Overcoming Challenges in Engineering the Genetic Code

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Published Version</td>
<td>10.1016/j.jmb.2015.09.003</td>
</tr>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:37308776">http://nrs.harvard.edu/urn-3:HUL.InstRepos:37308776</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Open Access Policy Articles, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#OAP</a></td>
</tr>
</tbody>
</table>
Overcoming challenges in engineering the genetic code

MJ Lajoie\textsuperscript{1,2,\textdagger}, D Söll\textsuperscript{3}, and GM Church\textsuperscript{1,4}

\textsuperscript{1}Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
\textsuperscript{2}Program in Chemical Biology, Harvard University, Cambridge, MA 02138, USA
\textsuperscript{3}Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520-8114, USA
\textsuperscript{4}Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA

Abstract

Withstanding 3.5 billion years of genetic drift, the canonical genetic code remains such a fundamental foundation for the complexity of life that it is highly conserved across all three phylogenetic domains. Genome engineering technologies are now making it possible to rationally change the genetic code, offering resistance to viruses, genetic isolation from horizontal gene transfer, and prevention of environmental escape by genetically modified organisms. We discuss the biochemical, genetic, and technological challenges that must be overcome in order to engineer the genetic code.

Graphical abstract

---

To whom correspondence should be addressed: mlajoie@genetics.med.harvard.edu.

\textdagger Present address: Department of Biochemistry, University of Washington, Seattle, WA 98195, USA.

Publisher’s Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of interest:

GMC has potentially relevant patent applications and connections to companies: arep.med.harvard.edu/gmc/tech.html
Keywords

génomically recoded organism (GRO); expanded genetic code; genome engineering

Introduction

Nearly all organisms share a common genetic code—the language that specifies how genetic information is interpreted to produce proteins. Although some organisms and mitochondria use non-canonical genetic codes, they typically involve a small number of codon reassignments in response to strong selective pressures [1]. For example, mitochondria are under selection to minimize genome size [1,2] and G+C content [3], while bacteria may use codon reassignment to evade viruses or to restrict horizontal gene transfer [4,5].

Why has the genetic code been so refractory to change? While it is possible that more divergent genetic codes have yet to be discovered [5], several factors help to conserve the genetic code. Evolution tends to increase biological complexity [6], leading to the large genomes of modern free-living organisms (the smallest known genome with a full complement of essential genes has 580,070 base pairs encoding 470 predicted open reading frames [7]). With a few exceptions [4,8–10], these organisms use all 64 codons to encode their proteins while simultaneously accommodating overlapping sequence features such as protein binding sites, promoters, splicing signals, and RNA secondary structure [11]. In this context, the genetic code provides fundamental biochemical constraints that guide how a genome is put together. Therefore, the genetic code shapes how mutations affect an evolving genome, while the genome relies on a stable genetic code to faithfully produce proteins essential for life. Any change in codon function must be tolerated at all instances genome-wide. Therefore small changes in the genetic code must be accompanied by many compensatory changes in the rest of the genome. This scenario is unlikely to occur by random mutagenesis, but current genome engineering technologies are now making it possible to rationally change the genetic code.
What does it mean to “change the genetic code?”

The term “genetic code” has been used with different meanings in different contexts. Since this can lead to confusion, we propose less ambiguous alternatives for “changing the genetic code.” (1) Changing the genome sequence (e.g., synonymous codon swaps in one or more genes) will be referred to as genome editing. (2) Introducing new amino acid assignments of one or more codons without removing the original function (e.g., UAG decoded as both a stop and an amino acid [12]) will be referred to as codon suppression. (3) Changing the amino acid assignments of one or more codons genome wide (e.g., genomically recoded organisms [13]) will be referred to as codon reassignment. (4) Adding a new codon to the translation code table (e.g., using quadruplet codons [14–16] or codons composed of unnatural bases [17]) will be referred to as codon creation. Broadly speaking, the term “genetic code” will be used throughout this review to describe the codon assignments in the translation code table. Codon suppression, reassignment, and creation are all ways of changing a genetic code.

Applications of changing the genetic code

Since the 1970s, biotechnology has relied on a ubiquitous code to permit the transgenic production of drugs [18], materials [19], and food [20]. Current DNA synthesis technologies [21–23] have liberated us from reliance on the canonical genetic code, yet we are still subjected to its challenges—viral infection and undesired horizontal gene transfer (e.g., antibiotic resistance [24], recombinant DNA dissemination [25–27]). Furthermore, while the 20 canonical amino acids support an astonishing diversity of biochemistry, they nevertheless limit the potential for new and useful protein functions. This fact is exemplified by the natural existence of selenocysteine and pyrrolysine as specialized catalytic residues for redox reactions and methanogenesis [1].

Organisms possessing alternate genetic codes have the potential to overcome these challenges (Figure 1). This hypothesis motivated the development of genomically recoded organisms (GROs), whose codons have been unambiguously reassigned to create alternate genetic codes [13]. GROs are able to efficiently incorporate non-standard amino acids (nsAAs) [28] that have been developed to enhance enzyme activity [29,30], to improve the performance of drugs [31,32], and to function as molecular probes [33]. Redesigning essential proteins to depend on these nsAAs for proper translation and function provides a robust strategy for restricting undesired survival outside of controlled environments [34,35]. Additionally, by interpreting genetic information differently, GROs would mistranslate foreign genes from natural organisms. This would prevent viruses from hijacking their translation machinery (potentially saving hundreds of millions of dollars from lost productivity [36,37]) and would thwart transfer of functional genetic information with natural organisms [13]. Thus, GROs have the potential to be safe and powerful chassis for biofermentation, bioremediation, and agriculture.

