The Ongoing Quest to Crack the Genetic Code for Protein Production

Molecular Cell

Nieuwkoop, Thijs; Finger-Bou, Max; Oost, John; Claassens, Nico J.

https://doi.org/10.1016/j.molcel.2020.09.014

Molecular Cell

Review

The Ongoing Quest to Crack the Genetic Code for Protein Production

Thijs Nieuwkoop,[1] Max Finger-Bou,[1] John van der Oost,[1] and Nico J. Claassens[1],*
[1]Laboratory of Microbiology, Wageningen University, Stippeneng 4, 6708 WE Wageningen, the Netherlands
*Correspondence: nico.claassens@wur.nl
https://doi.org/10.1016/j.molcel.2020.09.014

SUMMARY

Understanding the genetic design principles that determine protein production remains a major challenge. Although the key principles of gene expression were discovered 50 years ago, additional factors are still being uncovered. Both protein-coding and non-coding sequences harbor elements that collectively influence the efficiency of protein production by modulating transcription, mRNA decay, and translation. The influences of many contributing elements are intertwined, which complicates a full understanding of the individual factors. In natural genes, a functional balance between these factors has been obtained in the course of evolution, whereas for genetic-engineering projects, our incomplete understanding still limits optimal design of synthetic genes. However, notable advances have recently been made, supported by high-throughput analysis of synthetic gene libraries as well as by state-of-the-art biomolecular techniques. We discuss here how these advances further strengthen understanding of the gene expression process and how they can be harnessed to optimize protein production.

INTRODUCTION

The biosynthesis of proteins is one of the core processes in living cells, as well as in many biotechnological applications. It has already been 50 years since Francis Crick proposed the central dogma of molecular biology (Crick, 1970), explaining how DNA is transcribed to mRNA, which is then translated to protein. A characteristic feature of the conversion of the information stored in the nucleotide building blocks of DNA and mRNA into the amino acid building blocks of proteins is the redundancy in the number of codons on the nucleotide level. Although there are 64 unique codons (nucleotide triplets), only 20 different amino acids make up proteins in most organisms. This redundancy gives astronomical numbers of codon combinations to encode the same amino acid sequence, e.g., the medium-size green fluorescent protein (GFP, 238 amino acids) can be encoded by ±3 × 10^{110} different open reading frames (ORFs).

However, different sequences encoding an identical protein sequence can lead to dramatic variations in protein production levels, and sometimes even lead to differences in protein folding and functionality (Buhr et al., 2016; Kim et al., 2015; Zhou et al., 2013) (Figure 1). Apart from ORFs, non-coding regions with potential regulatory functions, such as promoters and untranslated regions (UTRs; Figure 1), add a vast sequence space. As the design principles of both the coding and non-coding sequences are only partly known, the design of synthetic genes for expression is still a major challenge.

Already, since the early days of gene sequencing in the 1980s, a bias has been recognized in the codon usage of highly expressed native genes; particular synonymous codons (i.e., different codons encoding the same amino acid) were observed to be used more frequently than others. This notion led to the formulation of the Codon Adaption Index (CAI) (Sharp and Li, 1987), and it was postulated that the codon bias within highly expressed genes allowed for more-efficient translation. An underlying hypothesis to this observation is that the (amino-acid-charged) cognate tRNAs for these frequent codons are more abundant and that they are more-efficient decoders during ribosomal protein biosynthesis (Ikemura, 1985; dos Reis et al., 2003). In recent decades, the advent of high-throughput sequencing technologies has revealed more codon usage signatures varying across organisms, tissue types, and genes (Hanson and Coller, 2018).

Following these observations, several types of codon bias and mechanistic explanations were introduced (Quax et al., 2015). The current view on codon usage is that it is related to a complexity of factors. The weight of those factors varies depending on the context, which includes the type of organism, tissue, or compartment; physiological control (e.g., pathway or growth phase); or even the position within an ORF (Hanson and Coller, 2018; Quax et al., 2015). It became clear that the notion of frequent versus rare codons, similar to good versus bad codons for protein production, is an oversimplification of biological reality. Consequently, codon optimization algorithms, which are all based on simplified assumptions and codon indices (Bourret et al., 2019), cannot warrant successful heterologous protein production (Parret et al., 2016). Because codon choices are related to diverse mechanisms and regulatory processes, we prefer to use the term "codon optimality" only when a range of factors acting at different levels of the expression process have been taken into account.

A couple of years ago, we reviewed the effect of codon usage within the ORF on expression (Quax et al., 2015). Impressive advances have been made in the field since then, because, on one

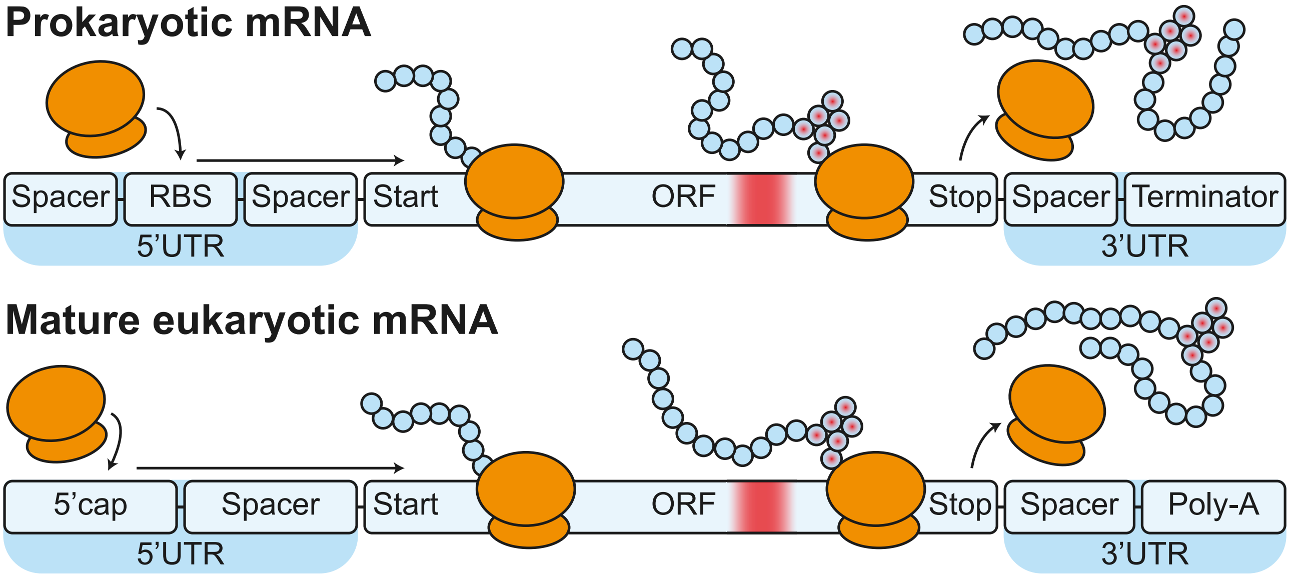


Figure 1. Schematic Overview of a Prokaryotic and Mature Eukaryotic mRNA Being Translated by Ribosomes (Orange)
RBS, ribosome binding site; ORF, open reading frame; 5′/3′ UTR, 5′/3′ untranslated region. The co-translational folding phenomenon is indicated with a red gradient in the mRNA and the associated amino acids.

hand, of the technical advances, including high-throughput analyses of large synthetic gene libraries (Cambray et al., 2018) and, on the other hand, because of innovative molecular biology approaches that unraveled additional details of transcription, translation, and protein folding (Buhr et al., 2016; Buschauer et al., 2020; Kim et al., 2015). These studies contributed to a further understanding of some of the factors involved and have also revealed relevant interactions among them.

Here, we provide a timely overview of the field of gene expression, discussing relevant features both in the regulation of noncoding regions and in ORFs. As transcription, mRNA decay, and translation (initiation and elongation) all have important roles in controlling protein production, we discuss all these stages (Figure 2). Furthermore, we highlight key controversies and knowledge gaps in the field and propose potential avenues to resolve these. Lastly, we discuss how our relatively poor understanding of optimal gene designs is a major limitation for biotechnology and synthetic biology. We examine how emerging tools and approaches can aid in overcoming challenges for engineering protein production.

