

Published in final edited form as:

*Trends Mol Med.* 2014 November ; 20(11): 604–613. doi:10.1016/j.molmed.2014.09.003.

## A critical analysis of codon optimization in human therapeutics

**Vincent P. Mauro** and **Stephen A. Chappell**

Department of Cell and Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA, USA 92037

### Abstract

Codon-optimization describes gene engineering approaches that use synonymous codon changes to increase protein production. Applications for codon-optimization include recombinant protein drugs and nucleic acid therapies, including gene therapy, mRNA therapy, and DNA/RNA vaccines. However, recent reports indicate that codon-optimization can affect protein conformation and function, increase immunogenicity, and reduce efficacy. We critically review this subject, identifying additional potential hazards including some unique to nucleic acid therapies. This analysis highlights the evolved complexity of codon usage and challenges the scientific bases for codon-optimization. Consequently, codon-optimization may not provide the optimal strategy for increasing protein production and may decrease the safety and efficacy of biotech therapeutics. We suggest that the use of this approach is reconsidered, particularly for *in vivo* applications.

### Keywords

codon optimization; gene therapy; mRNA therapy; vaccine; A-to-I editing; tRNA wobble

### Optimizing codon usage for increased protein expression

The polypeptide chain(s) of most proteins can be encoded by a seemingly infinite number of mRNA sequences due to the degenerate nature of the genetic code (see Glossary) [1]. Interestingly, mRNAs encoding the same polypeptide via different codon assignments can vary dramatically in the amount of protein expressed [2, 3]. The attempt to produce more protein by altering codon assignments has led to the broad use of codon-optimized mRNAs for bioproduction of protein pharmaceuticals and nucleic acid therapies. However, considerable evidence demonstrates that synonymous codon choices in natural mRNAs have evolved in response to diverse selective pressures at both the RNA and protein levels [4]. In addition, various studies have shown that synonymous codon changes can have unanticipated effects. Synonymous codon changes may affect protein conformation and stability, change sites of post-translational modifications, and alter protein function [5–9].

© 2014 Elsevier Ltd. All rights reserved.

Corresponding author: Mauro, V.P. (vmauro@scripps.edu).

**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.

Moreover, synonymous mutations have been linked to numerous diseases [4, 10–13]. Some potential risks associated with the use of codon-optimized mRNAs for producing recombinant protein drugs have been discussed recently [11, 12, 14, 15]. These risks include the production of anti-drug antibodies which can reduce drug efficacy and cause allergic reactions.

In this article we critically review the scientific bases for codon optimization and identify additional risks. These include two potentially serious side effects that pose unique risks for applications in nucleic acid therapies: (i) the production of novel peptides from alternative out-of-frame open reading frames (ORFs); and (ii) altered sites of post-transcriptional nucleotide modifications that can lead to the production of novel protein variants and ensembles. Understanding the potential risks of codon optimization so that they can be minimized or eliminated is critical as nucleic acid therapies begin to gain traction. We suggest that the use of these approaches for human therapeutics should be carefully considered to avoid introducing unnecessary problems.

## The genetic code, tRNAs, and wobble

The genetic code is degenerate as most amino acids are encoded by multiple synonymous codons (Figure 1). However, cells and organelles do not express 61 different tRNAs and vary dramatically in the relative expression of individual tRNAs [16, 17]. For instance, in humans around 500 tRNA genes correspond to 48 codons; there are no tRNA genes for the remaining 13 codons [17]. Interestingly, an overlapping but different set of tRNA genes is missing in Chinese hamster ovary (CHO) cells, a cell line that is often used to produce therapeutic proteins. Despite the absence of these tRNA genes, mRNAs use the full complement of codons and synonymous codon usage is not affected by the absence of a cognate tRNA. For example, two codons encode Aspartic acid (D) and have similar codon-usage even though there is no tRNA gene corresponding to the GAU codon. This is possible because of ‘wobble’, which enables both codons to be decoded by the same tRNA.

Wobble involves tRNAs with U or G in position 34, which base pair to the third base in the codon (Figure 2). Some tRNAs with U in position 34 are capable of U-A and U-G base pairing. Likewise, some tRNAs with G in position 34 are capable of G-C and G-U base pairing. In addition, for some tRNAs adenine in position 34 is deaminated to inosine (I), which can base pair to U, C, and A. This occurs for one tRNA in prokaryotes and 7–8 tRNAs in eukaryotes [18, 19]. Modifications at different positions in tRNAs can restrict wobble in some cases and expand it in others [20]. Experimentally, it has been demonstrated that 25 cognate tRNAs comprise a minimum set that can sustain protein synthesis by the use of extended wobble interactions referred to as ‘superwobbling’ [21, 22]. Superwobbling can explain how translation can occur with fewer tRNAs than are predicted by the wobble hypothesis. Although superwobbling has only been demonstrated in chloroplasts to date, its occurrence in mammals is plausible.

## The history, scientific basis, and art of codon-optimization

Degeneracy in the genetic code enabled the first recombinant peptide, a mammalian somatostatin, to be expressed in *E. coli* without knowing the peptide’s mRNA sequence

[23]. A coding sequence was obtained by reverse translating the amino acid sequence. Codon usage was biased with consideration of the effects of various gene sequences on translation and transcription, as well as to facilitate gene synthesis.

When the first gene sequences were determined, it was noted that codons are used in a non-random manner [24–26]. For highly expressed genes in *E. coli* and yeast, the non-random use of synonymous codons was found to be correlated with tRNA abundance [27–30]. The observation that some highly expressed genes preferentially use a subset of codons suggested that codon bias and protein expression are causally linked, and that it might be possible to enhance expression by mimicking the pattern of codon bias of highly expressed mRNAs. This prospect led to the development of numerous codon-optimization programs and commercial services. These approaches differ in how codon bias is measured, the number of variables considered, potential applications, and implementation. However, a general feature of these programs is that they avoid using rare codons, which are thought to decrease the rate of translation elongation. In addition, many programs contain features to facilitate cloning, gene synthesis, and gene modification, as well as to avoid features that may decrease protein expression. Indeed, many codon-optimization programs are not constrained by the natural codon usage of the gene at all, and require only an amino acid sequence as input.

Codon-optimized mRNA sequences that are produced using different programs or approaches can vary dramatically because different codon optimization strategies differ in how they quantify codon usage and implement codon changes. Some approaches use the most optimal (frequently used) codon for all instances of an amino acid, or a variation of this approach [31, 32]. Other approaches adjust codon usage so that it is proportional to the natural distribution of the host organism [32–38]. These approaches include codon harmonization, which endeavors to identify and maintain regions of slow translation thought to be important for protein folding [39]. Alternative approaches involve using codons thought to correspond to abundant tRNAs [40], using codons according to their cognate tRNA concentrations [41], selectively replacing rare codons [35], or avoiding occurrences of codon-pairs that are known to translate slowly [38, 42]. In addition to approaches that vary in the extent to which codon usage is considered as a parameter, there are hypothesis-free approaches that do not consider this parameter [43].

The same flexibility that has enabled scientists to modify codon usage has also enabled evolution to embed multiple levels of information into coding sequences. However, codon optimization most likely disrupts this information.

## Critical analysis of codon-optimization

Codon-optimization strategies for increasing protein expression are based on the assumptions that: (i) rare codons are rate-limiting for protein synthesis; (ii) synonymous codons are interchangeable without affecting protein structure and function; and (iii) replacing rare codons with frequently used codons increases protein production. Below we assess the validity of these various assumptions.

### Assumption 1: Rare codons are rate-limiting for protein synthesis

In bacteria, different codons are translated at different rates. For example, analysis of 29 codons in *E. coli* indicated that aminoacyl-tRNA selection rates vary by up to 25-fold [44]. It has also been reported that overexpression of some recombinant proteins can deplete one or more tRNAs and limit expression [45]. However, there is little evidence to substantiate the notion that rare codons limit protein production in mammalian cells. Even in bacteria, some studies have shown that the translation rates of specific codons do not correlate with either tRNA abundance or frequency of codon use [46, 47]. In one study, increasing expression of tRNAs corresponding to rare codons increased the translation rate, but led to protein misfolding and aggregation [48].

In humans, the extent to which different amino acids are encoded varies by up to 10-fold (Figure 3A); however, a general trend is that amino acids encoded less frequently have fewer synonymous codons than those encoded more frequently. In fact, if the amino acid frequency is normalized to the number of synonymous codons per amino acid, the differences are reduced to 3-fold (Figure 3B). Individual codon usage frequencies also vary by up to 10-fold (Figure 3C). However, whether a codon is rate-limiting is likely to depend on other variables including tRNA levels. If the number of tRNA genes, which varies from 0 to 33 for different tRNA isodecoders, and wobble are both considered, the normalized codon frequencies show a very different distribution (Figure 3D) and calls into question the definition of rare codons.