We recently reported the first GRO, which has an unambiguously reassigned UAG codon. However, while this GRO has demonstrated promising properties such as increased virus resistance [13] and metabolic dependence on nsAAs [34,35], evolution can readily
overcome genetic isolation from one reassigned stop codon. Indeed, UAG suppression is well-documented in natural organisms [5], and bacteriophage T7 readily evolves tolerance to a host with a reassigned UAG codon [38]. Therefore, radically altered genetic codes will be required in order to realize the full potential of these applications [39]. While only a subset of sense codons likely need to be reassigned in order to achieve robust genetic isolation, this review will consider the fundamental limits to changing the genetic code based on biochemical principles, genetic knowledge, and genome engineering technologies.

Engineering expanded genetic codes

In vitro translation systems offer the ultimate flexibility to implement alternate genetic codes [40–42]. Since translation components can be prepared separately, non-specific aminoacylation methods such as CA ligation [43] and flexizyme [44] can be used to incorporate a broad selection of nsAAs. As a result, in vitro translation has been the best way to produce unnatural backbones [45,46] for synthetic nonribosomal peptide mimetics [47,48] and polymer materials [46,49]. Furthermore, the production of ribosomes in vitro may accommodate extensive modifications that could otherwise compromise fitness in vivo [50]. Recent reports of orthogonal 16S rRNA [51], orthogonal tRNAs [52], tethering 16S to 23S rRNA [53], provide the infrastructure to evolve ribosomes with radically modified functions.

Codon suppression

In vivo systems are well-suited for inexpensive, simple, and scalable translation using nsAAs, but must be compatible with essential cellular processes. While sense codons have been transiently diverted to incorporate diverse nsAAs by metabolic labeling [54], persistent metabolic labeling is likely to be highly deleterious. Even evolving tolerance for structurally similar Trp analogs has met varying success in different systems [55–58]. In contrast, ambiguous decoding of stop codons is well-tolerated in E. coli [12,59], making it possible to introduce orthogonal translation machinery capable of producing high yields of nsAA-containing proteins in vivo [60–62]. The implementation of orthogonal translation machinery [33] has led to an explosion in the number of nsAAs (currently more than 167 nsAAs [28]) that can be site-specifically incorporated into proteins for applications in medicine [31] and bioremediation [30].

Codon reassignment

While ambiguous decoding has long made it possible to produce nsAA-containing proteins, only recently has the translation function of a codon been unambiguously reassigned, enabling the sustained expression of proteins containing one or more nsAAs [13,63,64]. While a surprisingly small number of changes permit the disruption of UAG termination [63,65], the remaining natural UAG codons provide a selective pressure for efficient UAG translation. This destabilizes the genetic code by selecting for spontaneous suppressor mutations that incorporate canonical amino acids at UAG codons [13]. This strategy could prove even more problematic for sense codon reassignment, since stop codons only occur once at the end of genes, limiting the impact of codon reassignment on the proteome [1]. Therefore, the most general strategy to expand the genetic code using reassigned codons
involves (1) identifying all genomic instances of a target codon, (2) replacing them with synonymous codons, (3) abolishing the target codon’s natural function by inactivating its translation factors, and (4) introducing new translation function by integrating orthogonal translation systems, and (5) introducing new instances of the target codon to specifically and efficiently incorporate nsAAs into desired proteins (Figure 2A–D) [13]. Using this strategy, expanded genetic codes can be stabilized by redesigning essential proteins to functionally depend on a specific nsAA for survival [34,35]. However, it remains a major biochemical, genetic, and technical challenge to reassign codons that are commonly utilized throughout a genome.

**Codon creation**

Beyond repurposing one or more of the existing 64 codons, it may be possible to add a new base pair [66–70] or to engineer quadruplet [14–16,71,72] or quintuplet [73] genetic codes, which could give $6^3 = 216$, $4^4 = 256$, or $5^4 = 625$ codons, respectively. Indeed, exciting progress has hinted at the promise of creating new codons, which would need to be replicated, transcribed, and translated. Several unnatural base pairs exhibiting high fidelity replication by PCR and compatibility with proofreading mechanisms of Exo+ polymerases have been developed [67–70]. Additionally, transcription (using T7 RNA polymerase) [74–77] and reverse transcription [77] have been demonstrated. Finally, codons containing unnatural base pairs have been implemented to translate peptides containing unnatural amino acids using an E. coli-derived in vitro translation system [17,78]. This means that codons containing unnatural base pairs can be immediately implemented for in vitro translation of proteins containing nsAAs.