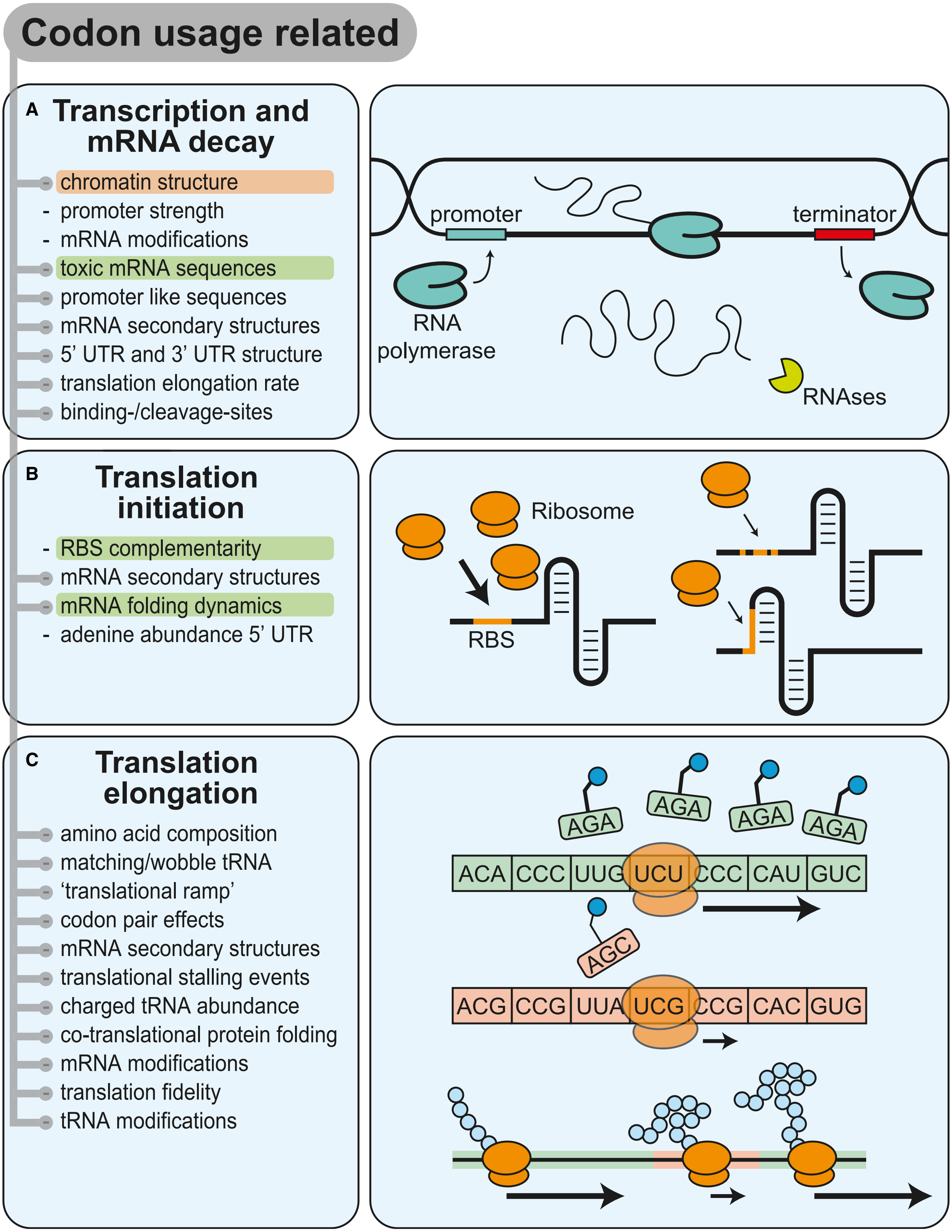
TRANSCRIPTION AND mRNA DECAY

Transcription Initiation

The first step in protein production is the transcription of DNA to mRNA by RNA polymerase (RNAP). Synthesis rates of mRNA are mediated by the binding affinity of RNAPs and related transcriptional factors with the promoter sequences; other factors, such as chromatin structures in eukaryotes, also have a role (Lenstra et al., 2016). In addition, the transition from transcription initiation to transcription elongation is important in determining mRNA synthesis rates. After the RNAP is bound, DNA is unwound, and an open complex is formed. During the open complex configuration, the first short RNA stretch is transcribed, and then, the RNAP either moves on to transcribe the full mRNA (promoter escape) or the initiation is aborted. Several promoter sequence features, for example, the length and nucleotides in the bacterial discriminator region (~4–7 bp upstream of the transcription start), determine the efficiency of the promoter escape (Henderson et al., 2017; Winkelman et al., 2016). Promoter sequence regions, as well as transcription initiation and elongation factors involved in promoter escape are reviewed in more detail elsewhere (Lee and Borukhov, 2016; Wade and Struhl, 2008). Although most of the key principles of transcription initiation and promoter escape are known, models to predict promoter strengths from sequences are still under development.

Recently, several groups investigated promoter properties and design constraints by expressing some reporter genes from libraries with randomized promoter sequences. Some studies in *Escherichia coli* reported that, of all fully randomized promoter sequences, 7%–10% resulted in detectable expression (Urtecho et al., 2020; Yona et al., 2018). Furthermore, it was found during laboratory evolution of random sequences in *E. coli* that 60% of those sequences became functional promoters with only one mutation (Yona et al., 2018). Functional promoters in *E. coli* were generally observed to have at least a canonical −10 or −35 motif for binding the RNAP-sigma subunit, which occurs relatively frequently in DNA sequences by chance. Another study randomized the yeast −90 to −170 promoter region, whereas the consensus TATA region was kept constant, which resulted in detectable expression for 83% of the sequences (de Boer et al., 2020).

The increasing data on characterized (random) promoters has also been used to create predictive models. Such *in silico* predictions have been successful for predicting promoter strengths of yeast, by modeling the transcription factor binding sites and their



(legend on next page)

accessibility (de Boer et al., 2020; Levo et al., 2017). However, the generation of predictive models for *E. coli* based on a set of fully randomized and native promoters by machine learning was still unsuccessful (Urtecho et al., 2020). This may be explained by the diverse sigma-factor-type promoters that are included in the training set. A previous study that performed machine learning and regression only on sigma-70 "household" promoters in *E. coli* did result in good predictive models (Urtecho et al., 2019).

The relatively high chance for random sequences to act as a promoter may also create "accidental promoters" in natural or synthetic sequences, which can cause transcriptional burdens and other distortions when they occur in undesired loci. Relatedly, an evolutionary selection against promoter-like sequences was observed within ORFs in *E. coli* (Yona et al., 2018). Promoters within ORFs may, however, also serve functional roles occasionally; it has been proposed that promoters in the reverse sequence of ORFs can produce antisense RNAs to downregulate protein production (Brophy and Voigt, 2016; Urtecho et al., 2020).

Apart from the influence of promoter regions on transcription, it was observed in some eukaryotes that the codon or nucleotide usage within an ORF might also affect transcription rates (Fu et al., 2018; Newman et al., 2016; Zhou et al., 2016). Proposed mechanisms through which nucleotide composition or codons could modulate transcriptional activity are related to histone modifications or the influence of GC-content on transcription elongation rates.

mRNA Decay

All cells harbor several endo- and exo-ribonucleases that are involved in degrading mRNA, providing additional control over mRNA levels and protein production (Schmid and Jensen, 2018). Furthermore, ribonucleases can clean up non-functional RNAs, e.g., from accidental transcription. The dynamics between mRNA transcription and mRNA decay result in a wide range of mRNA half-lives, serving as one of the key factors for protein production (Boël et al., 2016; Lahtvee et al., 2017; Presnyak et al., 2015).

One of the factors modulating mRNA stability is the presence of structural elements in their untranslated regions. Secondary structures and sequences of UTRs can influence mRNA decay rates, especially in bacteria (Mohanty and Kushner, 2016). Recently an increasing number of studies demonstrated the important role of the 3′ UTR region in controlling mRNA decay (Menendez-Gil et al., 2020; Zhao et al., 2018). For the 5′ UTR, it is harder to determine the effect of the sequence itself on mRNA stability because that region also has a key effect on translation initiation. In eukaryotes, 5′ caps and 3′ poly-A tails (Figure 1) are the primary features of the UTR regions that protect mRNAs from degradation (Mugridge et al., 2018).

Diverse, alternative polyadenylation mechanisms in eukaryotes are activated by different signals in 3′ UTR sequences and lead to differing poly-A tails and 3′ UTR lengths; this region is highly interactive with RNA binding proteins, microRNA and long noncoding RNAs. These interactions and the 3′ UTR length influence mRNA stability and decay, but also influence mRNA translation, as extensively reviewed elsewhere (Tian and Manley, 2017).

In the past decades, it has been suggested that the translation process may influence mRNA stability in yeast, as reviewed previously (Hanson and Coller, 2018). More recently, this connection gained additional attention in extensive studies in a range of eukaryotes, which all clearly demonstrated a positive correlation between the presence of certain codons in ORFs and the stability of the corresponding mRNAs (Bazzini et al., 2016; Burow et al., 2018; Forrest et al., 2020; Harigaya and Parker, 2016; Hia et al., 2019; Jeacock et al., 2018; Mishima and Tomari, 2016; Narula et al., 2019; de Freitas Nascimento et al., 2018; Presnyak et al., 2015; Wu et al., 2019). In particular, specific codons are observed to be more abundant in mRNAs with a longer half-life. This observation was captured by a newly proposed codon index, the codon stability coefficient (CSC), which can be calculated for each codon as the correlation coefficient between the codon frequency in transcripts and their mRNA half-life (Presnyak et al., 2015) (Figure 3A). In several studies, it was found that this coefficient correlates moderately with the tRNA availability index (tAI). The latter index is based on the gene copy number of tRNAs available to decode a certain codon (Presnyak et al., 2015; dos Reis et al., 2003). The observation that codons leading to high mRNA stability seem related to more-abundant tRNAs, remarkably suggests that the translational process may influence the stability of mRNAs. This was further supported by experiments that compared the mRNA stability with and without blocking the translation process (Bazzini et al., 2016; Presnyak et al., 2015; Wu et al., 2019). These experiments showed that when translation is inhibited, the mRNA half-life times are reduced, especially for transcripts with high "codon optimality."