We suggest that an additional complication is that the number of tRNA genes is not necessarily directly related to tRNA levels. Moreover, codons have been categorized as rare or abundant based on how often each codon occurs in mRNA coding sequences at the gene level, without considering mRNA levels, or tissue-specific differences in expression. Furthermore, codon usage tables do not consider other variables that can affect codon frequencies, including codon usage by out-of-frame or alternative in-frame initiation events, which may dramatically skew codon usage. It is worth noting that there is little evidence that protein synthesis is limited by codons that have been designated as “rare”. This notion is independently supported by studies which indicate that translation initiation, not elongation, is rate limiting for protein synthesis [49]. Based on the preceding argument, it seems likely that codons designated as “rare” may be incorrectly categorized and may not be rate limiting for translation.

### Assumption 2: Synonymous codons are interchangeable without affecting protein structure and function

There are many studies which indicate that this assumption is false [4–13]. One reason is that codon usage is thought to determine the elongation rhythm, which can cause ribosomes to slow down or pause at certain sites, and which may be necessary in some cases for correct protein folding [50, 51]. Although rare codons have been implicated in slowing translation and forming pause sites, the situation appears to be more complex, as other studies have shown that rare codons do not necessarily diminish local translation rates [51, 52].

An alternative explanation for ribosomal pausing is provided by a study showing that translation by cognate and wobble tRNAs occurs at different rates, with wobble pairing occurring more slowly, by up to three-fold in HeLa cells [53]. The authors suggested that wobble-dependent slowing of elongation may have been selected as a mechanism for protein folding as it is largely independent of tRNA levels. Variations in decoding efficiencies may provide a mechanism to fine-tune the temporal pattern of elongation, which may be important for protein conformation.

In *E. coli*, ribosomal pausing involves base-pairing of rRNA to Shine-Dalgarno-like sequences in coding regions of mRNAs [54]. Synonymous codon substitutions in codon-optimized mRNAs may disrupt information encoded in the primary sequence of a gene. For example, by removing complementary matches that affect translation by base pairing to rRNA or other RNAs, including noncoding RNAs such as microRNAs (miRNAs). These types of interactions can affect initiation, shunting, pausing, frameshifting, and reinitiation, as well as mRNA stability [54–60]. In addition to disrupting these types of interactions, codon optimization may unintentionally introduce new RNA binding sites.

A recent screen of protein function has highlighted the fact that synonymous codons are not necessarily interchangeable. This study tested 342 antibody constructs with synonymous codon variants and was able to identify altered expression, solubility, and binding affinities of the antibodies [61]. The effects of synonymous codon changes are further highlighted in a recent study using a fluorescent protein gene that was engineered to have different fluorescent properties depending on the folded structure [62]. This study showed that the fluorescent properties of the protein were altered by synonymous codon changes due to altered protein folding.

### **Assumption 3: Replacing rare codons with frequently used ones increases protein production**

Examples supporting this possibility are anecdotal. For example, expression of phosphoglycerate kinase was diminished when the major codons in the gene were systematically replaced with minor ones [63]. By contrast, expression of an immunoglobulin kappa protein in yeast was increased by replacing more than half of the codons with those used predominantly by abundant proteins [64]. Additional examples are compiled in Table 1 of reference [65]. However, a limitation with these types of studies is that they do not account for numerous variables that may inadvertently affect expression, or indicate whether other codon-optimized variants were tested. Without additional data, it is impossible to determine whether the reported effects were due to altered codon bias or other mechanisms. Other studies do not support the postulated effects of rare codons on protein expression. For example, studies using synonymous variants of the formaldehyde activating enzyme showed that enzyme expression and cell fitness are not correlated with the use of either rare or frequent codons [6]. In addition, other studies suggest other mechanisms. For example, an analysis of ribosomal footprint data indicated that the rate of translation is not slower when ribosomes translate rare codons or clusters of rare codons, but is affected by amino acid charge [66]. Moreover, other factors affecting elongation rate include mRNA secondary structure and adaptation of codons to the tRNA pool [67, 68].

A more empirical investigation of the contributions of different variables to protein expression comes from Plotkin's group [2]. In this study, a library of 154 *green fluorescent protein (GFP)* genes was synthesized. These genes varied randomly in codon usage but all encoded the same GFP protein. When expressed in *E. coli*, fluorescence and green fluorescent protein levels varied by 250-fold across the library. However, there was no correlation between expression level (fluorescence) and codon bias (which was assessed using two measures), or between fluorescence and the number of pairs of rare codons. This study did find a correlation between codon usage and cell fitness, which was lower for cells expressing mRNAs with large numbers of rare codons. The authors reason that codon usage in highly expressed mRNAs affects the numbers of free ribosomes and global translation, which ultimately affects fitness. Fast growing bacteria and yeast both exhibit strong codon bias in highly expressed genes. By contrast, higher eukaryotes exhibit much less codon bias [69]. For highly expressed genes, codon bias is inversely related to species generation time, varying by more than four orders of magnitude, with the lowest bias occurring in mammals. Therefore, even though codon usage in mammals has not yet been studied systematically as in *E. coli*, there is little reason to expect that optimizing codon usage of mammalian genes should enhance protein expression.

The fact that frequently used codons do not cause high expression, even though highly expressed genes (at least in microbes) evolve an optimal codon bias, suggests that codon bias *per se* does not necessarily yield high expression but requires other features (also see [70, 71]). An athletic analogy illustrates this principle. Elite runners have certain features that are optimal for running, including a low body mass index. However, a low body mass index alone does not enable a nonathletic individual to achieve elite runner status without other features, including high endurance, a suitable physiology, and extensive training [72].

## Codon-optimization, novel peptides, unknown consequences

An implicit assumption associated with codon-optimization approaches is that protein expression is limited predominantly or exclusively to the initiation codon of full-length cistron. However, this assumption is unfounded. Translation typically starts from multiple sites within mRNAs, including both in-frame and out-of-frame ORFs, which initiate at AUG and non-canonical start sites. The selected examples discussed below provide evidence for alternative initiation, as well as an indication of its scope and importance.

There are numerous examples of individual mRNAs that initiate translation from multiple start sites; in some cases the alternative ORFs express more protein than the main cistron. In one example, translation was initiated efficiently from two AUG codons in a synthetic mRNA, however the relative use of the downstream AUG codon could be increased dramatically by various factors, including short oligonucleotides and protein complexes which mask the first AUG codon [73]. Important examples of alternative initiation also come from the Major Histocompatibility Complex I (MHC) of the immune system. For example, an antigenic peptide was shown to be translated from an out-of-frame CUG start codon [74]. In addition, ribosome-profiling studies in yeast [75] and mammalian cells [76, 77] have shown that translation initiation complexes form at multiple start sites in eukaryotic mRNAs. These studies demonstrate that up to a third of ribosomes are enriched at multiple

near-cognate initiation codons located within coding sequences. Proteomic studies that enable the isolation of N-terminal peptides provide corroborating data regarding the scope of alternative initiation events [78, 79].

Various mechanisms have been proposed to account for translation initiation at multiple start sites (Figure 4), including leaky scanning and reinitiation [80], and ribosomal tethering and clustering [81]. However, regardless of the proposed mechanism, thousands of known peptides are synthesized, some with biological activities [74, 79]. Codon-optimization can disrupt alternative translation start sites and also generate new sites that encode a different set of novel peptides (Figure 4B). These peptides will vary for different codon-optimized variants of the same mRNA and will likely include numerous novel bio-reactive peptides. Some peptides may trigger immune reactions or interfere with normal cellular functions; some may function as hormones, some may even be toxins. In the case of nucleic acid vaccines, naturally-occurring cryptic peptides that may contribute to a therapeutic immune response, may be lost upon codon-optimization.

## Codon-optimization, altered mRNA editing, new protein ensembles

An additional potential problem arising from codon modifications is the disruption of natural sites of post-transcriptional modifications and the introduction of new ones. Adenosine-to-inosine (A-to-I) editing is the most widespread form of RNA editing in higher eukaryotes and occurs most frequently within noncoding sequences; however, it also occurs in coding sequences [82, 83]. The extent of editing at these sites may change during development, or show cell or tissue-specificity [84, 85]. In addition, edited and non-edited forms of an mRNA may be present in the same cell and generate patterns of protein heterogeneity that contribute to normal cellular physiology.