Recently, Malyshev et al. [79] have taken a crucial step toward in vivo implementation of unnatural base pairs with the demonstration that the d5SICS-dNaM base pair can be replicated in vivo [79]. Still, several major challenges must be overcome to fully implement an unnatural base pair for in vivo translation. Malyshev et al. [79] demonstrated that the bioavailability of nucleoside triphosphates is crucial and that heterologous transporters provide one solution to this problem. Additionally, stability is crucial to prevent loss of information. Replication error rates below $10^{-3}$ per bp per replication have been recommended for PCR [80], but replication fidelity better than $10^{-8}$ per bp per replication may be necessary in vivo in the absence of a strong selective pressure to maintain the unnatural base pair (as has been demonstrated for nsAAs [34,35]). The fidelities of transcription and translation are more flexible as long as they do not interfere with normal cell function, but net translational fidelity should be comparable with that of current suppression systems [81,82]. Finally, unnatural base pairs must be compatible with essential components of the host replication, transcription, and translation machinery. Fortunately, previous studies have demonstrated that codons containing unnatural base pairs are compatible with the E. coli translation system reconstituted in vitro [17,78]. However, while thermostable PCR enzymes have been used for replication in vitro [67–70] and E. coli PolI has been implemented in vivo [79], compatibility with E. coli PolIII replication has yet to be demonstrated. Additionally, T7 RNA polymerase is routinely used for transcription in vitro [74–77], and E. coli RNA polymerase transcription has yet to be demonstrated.
Similarly, improved systems for quadruplet decoding have improved translation yields to a level rivaling reassigned triplet systems [16,83]. However, wobble rules for quadruplet and quintuplet systems are not yet well-understood [84,85], requiring an increased reliance on empirical validation of anticodon specificities. Furthermore, triplet codons must still be removed to prevent ambiguous decoding in a genetic code that utilizes both triplet and quadruplet codons [16]. Taken together, significant challenges must still be overcome to implement sustained in vivo translation of novel codons. We expect there to be many exciting advances in this field over the next several years.

Biochemical barriers

Ribosomes translate proteins by adding an amino acid to the nascent polypeptide in response to three-nucleotide codons on messenger RNAs (mRNAs). The identity of the amino acid is controlled at multiple discrete steps. First, aminoacyl-tRNA synthetases (aaRSs) charge transfer RNAs (tRNAs) with their correct amino acids. The aminoacyl-tRNA is then shuttled into the active site of the ribosome by elongation factor Tu (EF-Tu), where base pairing between the mRNA codon and tRNA anticodon allows transfer of the amino acid onto the nascent peptide chain regardless of the amino acid identity. Translation involves more than 100 proteins and RNAs, a subset of which have been engineered to expand the genetic code (e.g., [33,51–53,62,86,87]) (Figure 3). Together with insights from natural non-canonical genetic codes [1,88–91], this work suggests that the biochemistry of the genetic code is remarkably flexible.

Studying natural codon reassignment can provide insights into how to change the genetic code. Interestingly, many of the same codon reassignments appear to have independently evolved several times, suggesting that certain codons have a predisposition for reassignment [1,92]. Stop codons may be favored because they are only used once at the end of genes, so their reassignment is expected to cause minimal damage to the proteome [1]. Indeed, suppression of the stop codons is well-tolerated in E. coli [12] and broadly observed in metagenomic data [5]. Alternatively, mitochondrial genomes are under strong selective pressure to reduce genome size and G+C content, which can lead to the genomic depletion of certain codons [2,3]. When this occurs, small tweaks to post-transcriptional anticodon modifications can change codon assignments without affecting aaRS recognition [1,92–94]. For example, a 7-methylguanosine modification on tRNA*_{GCU} allows it to decode all four AGN codons in squid mitochondria [1]. In this way, small mutations in tRNA anticodons or changes to their post-transcriptional modifications can change the genetic code [92]. Our analysis will focus on the minimal and maximal variants of the canonical genetic code using E. coli as an example (Figure 4). Although the proposed genetic codes may be far from optimal [11,92,95] and challenging to implement, it is instructive to consider how to change the number of unique anticodons that can be achieved based on existing translation machinery.

Minimal genetic code

E. coli strain MG1655 has 87 tRNA genes (with 42 unique tRNA anticodons; including fMet and selenocysteine) and two release factors that unambiguously decode all 64 codons and incorporate all 20 canonical amino acids (Figure 4A) [94]. In contrast, a minimal
genetic code requires one tRNA for each of 20 amino acids, a formylmethionine tRNA
fMet for translation initiation [96], and a release factor for translation termination (Figure 4B).
First, Cys, Trp, Met, fMet, Asp, Glu, Lys, Asn, Gln, His, and Tyr could use one of their
natural tRNAs without any anticodon modifications. This is particularly helpful for E. coli
tRNA\(^{Glu}\), tRNA\(^{Lys}\), and tRNA\(^{Gln}\), which utilize mm\(^5\)S\(^2\)U to recognize their full
complement of codons [93]. Therefore, even though GluRS requires mm\(^5\)S\(^2\)U for efficient
glutamylation [97], no aaRS/tRNA engineering is necessary for a single tRNA (naturally
modified with mm\(^5\)S\(^2\)U) to decode all necessary codons for these amino acids. Ile requires
two tRNAs with non-redundant anticodons, and selenocysteine is not essential in E. coli.
Additionally, as demonstrated by mitochondrial genetic codes, an unmodified uracil in the
anticodon wobble position can recognize all four codons in a family group (codons that are
identical at the first two positions and differ at the third position) [89]. Therefore, 10 tRNAs
with a uracil in the anticodon wobble position could unambiguously assign 40 codons to
decode 9 amino acids (Arg and Ser each have six codons, so the AGN family group is
redundant; see Figure 4B). This leaves all three stop codons, which can be recognized by a
single E167K release factor 2 variant [98]. Therefore, simply by mutating the anticodon
wobble position to uracil in 10 tRNAs and deleting all redundant tRNAs, a minimal genetic
code would require 23 tRNAs and one release factor in order to decode all 64 codons. More
radically, it may be possible to achieve adequate protein function using a code composed of
fewer than 20 amino acids [99–102]. Preliminary studies propose that Ile [103] and Trp
[104] could be replaced by natural amino acids that have similar side chains. If Ile and Trp
are removed and blank codons are tolerated, the minimal genetic code would require only 19
tRNAs and one release factor (Figure 4B).