On top of codon identity, a link is also suggested between amino acid identity and mRNA decay. A few amino acids are also specifically correlated to more or less stable mRNAs (Bazzini et al., 2016; Forrest et al., 2020; Narula et al., 2019; Wu et al., 2019). It is hypothesized that for these amino acids' higher or lower intracellular concentrations influence the amount of available tRNAs for translating those amino acids and hence influence translation elongation rates and consequently mRNA stability. In summary, several lines of evidence suggest that faster translation elongation leads to higher mRNA stability.

A potential molecular mechanism connecting translation elongation rates to mRNA decay has recently been unraveled (Figure 3B). Clear evidence was found in yeast that the de-adenylating Ccr4-Not complex directly interacts with ribosomes that are not loaded with a new tRNA in their A-site (Buschauer et al., 2020). Hence, this complex can sense slow-moving ribosomes and then triggers de-adenylating of the poly-A tail; after which, the RNA helicase Dhh1p activates de-capping, eventually

Figure 2. An Overview of Reported Factors Involved in Protein Production

(A–C) Factors at the level of (A) transcription and mRNA decay, (B) translation initiation, and (C) translation elongation. Factors that can be related to codon usage are connected by the gray bar. Factors that have only been well-described in eukaryotes (orange) or prokaryotes (green) are highlighted; other factors have been observed in both domains of life.

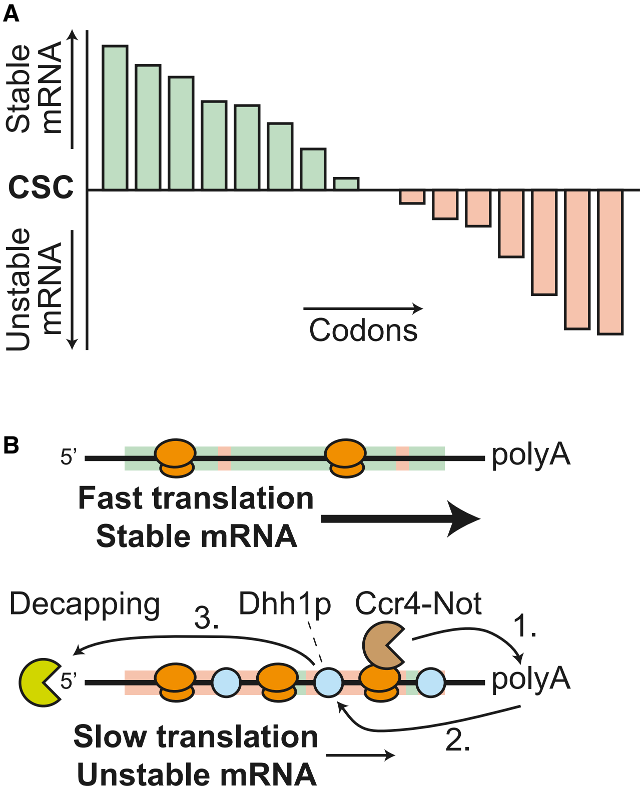


Figure 3. Codon Usage and Translation Elongation Are Related to mRNA Stability in Several Eukaryotes

(A) A schematic representation of a codon-stabilization coefficient (CSC) plot, based on recent studies in several eukaryotes, e.g., Presnyak et al. (2015). Bars for each codon represent the correlation between the codon frequency in the transcripts and the half-life of the transcripts. Positive correlations (green) indicate codons that are more abundant in mRNAs with a longer half-life time, whereas negatively correlated codons (red) are overrepresented in less-stable mRNAs. For illustrative purposes, only a few codons are depicted; in a real plot, the CSC value for all 61 amino-acid-encoding codons would be shown.

(B) mRNAs with more codons with a high, positive CSC value (green) are observed to be translated faster by the ribosomes because, for example, those codons have more abundant cognate tRNAs. In the eukaryotic model organism yeast, a molecular mechanism has been elucidated that can explain the connection between slowly translated mRNAs and mRNA decay rates. The de-adenylating Ccr4-Not complex can directly interact with ribosomes that are not loaded with a new tRNA in their A-site (Buschauer et al., 2020). Likely, this complex senses slow-moving ribosomes and then triggers de-adenylating of the poly-A tail, and next the RNA helicase Dhh1p activates de-capping and subsequent mRNA decay.

resulting in mRNA decay (Mishima and Tomari, 2016; Radhakrishnan et al., 2016; Webster et al., 2018).

A link between codon usage and mRNA stability was also suggested for the bacterium *E. coli* to have a major role in protein production efficiency (Boël et al., 2016). This study focused on expression data from a large set of plasmid-encoded heterologous genes transcribed by T7 RNAP. So far, no genome-wide analyses are available on such correlations in bacteria for native gene expression.

In relation to that, it is interesting to note here that recent structural studies in *E. coli* and *Mycoplasma pneumonia* clearly show that the RNAP complex can be linked to ribosomes in a so-called expressome, which leads to the coupling of transcription elongation to the translation process (O'Reilly et al., 2020). However, it was also recently reported that this coupling is not present in all bacteria because it was demonstrated in *Bacillus subtilis* that its RNAP moves faster than its ribosomes, in so-called runaway transcription (Johnson et al., 2020). The consequences of the presence and absence of this mechanism in different bacteria for the influence of codon usage and translation elongation on transcription deserve further analysis.

Lastly, another mRNA-mediated mechanism was discovered in *E. coli*, in which specific heterologous sequences of the mRNA appear to be toxic to the bacterial cells. It is not uncommon that the expression of heterologous proteins causes growth retardation in the expressing host, usually related to a protein production burden. However, a recent study surprisingly demonstrates that the growth retardation for specific heterologous mRNAs still happens when translation is blocked (Mittal et al., 2018). It is hypothesized that specific mRNA secondary structures cause toxic effects in the cell via a yet unknown mechanism.

Overall, our understanding of control mechanisms that determine mRNA concentrations is increasing. It is clear that mRNA abundance is affecting the downstream translational process and, remarkably, also vice versa translational processes seem to exert control on mRNA levels.

TRANSLATION INITIATION

For transcripts to be translated into protein, ribosomes need to associate with the 5′ UTR of the mRNA and start translating the ORF from the start codon. The translation initiation process is considered one of the most influential steps in translation efficiency.

In prokaryotes, it is generally assumed that translation initiation begins when the 30S ribosomal subunit recognizes a ribosome binding site (RBS) in the 5′ UTR. The RBS usually contains a Shine-Dalgarno (SD) sequence, which has high complementarity to the 3′ end of the 16S rRNA of the 30S ribosomal subunit, the so-called anti-Shine-Dalgarno sequence (aSD) (Shine and Dalgarno, 1974). In eukaryotes, the ribosome binds the 5′ cap or an internal ribosome entry site (IRES) and usually translation initiation is further controlled by a Kozak sequence (Kozak, 1981), a motif surrounding the start codon with a relatively high abundance of adenines (Leppek et al., 2018). However, because most recent studies on translation initiation used *E. coli* as a model, we mostly discuss prokaryotic translation initiation. For detailed insights on translation initiation and the 5′ UTR in eukaryotes, we refer to other recent reviews (Leppek et al., 2018; De Nijs et al., 2020).

Numerous studies, mostly investigating heterologous protein production in *E. coli*, have found that strong mRNA secondary structures around the RBS/SD region severely hamper translation initiation (Boël et al., 2016; Cambray et al., 2018; Goodman et al., 2013; Kudla et al., 2009). The mRNA folding in this region is also regularly observed to be influenced by the codon usage at the start of the ORF. A recent study aimed to quantify the influence of mRNA secondary structures more accurately by designing strong RNA hairpins in the 5′ UTR region of a reporter protein. Although secondary structures located far from the SD only result in less than 2-fold repression of translation, secondary structures close to the SD were shown to repress translation more than 100-fold; the repression levels are proportional to the free energy needed to unfold the RNA hairpins (Espah Borujeni et al., 2017). Furthermore, a study that introduced synonymous codon mutations throughout ORFs of two native *E. coli* genes revealed that, especially mutations leading to relatively strong, predicted mRNA secondary structures that include the RBS, result in significantly decreased protein production levels (Bhattacharyya et al., 2018).