Sequence changes in RNA due to A-to-I editing can have significant effects on expression because inosine is recognized as guanosine by much of the gene expression machinery. Within mRNA coding sequences, A-to-I editing can lead to amino acid substitutions that may have functional consequences. For example, the brain-specific alternative splicing factor *Nova 1* is spatiotemporally edited during embryonic development in mouse and chicken brains [82]. Although A-to-I editing did not affect *Nova 1* functional activity, protein half-life was increased. Similarly, tissue-specific A-to-I editing appears to affect the proteolytic processing of the human insulin-like growth factor-binding protein 7 (IGFBP7) which yields variants with different biological activities that may alter cell-extracellular matrix interactions [83].

Dysregulated A-to-I editing appears to contribute to the pathogenesis of various diseases including Amyotrophic Lateral Sclerosis (ALS) and various cancers of the nervous system. In many cases changes in global levels of A-to-I editing appear to occur as a consequence of changes in expression of the Adenosine Deaminases Acting on RNA (ADAR) enzymes [86]. Altered levels of A-to-I editing have also been identified in other pathologies with no clear link to expression levels of the ADAR enzymes, including schizophrenia and bipolar disorder [87]. Another example is seen in high-grade gliomas [88], which is particularly interesting since impaired A-to-I editing of microRNA miR-376a\* leads to a single



nucleotide difference in its seed sequence that alters its target specificity from *Autocrine Motility Factor Receptor (AMFR)* to *Ras-Related Protein 2A (RAP2A)*, two mRNAs with opposing roles in regulating cell invasion, resulting in the promotion of glioma cell migration and invasion.

Currently, the most accurate way to identify A-to-I editing sites is by sequencing of RNA populations [84]. Although computational approaches have been attempted to predict editing sites (see for e.g. [89]) they appear to be insufficient for many reasons, most notably the inability to accurately predict higher order structures that form *in vivo*. Consequently, synonymous codon changes in nucleic acid therapeutics may result in unanticipated altered editing of RNA sequences by removing a preexisting editing site and/or introducing a novel editing site. These alterations may change coding potential, splicing, or even affect expression by introducing or disrupting miRNA seed sequences. These changes in codon-optimized mRNAs may generate an altered repertoire of proteins, with immunological consequences. In addition, as with the cryptic protein products described above, there may be unanticipated functional effects or toxicity associated with some variants.

### **An additional consideration: tRNA channeling**

An additional theoretical consideration associated with codon-optimization involves tRNA channeling. Experiments performed in yeast showed that once a particular codon is used, subsequent occurrences of codons for the same amino acid do not occur randomly, but favor those that use the same tRNA [52]. This effect is most pronounced in rapidly induced genes and involves both frequent and rare codons. Furthermore, they showed that codon correlation accelerates translation elongation.

### **Concluding remarks**

Codon-optimization is often suggested as a primary consideration for generating high-expressing constructs suitable for gene therapy and genetic vaccines. Although protein expression can be increased using these approaches, it is evident, however, that mRNAs contain numerous layers of information that overlap the amino acid code and that this complexity can be disrupted by codon-optimization. A more serious problem is that the scientific basis of codon optimization, at least in mammals, does not support that codon usage is rate limiting for protein expression. In addition, there are potentially serious consequences associated with using codon-optimization, particularly for nucleic acid therapeutics. In the absence of analysis; these potential problems include: (i) disrupting the normal patterns of cognate and wobble tRNA usage, affecting protein structure and function; (ii) producing novel peptides with unknown biological activities; and (iii) altering post-transcriptional modifications that may modify protein ensembles. In view of these problems, we suggest further studies of current codon-optimization approaches for genetic therapeutics and vaccines as the potential risks of these approaches to patients may outweigh their usefulness (Box 1). One may be tempted to suggest that constructs for human therapeutic applications should be restricted to unmodified natural gene sequences; however, such a restriction may be too limiting, particularly for genes that express poorly. We expect that a more constructive path forward will include testing the possible effects of codon



optimization, which may include mass spectrometry analysis of cryptic peptide expression from constructs intended for *in vivo* nucleic acid therapies. In addition, the development of new approaches that may involve introducing a minimal number of specific targeted modifications into natural gene sequences in order to address specific potential problems that may decrease expression, for example splice sites, while avoiding problems that may arise from wholesale changes, including the introduction of novel ORFs.

#### Box 1

##### Outstanding questions

- To what extent does codon usage specify protein conformation?
- Can codon usage be altered to increase expression and maintain protein structure?
- What codons are limiting for translation in different cell types?

## Acknowledgments

We acknowledge the late Dr. Gerald M. Edelman and Drs. Joseph Gally and Kathryn Crossin for critical discussions and valuable comments. Funding was provided by the National Institutes of Health (GM078071) and Promosome, LLC (SFP 1539).

## Glossary

|                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
|---------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Cistron</b>            | historically refers to a gene. In more contemporary usage, a cistron refers to the nucleic acid sequence that encodes a polypeptide chain                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
| <b>Codon optimization</b> | refers to experimental approaches designed to improve the codon composition of a recombinant gene based on various criteria without altering the amino acid sequence. This is possible because most amino acids are encoded by more than one codon. Most codon-optimization approaches avoid the use of rare codons. However, different approaches vary in the extent of other features considered. Features include mRNA elements that can inhibit expression, for example mRNA instability elements, nucleotide context of the initiation codon, mRNA secondary structures, sequence repeats, nucleotide composition, internal ribosome entry sites, promoter sequences, and putative splice donor and acceptor sites [33, 37, 38, 90]. In addition, some programs consider protein structural information, intragenic poly(A) sites, stop codons in alternative reading frames, and dinucleotides that are targets for RNase cleavage, mutation, and methylation-dependent gene silencing [38, 90, 91]. Moreover, some approaches have features that facilitate cloning, for example by adding or removing restriction sites [31, 34, 92]. Some approaches also allow |

|                                                 |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                                                 | oligonucleotides to be designed and optimized for gene synthesis using different strategies [31, 32, 34, 42, 91, 93]. In some cases, gene synthesis is the primary consideration and the ability to alter codon usage provides flexibility for good oligonucleotide design [94, 95]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                        |
| <b>Codon optimization applications</b>          | include optimizing mRNAs for expression in different organisms by using organism-specific codon usage frequencies, e.g. [36, 37, 93, 96], designing RNAi resistant genes, e.g. for gene rescue experiments [96], and embedding genetic watermarks into genes [97]. Specific codon-optimization approaches have also been developed for DNA vaccine and gene therapy applications [98, 99]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
| <b>Codon usage</b>                              | refers to the non-random use of codons in mRNAs. Codon usage in many organisms has been quantified using various calculations, including the frequency of use of optimal codons [28], the codon bias index [29], relative synonymous codon usage [30], the codon adaptation index [100], and the effective number of codons [101]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| <b>Genetic code</b>                             | refers to the nucleotide triplets termed codons that specify specific amino acids. Codons comprise the coding sequences of genes and are recognized at the mRNA level during the process of translation. The genetic code consists of 64 trinucleotide codons: 61 triplets specify 20 amino acids and 3 serve as stop codons. Only two amino acids, methionine (Met, M) and tryptophan (Trp, W) are encoded by single codons; other amino acids are encoded by 2,3, 4, or 6 synonymous codons                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| <b>Major Histocompatibility Complex I (MHC)</b> | in vertebrates, the MHC binds peptides that are typically derived from endogenous proteins and presents them on the cell surface, where they serve as ligands for antigen receptors of cytotoxic (CD8+) T cells. These MHC/peptide complexes allow the immune system to distinguish normal, healthy cells from those harboring pathogenic infection, or having undergone tumorigenic transformation. Peptides were originally believed to originate from the proteolytic cleavage of mature, functional proteins [102]. It was subsequently proposed that these peptides may also result from defective ribosomal products, prematurely terminated proteins, and misfolded polypeptides [103–105]. However, it is now evident that many peptides are derived from newly synthesized proteins [106]. In addition, there is evidence that many of these peptides are translated during the pioneer round of mRNA translation [107, 108]. Moreover, some peptides are encoded in alternative reading frames [74]. Interestingly, many of the epitopes that originate from alternative reading frames initiate |

### **Nucleic acid therapies**

translation at non-AUG codons. In the best studied case, a CUG start codon was shown to initiate translation using an elongator leucine tRNA (tRNA<sup>Leu</sup>CAG) as opposed to the initiator methionine-tRNA (tRNA<sup>Met</sup>CAU) [109]

refers to approaches that use DNA or RNA to mediate a therapeutic effect. Nucleic acid therapies include gene therapy, mRNA therapy, DNA vaccines, and RNA vaccines