Maximal genetic code

To expand the genetic code, codons that are naturally used for canonical amino acids must
be reassigned to have new functions. This requires the inactivation of natural translation
machinery and the introduction of orthogonal aaRS/tRNA pairs that can decode the
reprogrammed codons [62]. Despite the complexity of protein translation, simply modifying
tRNA anticodons could free up enough codons to more than double the number of amino
acids in the genetic code. While many different strategies could expand the genetic code, we
propose three tiers of tRNA manipulations to systematically maximize the number of unique
codon assignments (Figure 4C–D).

Tier 1—Simply by leveraging the degeneracy of genetic code, up to ten anticodons could be
reassigned in order to provide seven unambiguous and three ambiguous anticodons for
nsAA incorporation (Figure 4C). GUN, GCN, and CCN were not included in this list, since
the cmo\(^5\)U-containing anticodon has been empirically shown to recognize all four codons in
their respective family groups [94]. Although reassigning the codons chosen in Figure 4C
require considerably more genome modifications in E. coli MG1655 compared to the most
efficient strategy, we chose these codons because they can be further fractionated by
changing their anticodon modifications (Figure 4D). A conceptually simpler strategy is
outlined in Figure S1. These changes are likely adequate for complete genetic isolation from
multiple viruses and horizontal gene transfer.
Tier 2—Inconveniently, six family groups (CUN, GUN, UCN, CCN, ACN, and GCN) use anticodons with overlapping codon specificity, making it difficult to unambiguously reassign their functions. The overlapping tRNA specificities are caused by cmo\(^5\)U wobble bases, which are able to base pair with A, G, U, and sometimes C [105]. Unlike these six family groups, UUN and GGN use alternative wobble bases that allow two unambiguous anticodons in each family group. By inactivating tRNA-modifying enzyme CmoB (modifies U34 of tRNAs with cmo\(^5\)U) [105] and reengineering MnmE and MnmG to modify U34 of additional tRNAs (naturally modifies tRNA\(^\text{Gln}\), tRNA\(^\text{Lys}\), tRNA\(^\text{Glu}\), and tRNA\(^\text{Arg}\) with mnm\(^5\)) [106,107], six ambiguous codons could be disambiguated, accommodating up to 33 total amino acids (Figure 4D, blue features). Although no attempts to reengineer the specificity of tRNA-modifying enzymes have been reported, genetic code expansion might provide an incentive to try.

Tier 3—In wild type \(E.\ coli\), the AUA (Ile) and AUG (Met) codons are unambiguously decoded. Similarly, the other NNR codons could be split into unique singlet codons by exploiting anticodons modified with lysidine (specifically base pairs with A) and cytosine (specifically base pairs with G) to decode NNA and NNG codons, respectively (Figure 4D, magenta features). In order to accomplish this, TilS would need to be engineered to lysidinylate more anticodons in addition to its natural target, tRNA\(^\text{Ile}\) [108]. Although wobble codons do not usually coincide with tRNA identity determinants, lysidine is a crucial identity determinant for IleRS and a crucial antideterminant for MetRS [109]. Therefore, it could potentially impact the orthogonality of heterologous tRNAs introduced for genetic code expansion. Additionally, mnm\(^5\)-modified wobble bases are minor tRNA identity determinants for GluRS, GlnRS, and LysRS [110]. Reduced aminoacylation could be addressed by co-evolving the aaRS/tRNA pair to better recognize the modified anticodon loop [111]. Finally, by engineering one of the release factors to terminate translation exclusively at UAA codons, both UAG and UGA could be reassigned to incorporate nsAAs. Combined with the changes proposed for tier 1 and tier 2, these proposed changes could produce 27 new anticodons (up to 47 total amino acids). While the NNY codons cannot be split into singlet codons based on known anticodon modifications, such modifications have not been ruled out.

**aaRS/tRNA reassignment to expand the amino acid repertoire**

The aaRSs are evolutionarily ancient enzymes crucial for transmission of the genetic message [112]. They display superb specificity against all metabolites in the cell, and possess editing and proofreading mechanism to correct mistakes in aminoacylation [113]. However, they show little specificity against the large number of nsAAs that are used in genetic code expansion studies [114,115]. This ‘polyspecificity’ is observed in all the orthogonal aaRS/tRNA systems that have been developed [116–119]. This fact must be kept in mind when the design of better aaRSs is planned, and will be an issue in the production of proteins containing multiple different nsAAs. In such cases, it may be necessary to further evolve aaRS/tRNA pairs not simply to be orthogonal from the organism’s natural aaRS/tRNA systems, but also to be mutually orthogonal from each other [111].
Once codons have been liberated for reassignment, new translation machinery must be introduced in order to expand the amino acid repertoire. Orthogonal aaRS/tRNA pairs have been evolved to specifically and efficiently incorporate more than 167 nsAAs into proteins, while minimizing interactions with endogenous aaRS/tRNA pairs (extensively reviewed [4,28,33,120]). In some respects, the ability to incorporate more than 167 nsAAs with diverse functional groups suggests that creating custom orthogonal aaRS/tRNA pairs is a solved problem. However, nearly all nsAAs are analogs based on a small number of natural amino acids, and they have been developed for a small number of tRNA anticodons that lack strong aaRS specificity determinants. Furthermore, activity and specificity remain low compared to natural aaRS/tRNA pairs [121,122], mainly because current over-expression systems [81,82] do such a great job of outcompeting natural codon function that they compensate for low aminacylation activity. This results in orthogonal translation systems that are adequate for batch protein production, but are suboptimal for sustained translation with nsAAs [34,121]. Especially in strains with unambiguously reassigned codons [13], reducing expression levels to more closely match natural aaRS/tRNA pairs could provide an incentive to evolve orthogonal systems exhibiting superior aminacylation efficiency and specificity without having to change the standard selection systems (see ref. [33]) used for orthogonal aaRS optimization. Taken together, aaRS/tRNA engineering is certainly not a solved problem.