Although most studies base their mRNA structure predictions on *in silico* folding energy models, some recent studies have applied transcriptome-wide *in vivo* experiments to determine mRNA secondary structures. Experimental high-throughput measurements of mRNA secondary structures can be performed by cell-permeable chemicals that react selectively with non-paired RNA bases, e.g., SHAPE probes that acylate 2′ hydroxyl groups of unpaired nucleotides (SHAPE-MaP) (Siegfried et al., 2014) or dimethyl sulfate that modifies unpaired adenine and cytosine residues (DMS-seq) (Rouskin et al., 2014). As the next step, cDNA is generated from the chemically modified RNAs, and next-generation DNA sequencing allows for mapping of the modifications in non-structures regions and, hence, allows the elucidation of non-structured and structured mRNA regions. One of these studies, based on SHAPE-MaP in *E. coli*, demonstrated that the translation efficiency of native genes is, in large part (40%), determined by mRNA structures covering the RBS (Mustoe et al., 2018).

The improved resolution of mRNA structure measurements also allowed the study of two alternative models for translation initiation: the equilibrium model and the kinetic model. In the equilibrium model, the ribosome, once bound, remains and creates a new equilibrium mRNA secondary structure. In the kinetic model, however, there is a continuous competition between the unfolding and refolding of the mRNA and association and dissociation of the ribosome. Experimental data, as well as a theoretical biophysical approach, now suggest the kinetic model best explains translation initiation in *E. coli* (Espah Borujeni and Salis, 2016; Mustoe et al., 2018). This also allows for "ribosome drafting" in some highly translated mRNAs, a mechanism in which successive ribosomes bind an mRNA faster than the mRNA can refold.

In contrast with the ribosome drafting mechanism, in eukaryotes, it was observed that ribosome clearance around the translation initiation site is required for high-expressing genes. It is suggested that codons directly after the start codon need to mediate relatively fast translation elongation to free up space for the next ribosome to initiate translation (Chu et al., 2014).

Although it is generally accepted that SD-aSD interaction is the main player involved in prokaryotic ribosome loading, new findings hint at alternative mechanisms regulating ribosome recruitment and translation initiation. Several bacterial species, for example, *Flavobacterium johnsoniae*, naturally lack SD sequences. In this species it was observed that at some key nucleotide positions upstream of the start codon (−3, −6, −13, and −23), the presence of adenine nucleotides is a positive determinant for translation initiation (Baez et al., 2019). The molecular basis for this observation is currently not known but, as the authors state, it seems reminiscent of the eukaryotic Kozak sequence, which also shows a preference of adenine at position −3. Furthermore, some recent *E. coli* studies on native and reporter gene expression report an enrichment in adenines at sites mostly upstream, or shortly downstream of the start codon for well-expressed genes (Komarova et al., 2020; Saito et al., 2020). It was demonstrated experimentally that these A-rich sequences contribute to the identification of translational start sites, suggesting that these adenines could be highly conserved as an alternative mechanism for start site selection in bacteria (Saito et al., 2020).

TRANSLATION ELONGATION

Codon Usage and Translation Rates

After successful initiation, ribosomes continue with translation elongation, i.e., the sequential decoding of the codons of the mRNA to synthesize the corresponding amino acid sequence. The effect of codon usage during translation elongation has been extensively studied by multiple methods, however, often leading to contrasting conclusions. A popular hypothesis is that codon usage controls the speed of ribosomal translation elongation. The underlying assumption is that translating ribosomes slow down when they encounter "sub-optimal" codons, e.g., codons that are decoded by less-abundant (amino-acid-loaded) cognate tRNAs or by lower-affinity-matching tRNAs through wobble base-pairing.

A decade ago, the ribosome profiling technique was developed to monitor translation elongation rates in a high-throughput manner (Ingolia et al., 2009). This approach is based on the high-throughput sequencing of ribosome-protected mRNA fragments, providing a snapshot of ribosome density throughout the transcriptome. Initially, differences in experimental ribosome profiling protocols and subsequent data analysis led to conflicting conclusions on whether translation elongation speeds are influenced by codon usage or not (Charneski and Hurst, 2013; Gardin et al., 2014; Li et al., 2014; Quax et al., 2013). However, in recent years ribosome profiling protocols and data analysis were refined, e.g., by the use of flash freezing to stall translation, instead of the use of cycloheximide (Weinberg et al., 2016). Improved protocols led to a better consensus that codon usage may influence the translation elongation speed, but that this effect is rather weak and that a multitude of other factors are also involved (Hanson and Coller, 2018).

Recently, more-sensitive approaches using cell-free translation systems (Buhr et al., 2016; Yu et al., 2015) and *in vivo* imaging of nascent polypeptide synthesis (Chekulaeva and Landthaler, 2016; Yan et al., 2016) have been established. These methods all confirmed that heterologous mRNAs with "optimal" codon usage are translated faster. However, these studies monitored the strong contrast between synthetic genes that were designed to have almost only optimal codons with non-optimized genes. Within natural genes, which often have fluctuating use of optimal codons along the ORF, translational speed differences are generally more subtle.

It was also demonstrated for eukaryotic translation, both *in vivo* and *in vitro*, that rare codons sometimes not only slow down translation, but they can even stall part of the elongating ribosomes, leading to premature translation termination (Yang et al., 2019; Yu et al., 2015; Zhao et al., 2017b).

Does an mRNA Secondary Structure Influence Translation Elongation?

Besides the influence of codon usage on translation speed, the mRNA secondary structure within an ORF was also suggested as influencing translation elongation. However, until recently, it was hard to verify that hypothesis because only rough *in silico* predictions of mRNA folding energy were available to estimate mRNA structures. However, the aforementioned development of several experimental protocols allows for probing RNA structure *in vivo* at a transcriptome-wide scale. Two studies in this field used different methods to both reach the conclusion that translating ribosomes in *E. coli* dissolve RNA secondary structures (Burkhardt et al., 2017; Mustoe et al., 2018), which is in line with the demonstration that the *E. coli* ribosome exhibits helicase activity (Takyar et al., 2005).

Apart from that finding, the DMS-seq analysis by Burkhardt et al. (2017) reported a strong correlation between mRNA secondary structures in an ORF and its translation elongation efficiency, suggesting that at least some of those structures can still be an obstacle for translating ribosomes. In contrast, the SHAPE-MaP analysis by Mustoe et al. (2018) could not confirm that correlation. Hence, despite advances in *in vivo* RNA structure mapping, it remains unclear to what extent mRNA structures influence translation-elongation rates. Refinement and application of these methods throughout multiple organisms are required to clarify this matter.

Co-translation Folding Mediated by the ORF Sequence

For a few specific proteins, single-molecule approaches have been used to accurately monitor translation elongation rates and related co-translational protein-folding processes. In some cases, it was clearly shown that the slow-down of translation elongation is crucial to facilitate proper co-translational folding of the nascent protein (Buhr et al., 2016; Kim et al., 2015).

Similarly, it has been demonstrated *in vivo* for some eukaryotes that codon usage is crucial for the folding and functionality of some circadian clock proteins, especially for the unstructured domains of these proteins. When the sub-optimal codon usage in unstructured regions of these circadian clock genes, as well as in a luciferase reporter gene, was changed to a more-optimal codon usage, the *in vivo* functionality of these proteins was compromised (Fu et al., 2016; Yu et al., 2015; Zhou et al., 2013, 2015). This folding hypothesis is further supported by broad bioinformatic analyses of genes from several organisms, based on which correlations are reported between less-optimal codons in unstructured regions in between more-structured protein domains (Pechmann and Frydman, 2013; Zhou et al., 2015). Despite the fact that these unstructured domains do not form defined structures (α helices or β sheets), they seem to have certain folds (e.g., coils) that can be essential for their functionality. These studies suggest that translation slows down to facilitate folding either of these unstructured domains themselves or at structural junctions between structured and unstructured domains.

However, a broader analysis of clusters of rare codons throughout many genomes in all domains of life challenges this observation of rare codons within unstructured domains (Chaney et al., 2017). That study, in fact, reports an enrichment of rare codons within structural domains, suggesting that translational slow-downs may be specifically relevant for the folding of smaller structural sub-elements. As an example, they show conservation of rare codon clusters for two proteins at the same "structural" positions throughout different organisms. Providing such comparative analyses for more proteins, as well as performing functional experiments on these, could strengthen the proof that sub-optimal codons are also relevant within structural protein domains.

Overall, there is clear case-based evidence on the effects of codon bias and translational speed on co-translational folding for some specific proteins. However, interpretation of these effects on a genome-wide scale is complicated, given the limited understanding of the genetic features determining translational speed and the subjective definitions of optimal and non-optimal codons. Furthermore, determining the relevance of the coding sequence on protein folding is challenging, as it is currently not possible to experimentally determine protein structures or folding processes in a high-throughput manner.