### **Ribosomal tethering and clustering**

a hypothesis which involves direct binding of the initiator Met-tRNA to an accessible initiation codon. This binding occurs while the initiator Met-tRNA is associated with a ribosomal subunit that is either bound to the mRNA (tethered) or localized by more transient interactions (clustered). According to this model, alternative initiation is an inevitable consequence of initiation

### **Ribosome profiling**

a technique whereby the positions of translating ribosomes (ribosome footprints) can be mapped onto mRNAs at single nucleotide resolution. This technique has also been used to map ribosomal complexes at translation initiation sites by first treating cells with a drug (Harringtonine or Lactinomycin) that blocks translation initiation and freezes initiation complexes at the start site [110]

### **RNA editing**

a process that describes various post-transcriptional modifications that alter specific nucleotides in RNA molecules. For many organisms these modifications increase transcriptome complexity and contribute to a higher level of protein diversity than is indicated by the number of genes and alternative splicing variants. A-to-I editing involves the selective deamination to inosine of particular adenosines that are contained within imperfect, double-stranded regions of RNA. This process is catalyzed by Adenosine Deaminases Acting on RNA (ADAR) (reviewed in [111]). A-to-I editing occurs predominantly in tissues derived from the nervous system [112]; this editing occurs most frequently within non coding sequences and particularly in the RNA structures formed by inverted Alu repeats (reviewed in [113]). However A-to-I editing also occurs in other types of repetitive sequences, including microRNAs, and tRNAs [85, 111]. A-to-I editing within intronic sequences of pre-mRNAs may serve to modulate alternatively spliced variants by altering splice acceptor or donor sites, or by introducing new splice sites. Recent findings in *Drosophila* demonstrate that the extent of editing is determined cotranscriptionally [114] and supports a close relationship between splicing and editing [115]

|                            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Scanning hypothesis</b> | a hypothesis proposed by Marilyn Kozak which suggests that during translation initiation, the small ribosomal subunit scans from the mRNA cap-structure found at the 5' ends of mRNA to the initiation codon [116]. This model proposes 5' to 3' linear ribosomal movement in which the 5' leader is inspected nucleotide by nucleotide until the initiation codon is identified. At this point, scanning stops, the large ribosomal subunit attaches, and peptide synthesis begins. Kozak later extended and modified the model to include leaky scanning and reinitiation to accommodate examples of translation initiation that are inconsistent with the original model. In addition, the model has recently been modified by others to allow scanning from an internal mRNA recruitment site, scanning across the base of some stem-loop structures, and bidirectional movement resulting from backward scanning or diffusion |
| <b>Superwobble</b>         | this type of wobble pairing, also referred to as 4-way wobbling or hyperwobbling, occurs between an unmodified uridine in the tRNA at position 34 and the third nucleotide of the codon                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>Synonymous codon</b>    | refers to groups of codons that encode the same amino acid. A mutation that changes a codon to a synonymous codon is termed a silent mutation because the amino acid sequence is unaltered. However, the term silent mutation may be a misnomer as there are numerous diseases associated with synonymous codon changes                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
| <b>tRNA channeling</b>     | suggests that the translation machinery is organized during protein synthesis to facilitate charging, use, and recharging of tRNAs without their diffusion into the cytosol. This model is based on a series of studies from Deutscher in the 1990's [117, 118], and is consistent with various recent experimental observations [52, 119–121]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| <b>Wobble</b>              | Crick's wobble hypothesis [122] suggests that standard base pairing is used for the first 2 nucleotides of a codon, but that the stringency of base pairing is relaxed in the third position                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |

## References

1. Welch M, et al. You're one in a googol: optimizing genes for protein expression. *Journal of the Royal Society, Interface / the Royal Society*. 2009; 6(Suppl 4):S467–476.
2. Kudla G, et al. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science*. 2009; 324:255–258. [PubMed: 19359587]
3. Ward NJ, et al. Codon optimization of human factor VIII cDNAs leads to high-level expression. *Blood*. 2011; 117:798–807. [PubMed: 21041718]
4. Shabalina SA, et al. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic acids research*. 2013; 41:2073–2094. [PubMed: 23293005]

5. Tsai CJ, et al. Synonymous mutations and ribosome stalling can lead to altered folding pathways and distinct minima. *J Mol Biol.* 2008; 383:281–291. [PubMed: 18722384]
6. Agashe D, et al. Good codons, bad transcript: large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. *Mol Biol Evol.* 2013; 30:549–560. [PubMed: 23223712]
7. Spencer PS, et al. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J Mol Biol.* 2012; 422:328–335. [PubMed: 22705285]
8. Zhou JH, et al. The effects of the synonymous codon usage and tRNA abundance on protein folding of the 3C protease of foot-and-mouth disease virus. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases.* 2013; 16:270–274.
9. Zhang F, et al. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science.* 2010; 329:1534–1537. [PubMed: 20847274]
10. Hunt R, et al. Silent (synonymous) SNPs: should we care about them? *Methods Mol Biol.* 2009; 578:23–39. [PubMed: 19768585]
11. Katsnelson A. Breaking the silence. *Nat Med.* 2011; 17:1536–1538. [PubMed: 22146444]
12. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. *Nature reviews Genetics.* 2011; 12:683–691.
13. Chen R, et al. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PloS one.* 2010; 5:e13574. [PubMed: 21042586]
14. Kimchi-Sarfaty C, et al. Building better drugs: developing and regulating engineered therapeutic proteins. *Trends Pharmacol Sci.* 2013; 34:534–548. [PubMed: 24060103]
15. U. S. Food and Drug Administration. Paving the way for personalized medicine: FDA's role in a new era of medical product development. 2013. Retrieved from <http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/PersonalizedMedicine/UCM372421>
16. Dittmar KA, et al. Tissue-specific differences in human transfer RNA expression. *PLoS genetics.* 2006; 2:e221. [PubMed: 17194224]
17. Goldman E. tRNA and the Human Genome. *Proc Natl Acad Sci U S A.* 2011; 108:16980–16985. [PubMed: 21969541]
18. Murphy, FVt; Ramakrishnan, V. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nature structural & molecular biology.* 2004; 11:1251–1252.
19. Su AA, Randau L. A-to-I and C-to-U editing within transfer RNAs. *Biochemistry (Mosc).* 2011; 76:932–937. [PubMed: 22022967]
20. Agris PF. Decoding the genome: a modified view. *Nucleic acids research.* 2004; 32:223–238. [PubMed: 14715921]
21. Rogalski M, et al. Superwobbling facilitates translation with reduced tRNA sets. *Nature structural & molecular biology.* 2008; 15:192–198.
22. Alkatib S, et al. The contributions of wobbling and superwobbling to the reading of the genetic code. *PLoS genetics.* 2012; 8:e1003076. [PubMed: 23166520]
23. Itakura K, et al. Expression in *Escherichia coli* of a chemically synthesized gene for the hormone somatostatin. *Science.* 1977; 198:1056–1063. [PubMed: 412251]
24. Air GM, et al. Gene F of bacteriophage phiX174. Correlation of nucleotide sequences from the DNA and amino acid sequences from the gene product. *Journal of molecular biology.* 1976; 107:445–458. [PubMed: 1088826]
25. Efstratiadis A, et al. The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. *Cell.* 1977; 10:571–585. [PubMed: 558827]
26. Fiers W, et al. A-protein gene of bacteriophage MS2. *Nature.* 1975; 256:273–278. [PubMed: 806810]
27. Post LE, et al. Nucleotide sequence of the ribosomal protein gene cluster adjacent to the gene for RNA polymerase subunit beta in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America.* 1979; 76:1697–1701. [PubMed: 377281]
28. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon

- choice that is optimal for the *E. coli* translational system. *Journal of molecular biology*. 1981; 151:389–409. [PubMed: 6175758]
29. Bennetzen JL, Hall BD. Codon selection in yeast. *Journal of Biological Chemistry*. 1982; 257:3026–3031. [PubMed: 7037777]
30. Sharp PM, et al. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic acids research*. 1986; 14:5125–5143. [PubMed: 3526280]
31. Richardson SM, et al. GeneDesign: rapid, automated design of multikilobase synthetic genes. *Genome research*. 2006; 16:550–556. [PubMed: 16481661]
32. Villalobos A, et al. Gene Designer: a synthetic biology tool for constructing artificial DNA segments. *BMC bioinformatics*. 2006; 7:285. [PubMed: 16756672]
33. Gao W, et al. UpGene: Application of a web-based DNA codon optimization algorithm. *Biotechnology progress*. 2004; 20:443–448. [PubMed: 15058988]
34. Jayaraj S, et al. GeMS: an advanced software package for designing synthetic genes. *Nucleic acids research*. 2005; 33:3011–3016. [PubMed: 15911632]
35. Wu G, et al. The Synthetic Gene Designer: a flexible web platform to explore sequence manipulation for heterologous expression. *Protein Expr Purif*. 2006; 47:441–445. [PubMed: 16376569]
36. Bode M, et al. TmPrime: fast, flexible oligonucleotide design software for gene synthesis. *Nucleic acids research*. 2009; 37:W214–221. [PubMed: 19515937]
37. Raab D, et al. The GeneOptimizer Algorithm: using a sliding window approach to cope with the vast sequence space in multiparameter DNA sequence optimization. *Systems and synthetic biology*. 2010; 4:215–225. [PubMed: 21189842]
38. Gaspar P, et al. EuGene: maximizing synthetic gene design for heterologous expression. *Bioinformatics*. 2012; 28:2683–2684. [PubMed: 22847936]
39. Angov E, et al. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PloS one*. 2008; 3:e2189. [PubMed: 18478103]
40. Fuglsang A. Codon optimizer: a freeware tool for codon optimization. *Protein Expr Purif*. 2003; 31:247–249. [PubMed: 14550643]
41. Qian W, et al. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS genetics*. 2012; 8:e1002603. [PubMed: 22479199]
42. Hatfield GW, Roth DA. Optimizing scaleup yield for protein production: Computationally Optimized DNA Assembly (CODA) and Translation Engineering. *Biotechnology annual review*. 2007; 13:27–42.
43. Gustafsson C, et al. Engineering genes for predictable protein expression. *Protein expression and purification*. 2012; 83:37–46. [PubMed: 22425659]
44. Curran JF, Yarus M. Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *Journal of molecular biology*. 1989; 209:65–77. [PubMed: 2478714]
45. Kurland C, Gallant J. Errors of heterologous protein expression. *Current opinion in biotechnology*. 1996; 7:489–493. [PubMed: 8939626]
46. Bonekamp F, et al. Translation rates of individual codons are not correlated with tRNA abundances or with frequencies of utilization in *Escherichia coli*. *Journal of Bacteriology*. 1989; 171:5812–5816. [PubMed: 2509420]
47. Wu X, et al. Codon optimization reveals critical factors for high level expression of two rare codon genes in *Escherichia coli*: RNA stability and secondary structure but not tRNA abundance. *Biochemical and biophysical research communications*. 2004; 313:89–96. [PubMed: 14672702]
48. Rosano GL, Ceccarelli EA. Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted *Escherichia coli* strain. *Microbial cell factories*. 2009; 8:41. [PubMed: 19630980]
49. Hershey JW, et al. Principles of translational control: an overview. *Cold Spring Harbor perspectives in biology*. 2012; 4:a011528. [PubMed: 23209153]
50. Komar AA. A pause for thought along the co-translational folding pathway. *Trends in biochemical sciences*. 2009; 34:16–24. [PubMed: 18996013]

51. Rosenblum G, et al. Quantifying Elongation Rhythm During Full-Length Protein Synthesis. *Mol Cell*. 2013; 135:11322–11329.
52. Cannarozzi G, et al. A role for codon order in translation dynamics. *Cell*. 2010; 141:355–367. [PubMed: 20403329]
53. Stadler M, Fire A. Wobble base-pairing slows in vivo translation elongation in metazoans. *Rna*. 2011; 17:2063–2073. [PubMed: 22045228]
54. Li GW, et al. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature*. 2012; 484:538–541. [PubMed: 22456704]
55. Mauro VP, Edelman GM. The ribosome filter redux. *Cell Cycle*. 2007; 6:2246–2251. [PubMed: 17890902]
56. Dresios J, et al. An mRNA-rRNA base-pairing mechanism for translation initiation in eukaryotes. *Nat Struct Mol Biol*. 2006; 13:30–34. [PubMed: 16341227]
57. Chappell SA, et al. Ribosomal shunting mediated by a translational enhancer element that base pairs to 18S rRNA. *Proc Natl Acad Sci USA*. 2006; 103:9488–9493. [PubMed: 16769881]
58. Larsen B, et al. rRNA-mRNA base pairing stimulates a programmed-1 ribosomal frameshift. *Journal of Bacteriology*. 1994; 176:6842–6851. [PubMed: 7961443]
59. Luttermann C, Meyers G. The importance of inter-and intramolecular base pairing for translation reinitiation on a eukaryotic bicistronic mRNA. *Genes & development*. 2009; 23:331–344. [PubMed: 19204118]
60. Hausser J, et al. Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res*. 2013; 23:604–615. [PubMed: 23335364]
61. Hu S, et al. Genetic Code-Guided Protein Synthesis and Folding in *E. coli*. *Journal of Biological Chemistry*. 2013; 288:30855–30861. [PubMed: 24003234]
62. Sander IM, et al. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *Journal of the American Chemical Society*. 2014; 136:858–861. [PubMed: 24392935]
63. Hoekema A, et al. Codon replacement in the PGK1 gene of *Saccharomyces cerevisiae*: experimental approach to study the role of biased codon usage in gene expression. *Molecular and cellular biology*. 1987; 7:2914–2924. [PubMed: 2823108]
64. Kotula L, Curtis PJ. Evaluation of foreign gene codon optimization in yeast: expression of a mouse IG kappa chain. *Bio/technology (Nature Publishing Company)*. 1991; 9:1386–1389.
65. Gustafsson C, et al. Codon bias and heterologous protein expression. *Trends Biotechnol*. 2004; 22:346–353. [PubMed: 15245907]
66. Charneski CA, Hurst LD. Positively charged residues are the major determinants of ribosomal velocity. *PLoS biology*. 2013; 11:e1001508. [PubMed: 23554576]
67. Chen C, et al. Dynamics of translation by single ribosomes through mRNA secondary structures. *Nature structural & molecular biology*. 2013; 20:582–588.
68. Dana A, Tuller T. Determinants of translation elongation speed and ribosomal profiling biases in mouse embryonic stem cells. *PLoS computational biology*. 2012; 8:e1002755. [PubMed: 23133360]
69. Subramanian S. Nearly neutrality and the evolution of codon usage bias in eukaryotic genomes. *Genetics*. 2008; 178:2429–2432. [PubMed: 18430960]
70. Andersson SG, Kurland CG. Codon preferences in free-living microorganisms. *Microbiological reviews*. 1990; 54:198–210. [PubMed: 2194095]
71. Klumpp S, et al. On ribosome load, codon bias and protein abundance. *PloS one*. 2012; 7:e48542. [PubMed: 23144899]
72. Larsen HB. Kenyan dominance in distance running. *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology*. 2003; 136:161–170.
73. Matsuda D, Mauro VP. Determinants of initiation codon selection during translation in mammalian cells. *PloS one*. 2010; 5:e15057. [PubMed: 21124832]
74. Malarkannan S, et al. Presentation of out-of-frame peptide/MHC class I complexes by a novel translation initiation mechanism. *Immunity*. 1999; 10:681–690. [PubMed: 10403643]



