and polymers containing diverse nSAAs and nSM, leading to the production of entirely new forms of matter, genetically encoded materials (GEMs).

Acknowledgements

FJI gratefully acknowledges the National Science Foundation (CHE-1740549, MCB-1714860) and National Institutes of Health (1R01GM125951-01, 1R01GM117230-01) for funding research related to topics covered in this review.

References and recommended reading

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

- of special interest
- of outstanding interest


This work shows the functional replacement of s: coil TrpRS-tRNA(TTrp) with its engineered s: cerevisiae counterpart. In turn, this enabled repurposing of the E: coil TrpRS-tRNA(TTrp) for genetic code expansion with several Trp-analogues.


In this work, the selectivity and activity of aaRSs were improved through PACE. Using this advanced molecular evolution strategy, the authors reported up to a 45-fold increase in activity for PyrRS.


Building on advances in genome engineering methods and GROs, these authors developed a platform for OTS optimization, which resulted in up to 25-fold improvement in aaRS activity. Additionally, the work demonstrates the tunability of substrate specificity for aaRSs.


The authors tested, using a cell-free translation system, the incorporation during translation of several ω-amino acids. Multiple instances were introduced using those that showed higher incorporation efficiency.


This work reports the incorporation in vivo of a ω-amino acid into a protein using wild-type EF-Tu, PheRS and mutant ribosomes.


Successful tethering of the 23S rRNA to an orthogonal 16S rRNA establishes a parallel ribosome, which can be readily engineered without affecting cell viability. This work is paving the way for new opportunities for ribosome engineering, and ribosomal synthesis of novel biopolymers.


The authors constructed a tethered or “stapled” orthogonal ribosome to enable mutations in the 23S rRNA without affecting cell viability.

8 Synthetic biology


This work marks the completion of UAG to UAA whole-genome recoding in E. coli. The resulting GRO exhibited improved nsAA incorporation and increased viral resistance.


The authors design and work toward the construction of anE. coli with 57-codon genome.


The authors recoded a strain of Salmonella using a novel methodology called SIRCAS, which involves the direct iterative recombining of 10-25 kb synthetic DNA constructs. They replaced 1557 synonymous leucine codons across 176 genes.


This paper describes the redesign and synthesis of a yeast genome with an 8% reduction in size, removing introns, IS elements, reassigning UAG codons to UAA, and introducing an inducible “SCRaMbLe” evolution system. This work opens up the possibility of customizable genome synthesis.


64. Zhang Y, Ptacin JL, Fischer EC, Aerni HR, Caffaro CE, San Jose K, Feldman AW, Turner CR, Romesberg FE: A semi-synthetic...

The authors present a strain of E. coli in which nsNBs can be maintained in genomic DNA and used in transcription and in translation. The added nsNB pair has the potential to greatly expand the number of open codons that could be used to encode nsAA.