Translation Effects at the Start of the ORF

Another frequently reported and heavily debated observation is the slower translation at the 5′ end of an ORF. Some evidence for this has been based on ribosome profiling data and the higher frequency of rare codons in the first part of the ORF (Tuller and Zur, 2015; Tuller et al., 2010). A main hypothetical explanation for the presence of a so-called translational ramp at that location is the distancing between ribosomes to prevent detrimental ribosomal collisions. Still, there are alternative explanations for the observed codon bias at the 5′ of ORFs. A key alternative hypothesis is that a strong selection against mRNA secondary structures at the 5′ end to facilitate translation initiation of highly expressed genes is more important than the selection pressure for well-translated codons in that region of the ORF.

Interestingly, several studies that randomized synonymous codons in *E. coli*, usually for GFP as a reporter protein, found strong correlations between protein production and reduced mRNA secondary structures around the 5′ end of the ORF (Goodman et al., 2013; Kelsic et al., 2016; Kudla et al., 2009). A recent study tried to resolve the factors in the 5′ end of the ORF in a more systematic way by designing >200,000 different N-terminal tags for 32 codons, followed by a GFP reporter gene (Cambray et al., 2018). Several factors were varied in the N-terminal library design, including the presence of different-strength translational ramps, as well as the presence of mRNA secondary structures at different positions. Although no correlation was detected between translational ramps and expression, that study did demonstrate a major role of mRNA structural elements in RBS availability and, consequently, in overall protein production. However, as the authors admit, the conclusion that the presence of a translational ramp could not be detected in that study might have been the result of non-optimal design. Although it remains unclear to what extent translations ramps influence expression levels, it was demonstrated recently that a ramp can decrease the resource costs of expression (Frumkin et al., 2017), likely by preventing ribosome jamming and translational abortion events (Tuller et al., 2010).

Other Factors Observed at the Translational Level

Apart from the effect of single codons on translational dynamics, it was observed previously that specific codon pairs might also influence translational processes (Buchan et al., 2006; Gutman and Hatfield, 1989). In yeast, ribosomal stalling has been reported for a small subset of codon pairs, mostly when they occur in a specific order (Gamble et al., 2016). Recently, a mechanistic explanation for that observation was found. It was determined that interactions of specific codons pairs with their tRNAs, mostly involving wobble-base pairing, induce certain conformational changes in the ribosomes that lead to stalling (Tesina et al., 2020).

The use of sub-optimal pairs of codons has also been proposed as a strategy to create life-attenuated viruses for vaccine development. However, there has been a lively debate about whether the decreased expression of those viruses in eukaryotic host cells should be attributed to sub-optimal codon pairs or, alternatively, to sub-optimal dinucleotide pairs (Kunec and Osterrieder, 2016). A recent study that aimed to disentangle the effects of dinucleotide bias and codon-pair bias in virus attenuation concluded that sub-optimal codon pairs primarily caused the decreased translational efficiency (Groenke et al., 2020). That study shows that the influence of sub-optimal codon pairs can, at least partly, be related to decreased mRNA stability, in line with the previously discussed correlation between codon usage, translation efficiency, and mRNA stability in eukaryotes.

In bacteria, the presence of SD-like sequences within ORFs was previously suggested to result in a slow down of the translation-elongation process (Li et al., 2012). However, that observation was later toned down in a re-evaluation of ribosome-profiling data, which concluded that SD-like sequences have little or no effect on translational pausing (Mohammad et al., 2016). Recently, a bioinformatical analysis studying the evolutionary conservation of those SD-like sequences in ORFs of several bacterial species, concluded that they are less conserved than would be expected by random chance (Hockenberry et al., 2018). This suggests a negative evolutionary selection against SD-like sequences, hinting at a potential decrease in fitness caused by the presence of those sequences within ORFs, possibly because they could induce mistranslation or erroneous frameshifting. In conclusion, it seems that SD-like sequences are not frequently used in nature because of detrimental by-effects on translation and that they do not have a major role in controlling translation elongation rates.

Another recent study has revealed an interesting effect of certain short amino acid motifs on translation elongation. That study focused on mutating codons at position 3, 4, and 5 of a GFP reporter in *E. coli* and allowed non-synonymous mutations (Verma et al., 2019). They identified specific amino acid motifs at the start of the ORF that lead to high translation efficiency, independent of specific codons or mRNA structures. At the same time, they identified detrimental amino acid motifs in the 5′ region of the ORF, which can cause pausing of the translation and lead to increased translational abortion. This observation was explained by specific interactions of the nascent peptide motif with the ribosome exit tunnel that could lead to ribosomal stalling and drop-off.

There are more reports of specific peptide motifs that cause stalling or translational slow down, likely via interactions in the ribosome exit tunnel. Motifs such as poly-proline sequences can slow down or stall translation in organisms throughout all domains of life (Huter et al., 2017; Wilson et al., 2016). In addition, it was observed in *E. coli* that four specific amino acid triplets completely stalled translation and were avoided within its proteome (Navon et al., 2016). It is good to realize that both in evolution and in synthetic biology approaches, the flexibility to evolve or design acceptable changes in amino acid sequences, without altering residues that are critical for protein functionality, may sometimes result in improved translation efficiency.

Furthermore, translational speed can be influenced by the modifications of mRNA and tRNAs. It is well established that the great diversity of tRNA modifications, especially modifications of ribonucleotides in anticodon regions, can have a major effect on translation rates and fidelity (Chou et al., 2017; Kimura et al., 2020; Nedialkova and Leidel, 2015). Recently, it was also observed that modifications of mRNA, e.g., N^6-methyl-adenosine and N^4-acetylcytidine, influence translation elongation and mRNA decay in both eukaryotes and bacteria (Arango et al., 2018; Choi et al., 2016, 2018; Zhao et al., 2017a).

Translational Fidelity versus Translation Rate and Translation Termination

Apart from governing translational speed, ORF sequence features such as codon usage have been postulated to govern translational fidelity. Even though support for this theory has been provided by bioinformatic analyses (Drummond and Wilke, 2008), only very recently has experimental evidence for this hypothesis been obtained. Using a ‘‘deep proteomics’’ approach, translational errors have been identified in the proteomes of *E. coli* and *Saccharomyces cerevisiae* (Mordret et al., 2019). That study revealed that translation errors are relatively abundant, occurring on average once every 1,000 amino acids. Transcriptional error rates occur much less frequently, at about 1 in 25,000 nucleotides (Traverse and Ochman, 2016).

Both the misloaded tRNAs and tRNA-codon mispairing can cause translation errors, but the latter error is more abundant. In that case, wrong amino acids are delivered by near-cognate tRNAs, which have only one mismatch between codon and anti-codon (Mordret et al., 2019). Interestingly, the effect of mistranslation events is probably reduced because the genetic code has evolved such that these near-cognate tRNAs often deliver amino acids with similar chemical properties. Some codons are more sensitive to mistranslation than others, and that pattern was relatively similar both in yeast and in *E. coli*, suggesting that evolutionarily conserved mechanisms or universal chemical interactions lead to occasional mistranslation.

The same study also demonstrated a negative correlation between translation speed and translation fidelity, suggesting a trade-off between optimizing coding sequences for translational speed and fidelity. This fidelity theory (slow downs to reduce translational errors) is an interesting alternative explanation for the aforementioned occurrence of "slow" codons in structurally important regions, which, in many reports, is explained by the co-translational folding theory (Buhr et al., 2016; Kim et al., 2015).

Frameshifting during translation has an even bigger effect on protein function than amino acid misincorporation because the downstream sequence is completely mistranslated. However, the operation of ribosomes and its translation elongation factors seems to limit frameshifting. Recently, another mechanism for frameshift fidelity was observed in human cells (Wan et al., 2018). It was suggested that periodic pairing of certain "sticky codons" on the mRNA with complementary triplets in the rRNA, near the exit of the ribosomal mRNA channel, helps to prevent frameshifting. That conclusion was supported by substitution of sticky codons by synonymous counterparts, which led to a 4-fold increase in frameshifting, as well as mutating the complementary triplet at the exit of the ribosomal mRNA channel, which also influenced the frameshifting rate. Finally, it seems that these sticky codons are naturally underrepresented in a non-coding frame in eukaryotic genomes, which may be to prevent accidental frameshifting (Wan et al., 2018). This mechanism deserves further analysis throughout different types of organisms and may cause certain codon preferences to limit frameshifting.

At the end of the translation-elongation process, the ribosome encounters a stop codon, and upon binding of a release factor (a protein mimic of a tRNA), the translation is ended, and the ribosome is released from the mRNA. However, in rare cases, translation read through happens, generally leading to the synthesis of non-functional proteins. If such a read-through event takes place, the ribosomes either encounter an in-frame stop codon within the 3′ UTR or they get stalled at the end of the mRNA (Wilson et al., 2016). These read-through proteins are generally degraded co- or post-translationally (Arribere et al., 2016). Some organisms may prevent translational read through by using tandem stop codons, which are, for example, observed more frequently in the 3′ UTR of ciliates (Fleming and Cavalcanti, 2019).