In addition to reducing aaRS/tRNA expression levels, engineering improved aminacylation efficiencies and specificities will require more sophisticated design methods. To date, aaRS directed evolution libraries have been constructed based on manual inspection of crystal structures, and the combinatorics have limited the number of residues that can be randomized to those that directly contact the ligand (six fully randomized residues is \(20^6 = 6.4 \times 10^7\) unique sequences). Given that second and third shell interactions crucially stabilize binding competent conformations [121,123], expanding the search space could lead to better-optimized enzymes. Although enzyme design remains extremely difficult, notable successes [124–126] provide encouragement that protein design software [127] could narrow the search space enough to probe additional residues in the second and third coordination shells of nsAA binding pockets, while maintaining realistic population sizes for directed evolution. More complex mechanisms of amino acid and tRNA determination, such as the extensive hydrogen bond network in *E. coli* GlnRS [128], have been addressed by transplanting key elements from other aaRSs and tRNAs to alter specificity [121]. Taken together with the robustness of aaRSs (extensive redesign of aaRS cores is tolerated [34,121]), structural and mechanistic insights can provide actionable information for rational aaRS/tRNA engineering [129].

Fidelity during protein translation requires all aaRSs to be selective both for their amino acids and their tRNAs [129]. In order to meet these requirements, orthogonal aaRS/tRNA pairs must have tRNA identity determinants that differ from those in the target organism [62]. Unfortunately, the majority of *E. coli* tRNAs have identity determinants in their anticodons [130]. Therefore, isolating orthogonal tRNAs that are not charged by endogenous *E. coli* aaRSs is a considerable challenge. For instance, a heterologous tRNA\(^{\text{Pyl}}\)\(_{\text{CCG}}\) is mischarged by ArgRS, presumably due to anticodon recognition [131]. In contrast,
suppressors of the UAG stop codon have been more successful, since \textit{E. coli} lacks an aaRS that recognizes the CUA anticodon. For this reason, the \textit{M. jannaschii} TyrRS/tRNA\textsubscript{CUA} [28,33,62] and the \textit{M. barkeri} PylRS/tRNA\textsubscript{Pyl}\textsubscript{CUA} [132,133] pairs have become the most popular systems for nsAA incorporation in \textit{E. coli}, although orthogonal aaRS/tRNA\textsubscript{CUA} pairs have also been isolated for Lys [14], Glu [134], Leu [135], Pyl [136], Trp [137,138], Pro [139], and His [140]. Encouragingly, the \textit{E. coli} tRNA identity determinants have been extensively characterized [110,130], and their antideterminants have been predicted [141], providing a starting point for the directed evolution of additional orthogonal aaRS/tRNA pairs. Additionally, preliminary work with the orthogonal selenocysteine translation machinery provides encouraging results for 60 of the 64 codons [142,143].

Although allosteric amino acid discrimination mechanisms [144], tRNA anticodon recognition [130], and editing domains [129] may still prove difficult to overcome while reengineering aaRSs, metagenomics offers a rich source of aaRS/tRNA pairs that can differ considerably among organisms separated by large evolutionary distances [129]. From a synthetic point of view, domesticating and reengineering only a small subset of the available aaRS/tRNA pairs available in nature is enough to radically expand the genetic code.

**Optimal genetic codes minimize the impact of mutational and translational errors**

Even if codons can be liberated and reassigned by orthogonal aaRS/tRNA pairs, the error minimization theory suggests that extensively changing the genetic code could be deleterious [11]. Similar amino acids are grouped with similar codons in the canonical genetic code to minimize the effects of spontaneous mutations and translation errors, smoothening the fitness landscape to facilitate evolution [145,146]. The canonical code does this remarkably well [95,147], utilizing all 64 codons for translation throughout the proteome and providing a disincentive for genetic code expansion [116]. While the natural persistence of less-optimal genetic codes [92,95] demonstrates that error minimization is not strictly essential, it nevertheless suggests that GROs should be made explicitly dependent on the expanded genetic code [34,35] in order to overcome the countervailing evolutionary pressure for error minimization. It remains to be seen to what extent expanded genetic codes will compromise fitness and whether the resulting GROs will evolve more accurate ribosomes.

**Genetic barriers**

Despite the impressive biochemical flexibility of the genetic code, our inadequate understanding of how to design genomes remains a major barrier for creating organisms with radically new genetic codes. Even in the age of chemically synthesized chromosomes [148–150], genomes must be designed based on incomplete information, and even the best-annotated genomes remain incompletely understood at all levels of complexity from single nucleotide variants to genome architecture (Figure 5).