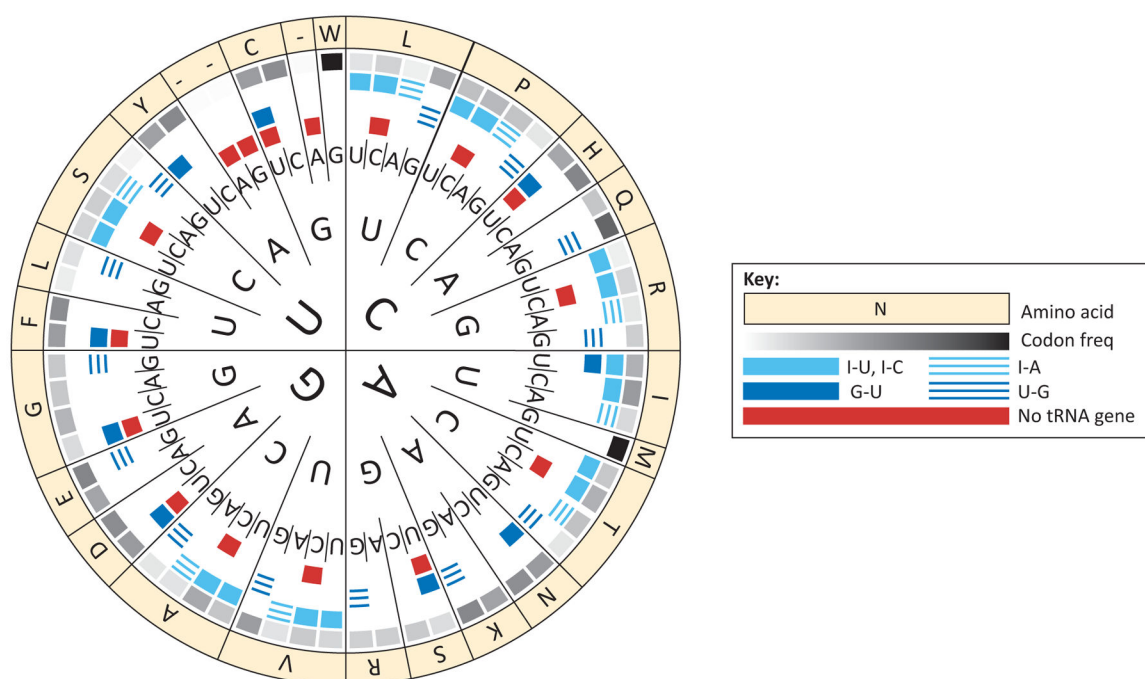
75. Ingolia NT, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*. 2009; 324:218–223. [PubMed: 19213877]
76. Ingolia NT, et al. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell*. 2011; 147:789–802. [PubMed: 22056041]
77. Lee S, et al. Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc Natl Acad Sci U S A*. 2012; 109:E2424–2432. [PubMed: 22927429]
78. Menschaert G, et al. Deep proteome coverage based on ribosome profiling aids MS-based protein and peptide discovery and provides evidence of alternative translation products and near-cognate translation initiation events. *Molecular & cellular proteomics*. 2013; 12:1780–1790. [PubMed: 23429522]
79. Slavoff SA, et al. Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nature chemical biology*. 2013; 9:59–64.
80. Kozak M. Initiation of translation in prokaryotes and eukaryotes. *Gene*. 1999; 234:187–208. [PubMed: 10395892]
81. Chappell SA, et al. Ribosomal tethering and clustering as mechanisms for translation initiation. *Proc Natl Acad Sci USA*. 2006; 103:18077–18082. [PubMed: 17110442]
82. Irimia M, et al. Evolutionarily conserved A-to-I editing increases protein stability of the alternative splicing factor Nova1. *RNA biology*. 2012; 9:12–21. [PubMed: 22258141]
83. Godfried Sie C, et al. IGFBP7's susceptibility to proteolysis is altered by A-to-I RNA editing of its transcript. *FEBS Lett*. 2012; 586:2313–2317. [PubMed: 22750143]
84. Ramaswami G, et al. Identifying RNA editing sites using RNA sequencing data alone. *Nature methods*. 2013; 10:128–132. [PubMed: 23291724]
85. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. *Annual review of biochemistry*. 2010; 79:321–349.
86. Hideyama T, et al. Profound downregulation of the RNA editing enzyme ADAR2 in ALS spinal motor neurons. *Neurobiology of Disease*. 2012; 45:1121–1128. [PubMed: 22226999]
87. Silberberg G, et al. Deregulation of the A-to-I RNA editing mechanism in psychiatric disorders. *Human Molecular Genetics*. 2012; 21:311–321. [PubMed: 21984433]
88. Choudhury Y, et al. Attenuated adenosine-to-inosine editing of microRNA-376a\* promotes invasiveness of glioblastoma cells. *J Clin Invest*. 2012; 122:4059–4076. [PubMed: 23093778]
89. Enstero M, et al. A computational screen for site selective A-to-I editing detects novel sites in neuron specific Hu proteins. *BMC bioinformatics*. 2010; 11:6. [PubMed: 20047656]
90. Fath S, et al. Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. *PloS one*. 2011; 6:e17596. [PubMed: 21408612]
91. Lorimer D, et al. Gene composer: database software for protein construct design, codon engineering, and gene synthesis. *BMC biotechnology*. 2009; 9:36. [PubMed: 19383142]
92. Raghava GP, Sahni G. GMAP: a multi-purpose computer program to aid synthetic gene design, cassette mutagenesis and the introduction of potential restriction sites into DNA sequences. *BioTechniques*. 1994; 16:1116–1123. [PubMed: 8074878]
93. Hoover DM, Lubkowski J. DNAWorks: an automated method for designing oligonucleotides for PCR-based gene synthesis. *Nucleic acids research*. 2002; 30:e43. [PubMed: 12000848]
94. Huang G, et al. An efficient and rapid method for cDNA cloning from difficult templates using codon optimization and SOE-PCR: with human RANK and TIMP2 gene as examples. *Biotechnology letters*. 2011; 33:1939–1947. [PubMed: 21660576]
95. Li MH, et al. De novo gene synthesis design using TmPrime software. *Methods Mol Biol*. 2012; 852:225–234. [PubMed: 22328437]
96. Kumar D, et al. Validation of RNAi silencing specificity using synthetic genes: salicylic acid-binding protein 2 is required for innate immunity in plants. *Plant Journal*. 2006; 45:863–868. [PubMed: 16460518]
97. Liss M, et al. Embedding permanent watermarks in synthetic genes. *PloS one*. 2012; 7:e42465. [PubMed: 22905136]

98. Satya RV, et al. A pattern matching algorithm for codon optimization and CpG motif-engineering in DNA expression vectors. *Proceedings / IEEE Computer Society Bioinformatics Conference. IEEE Computer Society Bioinformatics Conference.* 2003; 2:294–305. [PubMed: 16452805]
99. Harish N, et al. DyNAVacS: an integrative tool for optimized DNA vaccine design. *Nucleic acids research.* 2006; 34:W264–266. [PubMed: 16845007]
100. Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic acids research.* 1987; 15:1281–1295. [PubMed: 3547335]
101. Wright F. The 'effective number of codons' used in a gene. *Gene.* 1990; 87:23–29. [PubMed: 2110097]
102. Townsend A, et al. Defective presentation to class I-restricted cytotoxic T lymphocytes in vaccinia-infected cells is overcome by enhanced degradation of antigen. *The Journal of experimental medicine.* 1988; 168:1211–1224. [PubMed: 2459295]
103. Yewdell JW, et al. Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? *J Immunol.* 1996; 157:1823–1826. [PubMed: 8757297]
104. Yewdell JW, Nicchitta CV. The DRiP hypothesis decennial: support, controversy, refinement and extension. *Trends in immunology.* 2006; 27:368–373. [PubMed: 16815756]
105. Cardinaud S, et al. The synthesis of truncated polypeptides for immune surveillance and viral evasion. *PloS one.* 2010; 5:e8692. [PubMed: 20098683]
106. Reits EA, et al. The major substrates for TAP in vivo are derived from newly synthesized proteins. *Nature.* 2000; 404:774–778. [PubMed: 10783892]
107. Apcher S, et al. Major source of antigenic peptides for the MHC class I pathway is produced during the pioneer round of mRNA translation. *Proceedings of the National Academy of Sciences of the United States of America.* 2011; 108:11572–11577. [PubMed: 21709220]
108. Apcher S, et al. Translation of pre-spliced RNAs in the nuclear compartment generates peptides for the MHC class I pathway. *Proc Natl Acad Sci U S A.* 2013; 110:17951–17956. [PubMed: 24082107]
109. Starck SR, et al. Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I. *Science.* 2012; 336:1719–1723. [PubMed: 22745432]
110. Ingolia NT, et al. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nature protocols.* 2012; 7:1534–1550.
111. Mallela A, Nishikura K. A-to-I editing of protein coding and noncoding RNAs. *Critical reviews in biochemistry and molecular biology.* 2012; 47:493–501. [PubMed: 22988838]
112. Paul MS, Bass BL. Inosine exists in mRNA at tissue-specific levels and is most abundant in brain mRNA. *EMBO Journal.* 1998; 17:1120–1127. [PubMed: 9463389]
113. Penn AC, et al. Reciprocal regulation of A-to-I RNA editing and the vertebrate nervous system. *Frontiers in neuroscience.* 2013; 7:61. [PubMed: 23616744]
114. Rodriguez J, et al. Nascent-seq indicates widespread cotranscriptional RNA editing in *Drosophila*. *Mol Cell.* 2012; 47:27–37. [PubMed: 22658416]
115. Rieder LE, Reenan RA. The intricate relationship between RNA structure, editing, and splicing. *Seminars in Cell and Developmental Biology.* 2012; 23:281–288. [PubMed: 22178616]
116. Kozak M. How do eucaryotic ribosomes select initiation regions in messenger RNA? *Cell.* 1978; 15:1109–1123. [PubMed: 215319]
117. Negrutskii BS, et al. Supramolecular organization of the mammalian translation system. *Proc Natl Acad Sci USA.* 1994; 91:964–968. [PubMed: 8302874]
118. Stapulionis R, et al. Efficient mammalian protein synthesis requires an intact F-actin system. *J Biol Chem.* 1997; 272:24980–24986. [PubMed: 9312103]
119. Barhoom S, et al. Quantitative single cell monitoring of protein synthesis at subcellular resolution using fluorescently labeled tRNA. *Nucleic acids research.* 2011; 39:e129. [PubMed: 21795382]
120. Pavon-Eternod M, et al. Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. *Nucleic acids research.* 2013; 41:1914–1921. [PubMed: 23254333]
121. Hirschmann WD, et al. Scp160p is required for translational efficiency of codon-optimized mRNAs in yeast. *Nucleic acids research.* 2014; 42:4043–4055. [PubMed: 24445806]