THE INTERACTIONS BETWEEN DIFFERENT FACTORS

Cooperative and Counteracting Features

As discussed, distinct factors are involved in different steps of the gene-expression process, and they interact with each other in multiple ways. Some factors in the protein-production process act in a cooperative fashion. As a remarkable example of that, the translation-elongation efficiency and mRNA stability in eukaryotes have been demonstrated to be mechanistically linked, leading to positive feedback between translation elongation and mRNA stability (Buschauer et al., 2020; Radhakrishnan et al., 2016). However, other sequence features may also influence each other negatively. For example, a high-affinity SD sequence and well-translated codons in the 5′ region of the ORF could form a base pair and, consequently, form undesired mRNA secondary structures that hamper efficient translation initiation. These counteracting and cooperative features complicate the evaluation of individual factors.

Several studies have attempted to reveal new factors and to disentangle their connections in recent years. Many of those studies applied randomized or systematically designed reporter gene-variant libraries of GFP in *E. coli.* (Cambray et al., 2018; Frumkin et al., 2017; Goodman et al., 2013; Kudla et al., 2009). The consensus of those studies is that gene expression is significantly affected by strong (predicted) mRNA secondary structures in the 5′ UTR and the 5′ region of the ORF. However, a large part of the variation in expression levels in those studies is explained by a range of other factors, and a substantial part of the observed fluctuations cannot be explained at all. Furthermore, it is not certain that those studies properly reflect features that are relevant to native genes. Nevertheless, a number of recent studies on native gene expression in *E. coli* also suggest that mRNA structures and associated RBS availability are key factors that determine the expression rate of natural genes (Kelsic et al., 2016; Mustoe et al., 2018).

A combination of different experimental approaches to study native gene expression was recently performed in yeast, integrating multiple omics data and measurements of mRNA and protein half-life times (Lahtvee et al., 2017). The latter is an often-overlooked factor because proteins with shorter half-lives need to be translated at higher levels to sustain sufficient protein levels. That study found large differences in protein yield per mRNA, varying up to 400-fold among some proteins, suggesting an important role for the efficiency of the translation processes. However, when accounting for all proteins, translation-elongation efficiency only explained 15% of the protein abundance observed, whereas mRNA abundance was the most important explanatory factor for protein levels (explaining 61%). A large study on a diverse set of heterologous proteins in *E. coli* also reported mRNA abundance as the main predictor for protein abundance (Boël et al., 2016). However, it is important to realize that mRNA abundance can also be influenced by translation efficiency.

Influence of Gene Designs on Resource Consumption and Growth

An important, overarching aspect for protein production is the high metabolic costs associated with transcription and translation processes. Those additional costs include "materials," such as demands for ATP, nucleotides, and amino acids, but also the extra demand for the transcriptional and translation factors, such as RNAPs and ribosomes. There is an evolutionary pressure on the genome in general, and the architecture of genes and their regulation in particular, to reduce metabolic costs to optimize cellular fitness. Within synthetic-biology applications, the reduction of energy and resource requirements is of importance for gene design.

Hence, recent efforts studied growth parameters of microbial cells harboring codon-variant libraries of reporter genes (e.g., GFP) or of a growth-essential gene. The relative fitness of different variants was recorded, either by measuring growth curves for individual strains or by performing competition experiments between them (Cambray et al., 2018; Frumkin et al., 2017; Kelsic et al., 2016). One of the main conclusions is that, especially for highly expressed genes, a high level of protein produced per mRNA is a resource-efficient way for high expression.

So, even though, in nature, high mRNA levels are typically correlated to high expression, boosting expression solely by high mRNA levels is not the best strategy. Extremely abundant mRNAs potentially imply excessively high transcription costs or may sequester excessive amounts of ribosomes from the limited pool. In contrast, we note that the strategy to keep mRNA levels low and, rather, to couple it to highly efficient translation can increase the cell-to-cell variability in mRNA and protein concentrations (Taniguchi et al., 2010). Thus, to achieve both high resource efficiency and low cell-to-cell expression variability, nature and synthetic biologists need to properly tune the translation efficiency per mRNA.

BIOTECHNOLOGICAL CHALLENGES AND OPPORTUNITIES FOR GENE DESIGN

Innovations in DNA synthesis and genetic engineering have tremendously accelerated the capacity to express synthetic genes. However, based on data from consortia aiming to resolve large numbers of protein structures, it is estimated that only about one-half of the attempts for heterologous protein production led to successful expression (Parret et al., 2016). In practice, in molecular biology and synthetic biology projects, the expression of synthetic genes regularly leads to sub-optimal production or problematic growth because of the excessive expression burdens.

Limitations of Codon Optimization Algorithms

Synthetic genes for heterologous protein production are typically designed with codon-optimization algorithms, which generally optimize a particular ORF, adapting it to a codon-usage index of the expression host (Parret et al., 2016). Those codon indices are frequently determined with either the codon usage within a set of highly expressed reference genes (e.g., CAI) or the tRNA copy numbers (e.g., tAI) in the host cell. Some academic and commercial algorithms also take alternative parameters into account, such as GC content and avoidance of certain regulatory motifs, such as SD sequences or repeats (Gould et al., 2014). Only a few algorithms additionally aim to minimize mRNA secondary structures (Gould et al., 2014), even though the folding in the translation-initiation region, certainly in prokaryotes, is a key determinant of expression. A promising exception is the novel 31C-FO algorithm, which aims to minimize mRNA folding of the 5′ UTR and the first 48 bases of the ORF (Boël et al., 2016). At the same time, that algorithm optimizes codon usage by only including 31 codons that are correlated to high expression in *E. coli*. That algorithm was reported to lead to successful expression of several proteins by Boël et al. (2016) but has not been reported in other studies yet, and no easy tool for that algorithm is available so far.

Generally, the features involved in gene expression, individually or in concert with others, are still not understood in sufficient detail to compose robust optimization algorithms for relevant host organisms. Multi-parameter algorithms, such as EuGene or DNA-Tailor (D-Tailor) (Gould et al., 2014), typically leave the setting of specific objectives up to the users, which, in practice, is hard to decide upon, given the unknown weight of the different factors. Furthermore, it has been shown that a so-called design-of-experiments approach, which systematically varies multiple factors, is no guarantee for successful expression because not all relevant factors are known yet or are not known in sufficient detail (Cambray et al., 2018).

In addition to expression levels, proper protein folding is important for the functional production of proteins. The accumulating evidence on the role of codon usage in protein folding led to several approaches that aimed to include the translation-speed landscape to accommodate folding of structural elements. For example, codon-harmonization algorithms have been proposed to tackle this issue (Angov et al., 2008; Buhr et al., 2016). These algorithms have as their objective to copy the *native*-codon-usage landscape of a gene-of-interest (distribution of rare and frequent codons in the organism from which the gene originated natively) into a *heterologous*-codon-usage landscape (similar distribution of rare and frequent codons in the context of the expression host). However, codon harmonization does not always give the best expression levels in *E. coli* when comparing the production levels of codon-harmonized gene variants with native genes or CAI-codon-optimized genes for some membrane proteins (Claassens et al., 2017).

In some studies, sub-optimal codons or SD-like sequences have been included in ORFs to slow down translation in between structural domains, which was reported to improve protein solubility in a few cases (Hess et al., 2015; Vasquez et al., 2016). That, however, requires laborious, detailed studies to determine exactly the position and strength of the required translation pauses to optimize the folding of a specific protein. Furthermore, there is no full understanding yet on the role of the coding-sequence features for translational speed; this all restrains robust design approaches for proper folding of proteins.

In summary, improving heterologous protein production by codon-optimization algorithms often remains a trial-and-error approach. Success rates can be increased by testing multiple different codon-optimized variants, but that also increases experimental labor and costs.

UTR Optimization Strategies

In numerous studies, the 5′ UTR has been identified as a critical region that determines translation-initiation efficiency in protein production. As discussed, few of the available codon-optimization algorithms take the 5′ UTR into account and do not have integrated functionality to avoid detrimental mRNA structures in the translation-initiation region. Nonetheless, some specific tools have been developed to design optimized 5′ UTR regions for bacterial protein production, which generally try to design 5′ UTRs to have strong and accessible RBSs, taking into account the downstream ORF region. Hereto, these tools have used *in silico* mRNA folding energy calculations (Bonde et al., 2016; Jeschek et al., 2016; Salis et al., 2009). Despite their wide use and relatively successful predictions, they still suffer from the limited reliability of *in silico* RNA structural predictions. Recently emerging experimental tools for measuring *in vivo* RNA folding may become helpful to assess the validity of computational predictions (Rouskin et al., 2014; Siegfried et al., 2014).