**Single nucleotide variants**

To a large extent, the effect of single nucleotide mutations can be predicted based on how they change the coding DNA sequence—frameshifts, premature stop codons, and non-
conservative amino acid changes are more likely to disrupt function than are synonymous changes [151]. From this perspective, synonymous codon swaps should be unlikely to be deleterious [13,39]. However, while most synonymous mutations are well-tolerated [13,39], disrupting overlapping sequence features such as ribosome binding motifs [152] and small RNAs [153] can impact crucial cellular functions. Additionally, synonymous mutations are sometimes unpredictably rejected even when non-synonymous mutations are not [39], perhaps due to mRNA structure [154,155] or codon usage preferences [155,156]. Metagenomic conservation data provides a valuable opportunity to de-risk specific mutations [157], but only when the new sequence closely resembles a natural sequence.

### Isolated genetic components

Complex systems can be better understood by breaking them into component parts and establishing models to predict the function of each part. Automated annotation remains a major challenge, requiring empirical testing to characterize the function of genetic components. Furthermore, while recent advances have improved our understanding of promoters [158,159], ribosome binding sites (RBSs) [159,160], and terminators [161], sequence context can strongly impact function [155], requiring empirical testing for each genetic circuit.

### Isolated genetic pathways

Improved understanding of genetic parts has made it possible to refactor pathways in order to replace cryptic regulation mechanisms with more modular and predictable systems [162–164]. The premise of refactoring is that the functions of non-coding DNA sequences are difficult to predict, so it is easier simply to replace them. This becomes especially important when the goal is to transfer such pathways into heterologous hosts. However, since it remains difficult to predict how gene expression translates to pathway behavior, the iterative process of testing and redesigning biological pathways remains an important strategy for building systems from defined genetic parts [163,164].

### Genome construction

Engineering at the genome scale encompasses all of the challenges of engineering its parts plus the added challenge of preserving essential cellular functions, replication, transcription, DNA structure, and DNA repair. Some notable studies have taught important lessons about engineering genomes. For example, the separation of overlapping coding DNA sequences in T7 bacteriophage [162] provided a proof of concept for modular genomes that are easier to manipulate. The removal of mobile DNA elements and cryptic virulence genes from *E. coli* produced a more genetically stable genome [165]. Integration of the *Synechocystis* PCC6803 genome into *B. subtilis* demonstrated the importance of balancing replichore size, while highlighting the challenges of preserving the function of each species as part of a chimeric genome [166]. Finally, the introduction of large genome rearrangements was used to explore the effects of splitting genomes into multiple chromosomes [167] and to demonstrate the importance of co-orientating transcription with replication in circular chromosomes [168]. Future work will characterize additional crucial genome-scale features that are essential for viability (e.g., although each individual instance
of DNA gyrase [169] and chi [170] sites may not be essential, their presence throughout the genome is crucial for genome maintenance). While it has been difficult to produce a minimal genome based on information learned from single gene knockout studies [171], continued work in this area [7,148,171–175] is likely to uncover additional genome design rules involving synthetic lethality (two mutations are tolerated individually, but are lethal when combined [176]), operon structure, and genome structure.

### Genome engineering barriers

Despite much progress, it remains difficult to predict the correct changes to make at every level of genome complexity from single nucleotide changes to megabase/gigabase genome construction. Accepting the fact that existing information is inadequate, draft genomes must be a best guess based on as much information as possible. To put this in perspective, a logical next step in genome recoding is to reassign all instances of the rarest sense codons in E. coli [39]. However, with 4228 AGR codons, exhaustive sampling of all synonymous CGN codon alternatives would require testing of $4^{4228}$ genomes.

Given our tenuous understanding of how to design genomes, effective genome engineering technologies must integrate the information that is known, and overcome the inevitable design flaws that will arise based on our incomplete knowledge [39,177]. We know that all natural instances of a codon must be removed from the genome in order to abolish its natural translation function [13]. We know that orthogonal aaRS/tRNA pairs [62] can introduce new translation functions [28]. We know that we can stabilize an expanded genetic code by establishing functional dependence on an unnatural amino acid [34,35]. The challenge is to produce such genomes by making hundreds or thousands of changes without introducing any lethal design flaws.

The past decade has seen many impressive achievements in genome engineering (reviewed in [178]), although few attempts have been made to produce new sequences that cannot be found in nature. The de novo synthesis and transplantation of an intact Mycoplasma mycoides JCVI-syn1.0 genome demonstrated that a small, natural prokaryotic genome can be built from simple chemical components [148]. Such an approach could allow the synthesis of any user-defined genome sequence. However, genome design remains the major barrier because even a single design flaw could prevent the function of the entire genome [148]. Given the high stakes for design flaws, de novo genome synthesis is most effectively used in combination with exhaustive empirical tests [7,171,172,175] and complete computational models [179].

### Genome engineering technologies

Engineering the genetic code requires extensive genome manipulation that can affect fitness in unpredictable ways [39]. With this in mind, we have developed multiplex automated genome engineering (MAGE) [180] and conjugative assembly genome engineering (CAGE) [181] for rapidly prototyping and manufacturing genotypes in vivo. MAGE uses the λ bacteriophage β recombinase and ssDNA oligonucleotides [182] to simultaneously introduce multiple defined mutations at multiple locations throughout a replicating bacterial genome [180]. Meanwhile, CAGE uses bacterial conjugation to precisely transfer up to several
million base pairs of contiguous DNA [181], allowing the production of extensively modified genomes from small segments that are easily prototyped in parallel using MAGE. Together, MAGE and CAGE exploit evolution to combinatorially explore a broad pool of synthetically defined genotypes \textit{in vivo}, allowing natural selection to remove deleterious design flaws from the population.