122. Crick FH. Codon-anticodon pairing: the wobble hypothesis. *Journal of molecular biology*. 1966; 19:548–555. [PubMed: 5969078]
123. Agris PF, et al. tRNA's wobble decoding of the genome: 40 years of modification. *J Mol Biol*. 2007; 366:1–13. [PubMed: 17187822]
124. Sprinzl M, et al. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*. 1998; 26:148–153. [PubMed: 9399820]
125. Peabody DS. Translation initiation at non-AUG triplets in mammalian cells. *J Biol Chem*. 1989; 264:5031–5035. [PubMed: 2538469]

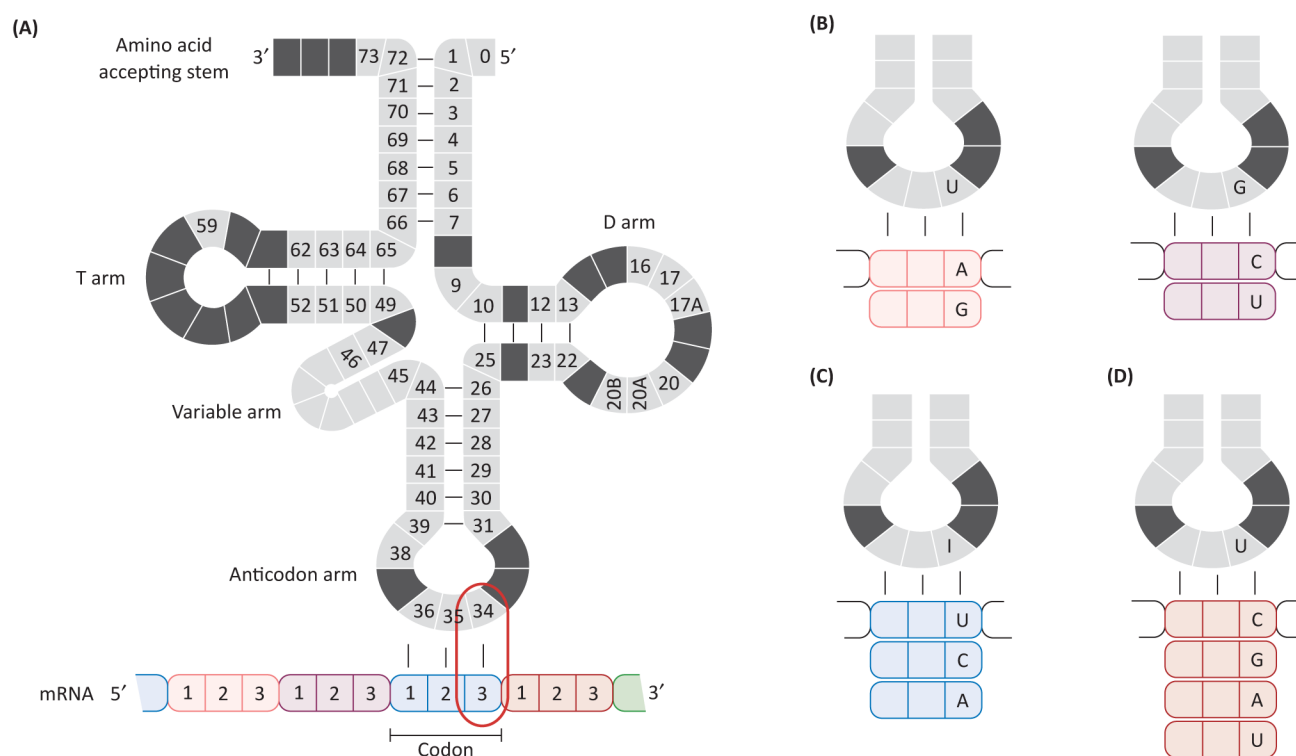
### Highlights

- Codon optimization is a gene engineering approach to increase protein production.
- Assumptions underlying codon optimization approaches may be invalid.
- Codon optimization for nucleic acid therapies presents potentially unique hazards.
- Hazards can arise by disrupting or introducing overlapping functions in mRNAs.



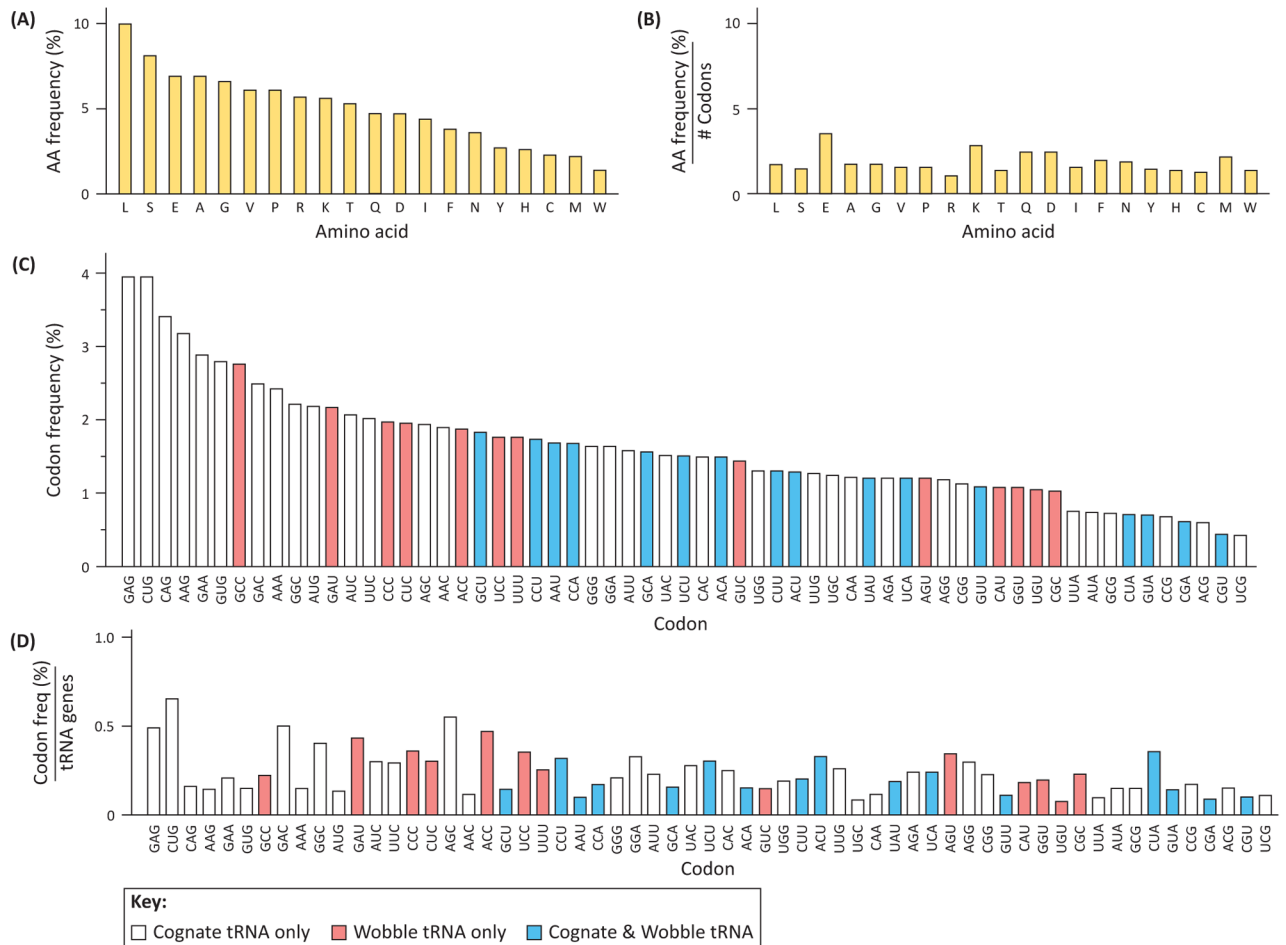
**Figure 1. The degenerate genetic code**