Alternatively, standardized 5′ UTR modules have been employed for robust gene expression, for example, by using combinations of well-expressed 5′ UTRs and N-terminal tags (Ki and

Pack, 2020). In addition, bicistronic RBS modules have proven highly useful because these modules partly uncouple translation-initiation efficiencies from the ORF sequence (Cambray et al., 2018; Mutalik et al., 2013). These bicistronic design elements (BCDs) have been shown to allow for tuned and improved expression levels in *E. coli* and *Corynebacterium glutamicum* (Claassens et al., 2019; Nieuwkoop et al., 2019; Sun et al., 2020).

The initiation mechanisms in the 5′ UTR in eukaryotes seem more diverse and complicated than do those for prokaryotes. However, recent studies have shown that the 5′ UTR sequence has great potential for tuning the expression in eukaryotes, such as *S. cerevisiae* or Chinese hamster ovary-S (CHO-S) cells (Ding et al., 2018; Petersen et al., 2018; Weenink et al., 2018). These studies provided modular 5′ UTRs designs that work relatively well with low-context dependence on the downstream ORF. One of the key factors that improve the performance of those 5′ UTR is the reduction of mRNA secondary structures in that region.

The 3′ UTR is less studied in relation to expression efficiency, but it has also been reported to influence mRNA stability and transcription termination efficiency, thereby modulating expression efficiency. Examples of 3′ UTR engineering in bacteria are scarce, so far. For yeast and human cell lines, some short synthetic 3′ UTR modules have been developed that relatively robustly increase expression for multiple genes throughout multiple species but also seem partly dependent on the upstream ORF sequence (Cheng et al., 2019; Curran et al., 2015).

An important part of the influence of 5′ UTRs and 3′ UTRs on protein production is explained by their roles in mRNA stability. An alternative, promising approach to improve mRNA stability for protein production, is through the circularization of mRNAs, which also occurs in nature. Synthetic circular mRNAs can, for example, be generated by harnessing the mechanism of self-splicing introns (Perriman and Ares, 1998; Wesselhoeft et al., 2018). A recent surge of research in this field showed promising applications for protein production driven by synthetic, circular mRNA transcripts in eukaryotes. Because canonical-eukaryotic translation initiation relies on the 5′ cap, alternative translation-initiation mechanisms, such as IRES or N^6-methyladenosine modifications, are required to ensure sufficient translation initiation in circular mRNAs. Furthermore, it has been proposed that the translation of circular mRNAs can be increased by creating an infinite ORF, by removing the stop codon of the ORF (Perriman and Ares, 1998); the same ribosomes will repeatedly translate the same sequence, leading to a multimeric protein, and individual functional proteins can be produced by introducing protease cleavage or self-cleavage sites in the polypeptide. A recent review elaborates in great detail on the developments of engineering of circular RNA (Costello et al., 2020).

Randomization, Smart Selection, and Machine Learning

The number of studies that randomly vary sequences in promoters, 5′ UTRs, and the start of the ORF have steadily increasing, mostly for GFP. This randomization approach may also be relevant for optimizing or fine-tuning the production of more biotechnologically relevant proteins (Figure 4). However, unlike expression levels of reporter proteins, levels of most proteins of interest are generally hard to screen with sufficient throughput from large randomized libraries. Still, some well-expressing modules identified in reporter-based screens, e.g., promoters or 5′ UTR-N-terminal tag peptide combinations, have been used successfully for the optimized production of other proteins.

When randomly optimizing the coding sequence, or at the junction of the 5′ UTR and coding sequence, novel approaches are required to screen for well-expressed gene variants in the case of non-reporter proteins. One simple approach is to fuse the protein of interest to a reporter protein, but such a fusion frequently distorts the function of the protein of interest. An alternative method, not based on a protein fusion, was recently established by translational coupling of the protein of interest to a selectable antibiotic-resistance reporter. This so-called TARSyn system was demonstrated for the high-throughput selection of optimized 5′ UTR:ORF junctions for the expression of antibody proteins in *E. coli* (Rennig et al., 2018) (Figure 4). We consider the development of selection and screening systems of well-expressed ‘‘randomized’’ sequences to be a very promising avenue for further exploration.

Alternatively, data collected from large-scale randomization studies on reporter proteins or growth-selectable markers may help to generate better predictive algorithms (Figure 4). These large-scale data could serve as training sets for machine learning. Different types of machine learning can be employed to generate more reliable algorithms to improve the design of synthetic genes (de Jongh et al., 2020).

A recent, innovative study that used machine learning focused on predicting the influence of different 5′ UTR sequences in *E. coli* (Höllerer et al., 2020). The study developed an innovative reporter system, based on a recombinase protein, to quantify the expression from a large library of randomized 5′ UTR sequences (Figure 4). At a certain expression level, that site-specific recombinase flips a DNA sequence, which is located directly next to the 5′ UTR on the same plasmid. Subsequent, high-throughput sequencing of short DNA fragments that contain both the 5′ UTR and the potentially flipped DNA sequence gives information on both 5′ UTR genotype and related expression phenotype, which provided data on the recombinase expression from 300,000 different 5′ UTRs, which were fed into machine learning. The analysis, surprisingly, revealed that, rather than mRNA secondary structures, the presence and positioning of the SD are most important for high protein production in this case, possibly because the 5′ end of the recombinase ORF was unlikely to form strong mRNA structures with any UTR. The machine-learning approach was used to develop a new 5′ UTR design algorithm that Höllerer et al. (2020) report outperformed currently available algorithms, which are mostly based on biophysical models. However, this algorithm has not yet been tested for ORFs other than the recombinase ORF in that study.

Likewise, successful 5′ UTR prediction algorithms based on multiple regression or machine learning approaches have been developed for yeast (Cuperus et al., 2017; Decoene et al., 2018; Ding et al., 2018). Such big-data analyses, based on randomized sequence libraries, seem a promising road toward better predictive algorithms for robust regulation of synthetic genes.

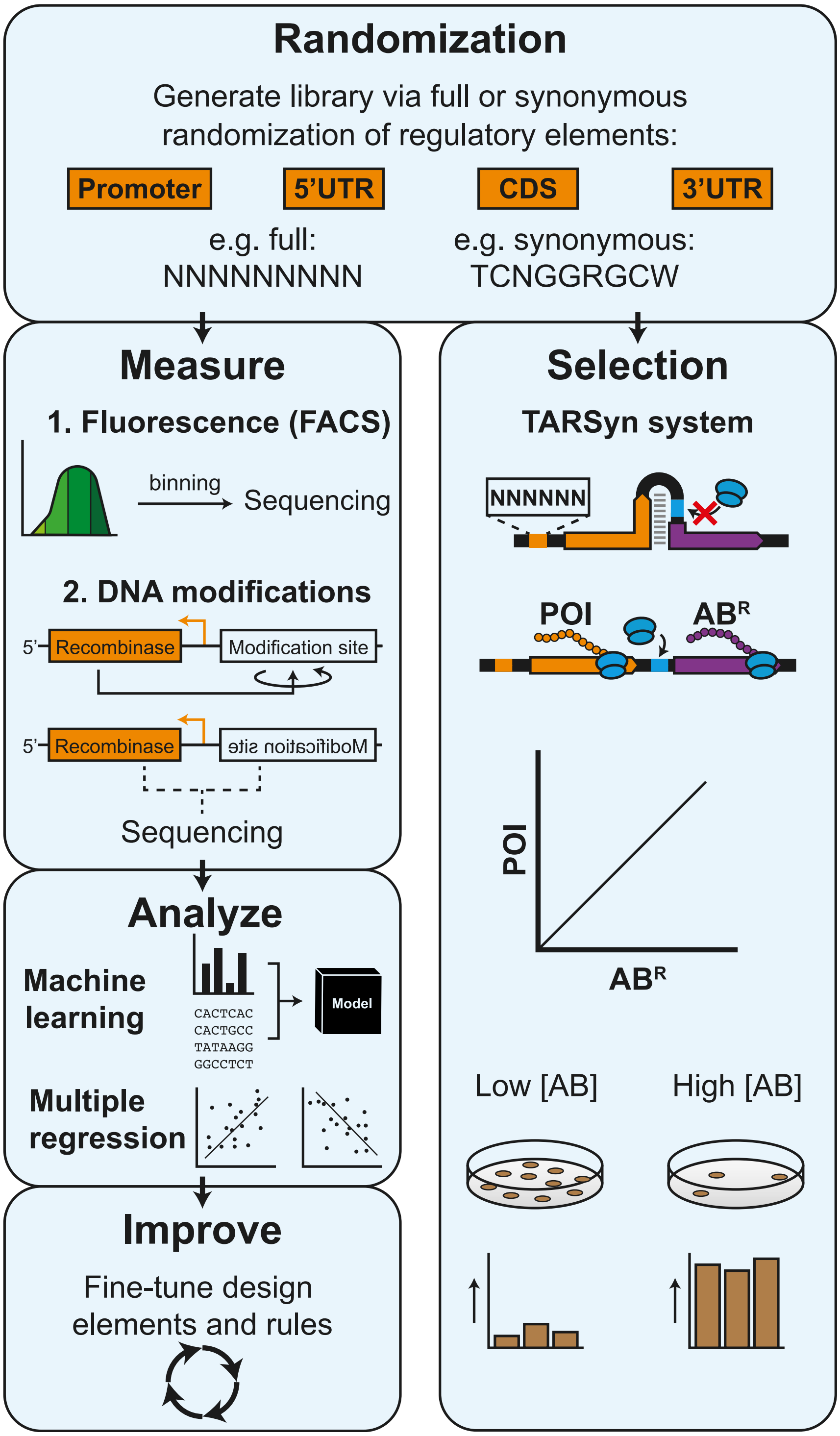


Figure 4. Overview of a Typical Workflow Randomizing Gene Regulatory and ORF Sequences