MAGE and CAGE were used to remove all 321 known instances of the UAG codon from \textit{E. coli} MG1655 at a fraction of the predicted cost for genome synthesis [13]. Still, DNA synthesis can be invaluable for extensively modifying genome sequences, provided that the synthetic genome fragments are small enough for efficient troubleshooting. For instance, we tested 6496 total mutations across 42 essential genes [39] using inexpensive, chip-based DNA synthesis [22]. Because we tested each essential gene individually, design flaws could be rapidly mapped and overcome using MAGE [39]. A similar strategy has been successful for the synthetic yeast 2.0 project [150] and could be extended to diverse organisms using an evergrowing arsenal of powerful genome engineering tools [183].

As genome designs increase in complexity, integrated CAD/analysis strategies will become essential for monitoring design clashes, managing genome builds, and analyzing genotypes. There are many useful genome engineering design and analysis tools available: searchable genome annotation databases [184], automated MAGE oligo [181] or CRISPR [185] design tools, sequence manipulation and synthetic circuit design tools [186], sequencing analysis tools for single nucleotide variants [151,187] and structural variants [188–190], and computational models of whole organisms [179]. Integrating these design and analysis tools into a cohesive and efficient software platform will greatly benefit efforts to produce GROs with radically altered genetic codes.

\section*{Outlook and conclusions}

While more than 167 nsAAs have already vastly expanded protein function [28], radically different genetic codes will be required to achieve virus resistance, genetic isolation, and stable expansion of the genetic code. Thirteen out of thirteen codons tested have already shown promise for reassignment [39]. To implement radically expanded genetic codes, a mechanistic understanding of biochemical principles will be crucial to engineer orthogonal translation machinery that is capable of reassigning such sense codons. Additionally, genome engineering methods capable of interrogating genetic landscapes containing thousands of potentially deleterious changes will be crucial for producing organisms with reassigned sense codons [13,39]. Advances in understanding codon usage [155,191], gene function [171,172,175], operon structure [159,163,164], and genome architecture [162,165–168] will help establish better guidelines for genome design, but diversity will remain a crucial aspect in prototyping genomes with new and useful biological functions.

\section*{Supplementary Material}

Refer to Web version on PubMed Central for supplementary material.

\textit{J Mol Biol.} Author manuscript; available in PMC 2017 February 27.
Acknowledgements

We thank J. Ling, F. Isaacs, D. Mandell, A. Chatterjee, M. Jewett, A. Forster, M. Sismour, M. Napolitano, and K. Lajoie for helpful comments. Funding was from the U.S. Department of Energy (DE-FG02-02ER63445 to G.M.C. and DE-FG02-98ER20311 to D.S.), NSF (SA5283-11210 to G.M.C.), NIH (GM22854 to D.S.), and Defense Advanced Research Projects Agency (N66001-12-C-4040 to G.M.C., N66001-12-C-4020 to G.M.C. and D.S., N66001-12-C-4211 to G.M.C. and D.S.).

Abbreviations

GRO genomically recoded organism
nsAAs non-standard amino acids
aaRSs aminoacyl-tRNA synthetases
MAGE multiplex automated genome engineering
CAGE conjugative assembly genome engineering

References


• Barriers to changing the genetic code
  – biochemical barriers
  – biological knowledge (ability to design functional genomes)
  – biotechnology (ability to create the genomes)
• Approach to engineer genomes with new biological functions
Figure 1. Expanded biological functions of GROs
GROs provide dedicated codons for efficient translation of nsAAs at multiple sites in a protein [13,64], increase resistance to bacteriophages [13], and enable biocontainment/niche restriction [34,35].
Figure 2. Engineering a GRO and its properties
GROs are produced by (A) identifying all instances of a target codon (e.g., UAG = stop) in the genome of the starting organism, (B) replacing all instances of the target codon with a synonymous codon (e.g., UAA = stop), (C) deleting translation factors responsible for decoding the target codon (e.g., release factor 1 terminates translation at UAG and UAA codons), (D) introducing orthogonal translation machinery capable of decoding the target codon with a nsAA and then reinserting the reassigned codon into the genome [13].
Figure 3. Components of the translation system have been engineered in vivo
The amino acid repertoire has been expanded by more than 167 nsAAs through directed evolution of orthogonal aaRS/tRNA pairs [28,33]. Amino acid specificity has been tuned by directed evolution of the aaRS amino acid binding pocket [62]. Meanwhile, target codon specificities have been controlled by altering the acceptor stem of the tRNA [14], or by mutually evolving the tRNA anticodon and the aaRS anticodon loop binding domain [111]. Additionally, orthogonal ribosomes have been engineered to exhibit useful properties that would also be deleterious for translating the rest of the proteome. An orthogonal 16S rRNA with a modified anti-Shine-Dalgarno sequence has been evolved to promote better UAG [51] and AGGA [15] suppression, an orthogonal 23S rRNA has been modified to accept orthogonal tRNAs with modified 3’ ends [52], and a 16S-23S tethered rRNA has been developed to facilitate engineering of peptidyl transferase activity [53]. Finally, release factors have been engineered to recognize additional codons [98], and EF-Tu has been modified to accept highly charged amino acids [87]. Although, tRNA-modifying enzymes such as CmoB, MnmE, and MnmG have not yet been reengineered, they offer the potential to radically alter the genetic code (see Figure 4).
Figure 4. Minimal and maximal genetic codes using triplet codons composed of four nucleotide types (U, C, A, G)