The circular representation of the genetic code indicates wobble, tRNA gene presence, and codon usage in humans. Amino acids are indicated in the outside yellow ring, using the one letter notation. Stop codons are indicated by a dash. The 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> nucleotides of codons are indicated in the inner, middle, and outer nucleotide circles, respectively. Codons that lack a corresponding tRNA gene in humans are indicated by red bars [17]. For illustration, potential wobble codons are indicated, based on Crick's wobble rules. G-U wobble base pairing codons are indicated by dark blue bars. Potential U-G wobble base pairing codons are indicated by dark blue striped bars. Codons capable of I-U and I-C base pairing are indicated by light blue bars. Potential I-A wobble base pairing codons are indicated by light blue striped bars. Note that 7 possible inosine modifications have been reported for yeast and 8 for mammalian tRNAs [18, 19]. Codons that recognize tRNAs with other modifications that may extend or restrict wobble are not indicated [20, 123]. The frequency of codon usage in human is indicated by the grey bars. The density of the bars corresponds to codon usage ranging from 0% usage (white) to 100% usage (black). Human codon usage data is from the "Codon Usage Database" (<http://www.kazusa.or.jp/codon/>).



**Figure 2. tRNA-mRNA base pairing and wobble base pairing**

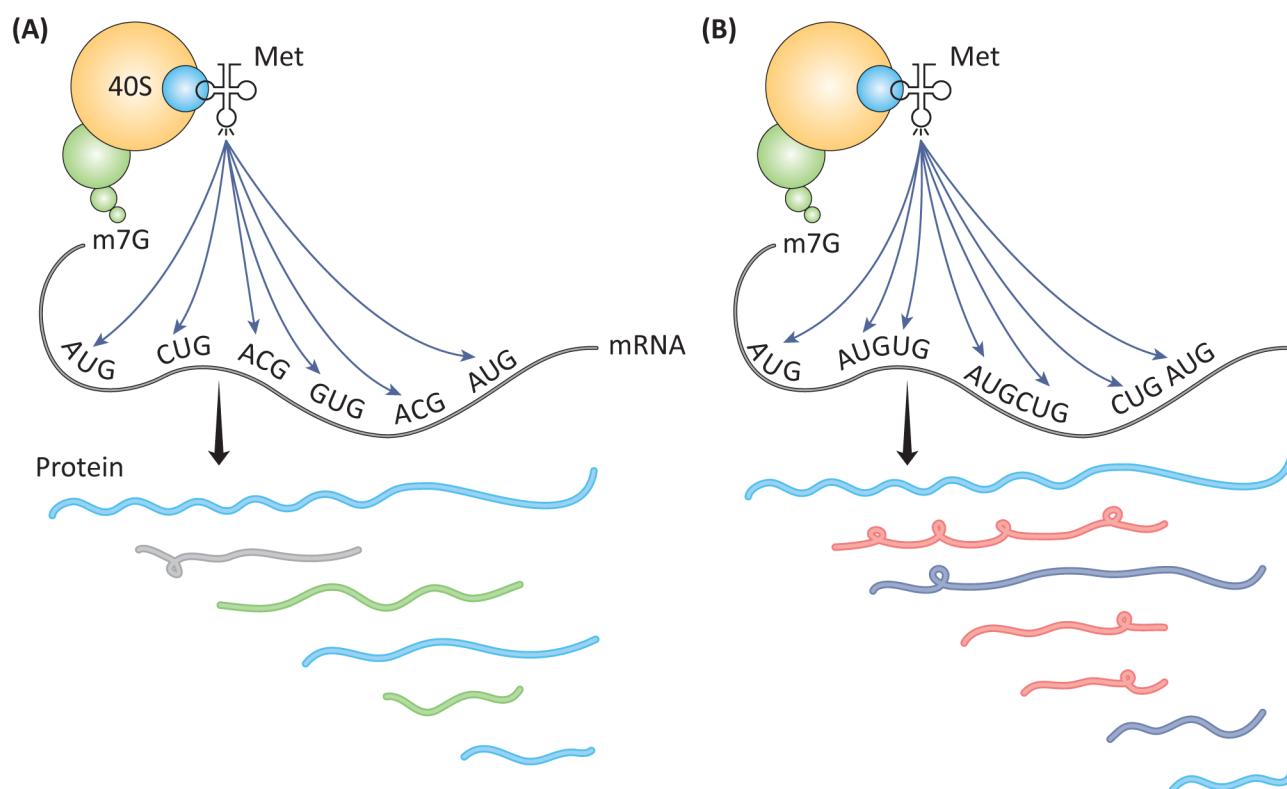
**(A)** Nucleotide positions. tRNA numbering is according to [124]. Structural features in the tRNA are indicated; the anticodon occupies nucleotide positions 34–36. Conserved nucleotide positions are highlighted in black. Base pairing of a tRNA to a codon in mRNA is indicated, with the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> positions of the codons labeled 1, 2, 3. Different colors represent different codons. Wobble base pairing occurs between position 34 of the tRNA and the 3<sup>rd</sup> position of the codon, indicated in red. **(B)** A tRNA with a U or G at position 34 can Watson-Crick base pair and wobble base pair as shown. Only the anti-codon loop of the tRNAs is shown in panels B-D. **(C)** A tRNA with I at position 34 can wobble base pair to U, C, or A as shown. **(D)** Superwobbling. A tRNA with an unmodified U in tRNA position 34 can base pair to C and wobble base pair to G, A, and U. To date, superwobbling has only been reported in chloroplasts [21, 22].



**Figure 3. Codon frequencies in human**

**(A)** The frequency of occurrence of amino acids (AA). AAs are listed on the abscissa in one letter notation. The frequency with which each AA is encoded is listed as a percentage on the ordinate. **(B)** Normalized frequency of occurrence of AAs. For each amino acid, the amino acid frequency (%) is normalized to the number of synonymous codons. **(C)** The frequency of occurrence of codons. Codons are listed on the abscissa of the bar graph; in each case, the frequency of usage (as a percentage) is plotted on the ordinate. The codons are ordered according to their reported frequency of occurrence. Stop codons are not listed. White bars indicate codons that can be decoded only by corresponding cognate tRNAs. Black bars indicate codons that lack a cognate tRNA gene and can be decoded only by wobble tRNAs. Grey bars indicate codons that can be decoded by both cognate and wobble tRNAs. Note that potential U-G and I-A wobble interactions are not considered in this bar graph as these wobble interactions do not yet appear to have been confirmed in human. **(D)** Normalized codon frequency. The codon frequency (%) has been normalized by dividing the codon frequency by the average number of cognate and wobble tRNA genes for each codon. The human codon usage data is from “Codon Usage Database” (<http://www.kazusa.or.jp/codon/>).





**Figure 4. Translation of full-length protein and cryptic peptides**

**(A)** Translation from a natural cap-dependent mRNA. Schematic represents an mRNA that initiates translation from multiple start sites, including AUG and noncanonical start sites. A ribosomal complex assembled at the 5' m7G cap-structure of an mRNA is indicated. The 40S subunit is tethered to the mRNA via the eIF4F complex of initiation factors and initiation factor eIF3, indicated as green circles [81]. The 40S subunit is bound to the ternary complex, which contains initiation factor eIF2, indicated in blue, GTP, and the initiator Met-tRNA, indicated as a cloverleaf structure. The arrows indicate possible start sites for translation initiation. Start sites include AUG and non-canonical start sites such as CUG, ACG, and GUG [125]. Protein synthesis starting at the initiation codon gives rise to full-length protein, which is represented by the long blue bar. Shorter peptides initiating from alternative start sites in the same reading frame are indicated by the shorter blue bars. Peptides initiating from out-of-frame start sites will generate peptides with different amino acid sequences; these are indicated by the grey and green bars, which represent the two alternative (out-of-frame) reading frames. **(B)** Translation from a codon-optimized mRNA. In this case, the full-length protein and any in-frame peptides will be the same as those from the natural mRNA. Some in-frame peptides may be lost if a non-AUG codon is modified (e.g. the peptide arising from the internal GUG). In addition, most if not all of the out-of-frame peptides will be lost and a new set of out-of-frame peptides will be encoded. These new peptides from the two new alternative reading frames are indicated in fuchsia and dark orange.