The proposed genetic codes are one possible permutation representing several possible ways to reassign redundant codons. Dashed brackets represent codon recognition ranges for a given anticodon: black is codon recognition agreeing with wobble rules [192,193]; gray is empirical data that supersedes wobble rules [94]; blue and magenta are proposed new tRNAs that can be assigned to nsAAs. Labels correspond to the wobble nucleotide at tRNA position 34 (cmo$^5$U = uridine 5-oxyacetic acid, mnm$^5$U = 5-methylaminomethyluridine, cmmn$^5$U = 5-carboxymethylaminomethyluridine, cmmn$^5$Um = 5-carboxymethylaminomethyl-2′-O-methyluridine, mnm$^5$s$^2$U = 5-methylaminomethyl-2-thiouridine, cmmn$^5$s$^2$U = 5-carboxymethylamino-methyl-2-thiouridine, I = inosine, k$^2$C = lysidine, Q = queuosine,
GluQ = glutamylqueuosine) [173]. Green letters indicate natural tRNA identity determinants that may be difficult to change without disrupting aminoacylation. Red letters indicate natural anticodon modifications that increase anticodon promiscuity. Blue and magenta letters represent proposed changes in the tRNA wobble position that would alter codon recognition. Amino acid assignments are indicated in the yellow sidebars. M refers to both Met and fMet (translation initiation). Codons available for new amino acids are indicated by blue and magenta boxes with white numbers. Selenocysteine is not shown.

(A) The E. coli genetic code is presented based on Björk et al. [94] and tRNA identity determinants are from Giegé et al. [110]. All 64 codons are used to encode 20 amino acids.

(B) A minimal genetic code utilizing all 64 codons would require initiation at AUG, one release factor (RF2 E167K mutants can terminate all 3 stop codons [98]), and one tRNA for each of the 20 amino acids. Unmodified uracils in the wobble positions would allow tRNAs to recognize all codons in a family group, allowing redundant tRNAs to be deleted. Gray shaded boxes represent additional anticodons that could be potentially deleted (tRNA^Arg^UCG would encode the same amino acid as tRNA^Arg^UCU and it may be possible to remove Ile [103] and Trp [104] from the genetic code). Conveniently, the wobble nucleotide is rarely a tRNA identity determinant [110]. The two relevant exceptions, tRNA^Phe^GAA (G34) and tRNA^Glu^UUG (cmnm^5^s^2^U34), are weak identity determinants [110], so the proposed changes may be tolerated by their respective aaRSs.

(C) The genetic code can be expanded to provide 7 unambiguous and 3 ambiguous anticodons by simply deleting tRNAs and introducing orthogonal aaRS/tRNA pairs encoding new amino acids. This analysis assumes that the original aaRS/tRNA identity determinants/antideterminants can be overcome by a metagenomic search for an orthogonal aaRS/tRNA pair and subsequent directed evolution to optimize their orthogonality. Red shaded boxes represent the three codons that would gain ambiguous translation function upon introduction of an orthogonal aaRS/tRNA pair. GUN, GCN, and CCN were not included, since the cmo^5^U-containing anticodon has been empirically shown to recognize all four codons in their respective families [94]. The UAG codon has already been reassigned [13].

(D) Replacing the cmo^5^U and inosine wobble nucleotides with mmn^5^U nucleotides could liberate 13 unambiguous anticodons for reassignment (33 total amino acids; changes indicated in blue). Doing so would require the inactivation cmoB [105] and the engineering of mmnE and mmnG to recognize additional tRNAs [106,107]. Taken a step further, the maximal genetic code would have unique amino acid assignments for all NNA and NNG codons (NNA: engineer tilS [108] to lysidylate additional tRNAs so that they only base pair with A; NNG: change the tRNA wobble bases to cytosine so that they only base pair with G). Anticodon modifications capable of splitting NNY codons into unambiguous NNU and NNC codons have not been reported, but such modifications have not been ruled out. Additionally, it may be possible to engineer a release factor to terminate translation only at UAA codons, thereby liberating both UAG and UGA codons. The proposed changes would liberate 27 unambiguous anticodons (47 total amino acids; changes indicated in magenta). This strategy may require directed evolution to overcome the tRNA identity determinants for Glu, Gln, and Lys [110]. Another potential complicating factor is that G + C anticodon content may affect cognate and near-cognate decoding efficiencies, just as the G + C rich anticodons for Val, Ala, and Pro break the wobble rules [94].
Genome design flaws can impair fitness or alter the desired functions of an engineered organism. Genome engineering can introduce such design flaws by several mechanisms, including intentional genome changes [39,166], spontaneous mutations [13,148], transposition [165], and genome rearrangements [13]. Point mutations produced during genome engineering can introduce frame shifts, cause amino acid substitutions, de-optimize codon usage, or disrupt the function of overlapping non-coding sequences. Genetic parts such as genes and expression signals can impact crucial cellular functions or the desired function of the engineered organism (e.g., nsAA incorporation, virus resistance, biocontainment). Refactoring genetic pathways provides an opportunity to increase modularity [163], but cryptic regulation mechanisms and polar effects make it difficult to design a de novo pathway architecture with optimal activity [164,180]. Finally, while genome-scale design rules will continue to be discovered, we already know that it is important to balance the size of replichores in circular chromosomes [166], to co-orient transcription of essential operons with translation [168], to preserve sequences involved in DNA structure [169] and repair [170], and to consider how chromosome size impacts its structural integrity [167].