After randomization of genetic regulatory sequences or (part of) the codons of an ORF, the protein production by the resulting (large) variant library can be measured and binned according to fluorescence levels by fluorescent activity cell sorting (FACS). As an alternative to fluorescent-reporter proteins, a DNA-modifying enzyme can be used as a reporter because its expression can be assessed by high-throughput sequencing of modifications in the DNA. The latter approach was demonstrated for the expression of a randomized 5′ UTR library mediating expression of a recombinase that flips a nearby DNA modification site. In the same single-sequencing read, the 5′ UTR variant can be identified and whether the site was flipped or not, allowing high-quality, large-scale data on expression levels (Höllerer et al., 2020). Analysis of generated large-scale data is typically performed by multiple regression analysis and, recently, by machine-learning algorithms. Next, understanding of expression levels can be further improved by correlations or rules derived from the analysis, and expression could be further studied during next-iteration rounds in which randomized sequence space can be limited, based on the results of the previous iterations. As an alternative to the learning cycle, a direct-selection system can be used for the selection of high-expressing variants. For example, the so-called TARSyn system allows selection of high-expressing clones based on antibiotic resistance (Rennig et al., 2018). The expression of a (non-reporter) protein of interest is translationally coupled to downstream antibiotic resistance, allowing for easy selection for high expression under high antibiotic concentrations.

CONCLUSIONS

Despite significant efforts to elucidate the effect of codon usage and other gene features on protein production, it is still not completely understood. During the past decade, genome, transcriptome, proteome, and translatome (ribosome profiling) data became increasingly available. Bioinformatic analysis of those data has provided relevant insights into coding features and their relation to protein production. Recently, such analyses, combined with half-life measurements of mRNA, led to the discovery that optimal translation of an mRNA increases its stability in eukaryotes. However, many factors and their relevance are still unclear and require further investigation and, possibly, new experimental approaches.

One of the key knowledge gaps is the role of mRNA secondary structures, which is suggested to have a pivotal role in translation initiation and elongation, but its true effect is still unsettled. Recently, emerging protocols enabled the generation of transcriptome-wide *in vivo* mRNA structural data. However, groups using such methods report partly contradicting results for the role of mRNA secondary structures on translation-elongation efficiency (Burkhardt et al., 2017; Mustoe et al., 2018). Further refinement and validation of those protocols are required to improve understanding of mRNA structures on translation. Another poorly explored territory is the influence of the ORF's codon sequence on co-translational folding and fidelity. Bioinformatic analysis of genome and translatome data suggested important roles for translation speed on protein folding, at least for some proteins. Detailed molecular studies focusing on some specific proteins have confirmed that codon usage has a crucial role in folding. However, data and protocols to test this hypothesis experimentally for larger sets or proteins or on a proteome-wide scale are lacking.

A general limitation of studying genetic features within native genes (in a certain organism or under certain conditions) is the complexity in detecting "weak signals" from relevant factors within sequences that underwent optimization during millions of years of evolution. Alternative approaches, based on synthetic gene libraries, represent strong complementary methods in which many variants for a single gene can be generated to probe relevant factors. However, these "controlled" studies have, so far, been able to provide generic explanations for variable protein production levels only to some extent and are mostly based on correlating expression with known factors. In addition, those studies have mostly focused on a few highly expressed reporter proteins (mostly GFP), which may make conclusions biased.

Machine-learning approaches may help to further elucidate unknown features and factors in a more-unbiased way. Such approaches have recently been applied to analyze expression data from randomized synthetic libraries of promoters and 5′ UTRs. Such approaches may be promising for developing better predictive algorithms. However, large datasets are required for machine-learning algorithms to generate predictive models, and machine learning does not necessarily lead to increased biological understanding because, sometimes, such machine-learning approaches generate a predictive "black box."

The limited understanding of the fundamental rules in protein production remains a significant challenge for its applications. Problems in synthetic gene design are regularly observed for tuning and optimizing production of biotechnological or medical relevance. These challenges become even more pressing for synthetic biologists trying to construct designer genomes, which require tuning of many synthetic genes simultaneously.

Specific methods have been proposed that can, to some extent, increase the predictability of synthetic gene design. Typically, commercial or academic codon optimization algorithms are used to design ORF regions for heterologous expression, often with limited success, which is not surprising given the current knowledge gaps. However, promising design and randomization approaches have been established regarding the engineering of the highly influential region comprising the 5′ UTRs and the first few codons of an ORF.

Overall, both for the understanding of the fundamental natural principles of gene design and expression and for diverse applications, there remains a need to delve further into the outstanding questions in this field. Despite the impressive recent progress, further refinement of recently launched techniques, as well as the development of new experimental and computational approaches, will be essential to address key questions that have intrigued many biologists for decades.

ACKNOWLEDGMENTS

T.N., M.F.-B., N.J.C., and J.v.d.O. acknowledge the support of the Netherlands Organisation for Scientific Research (NWO) via the Gravitation Project BaSyC (024.003.019) and Spinoza (SPI 93-537), awarded to J.v.d.O. In addition, N.J.C. acknowledges support from his NWO Veni fellowship (VI.Veni.192.156).

AUTHOR CONTRIBUTIONS

T.N., M.F-B., J.v.d.O., and N.J.C. reviewed the literature. T.N., M.F.-B., J.v.d.O, and N.J.C. wrote the review.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Angov, E., Hillier, C.J., Kincaid, R.L., and Lyon, J.A. (2008). Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. PLoS ONE *3*, e2189.

Arango, D., Sturgill, D., Alhusaini, N., Dillman, A.A., Sweet, T.J., Hanson, G., Hosogane, M., Sinclair, W.R., Nanan, K.K., Mandler, M.D., et al. (2018). Acetylation of cytidine in mRNA promotes translation efficiency. Cell *175*, 1872–1886.e24.

Arribere, J.A., Cenik, E.S., Jain, N., Hess, G.T., Lee, C.H., Bassik, M.C., and Fire, A.Z. (2016). Translation readthrough mitigation. Nature *534*, 719–723.

Baez, W.D., Roy, B., McNutt, Z.A., Shatoff, E.A., Chen, S., Bundschuh, R., and Fredrick, K. (2019). Global analysis of protein synthesis in Flavobacterium johnsoniae reveals the use of Kozak-like sequences in diverse bacteria. Nucleic Acids Res. *47*, 10477–10488.

Bazzini, A.A., Del Viso, F., Moreno-Mateos, M.A., Johnstone, T.G., Vejnar, C.E., Qin, Y., Yao, J., Khokha, M.K., and Giraldez, A.J. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. EMBO J. *35*, 2087–2103.

Bhattacharyya, S., Jacobs, W.M., Adkar, B.V., Yan, J., Zhang, W., and Shakhnovich, E.I. (2018). Accessibility of the Shine-Dalgarno sequence dictates N-terminal codon bias in *E. coli*. Mol. Cell *70*, 894–905.e5.

Boël, G., Letso, R., Neely, H., Price, W.N., Wong, K.H., Su, M., Luff, J., Valecha, M., Everett, J.K., Acton, T.B., et al. (2016). Codon influence on protein expression in *E. coli* correlates with mRNA levels. Nature *529*, 358–363.

Bonde, M.T., Pedersen, M., Klausen, M.S., Jensen, S.I., Wulff, T., Harrison, S., Nielsen, A.T., Herrgård, M.J., and Sommer, M.O.A. (2016). Predictable tuning of protein expression in bacteria. Nat. Methods *13*, 233–236.

Bourret, J., Alizon, S., and Bravo, I.G. (2019). COUSIN (COdon Usage Similarity INdex): a normalized measure of codon usage preferences. Genome Biol. Evol. *11*, 3523–3528.

Brophy, J.A., and Voigt, C.A. (2016). Antisense transcription as a tool to tune gene expression. Mol. Syst. Biol. *12*, 854.

Buchan, J.R., Aucott, L.S., and Stansfield, I. (2006). tRNA properties help shape codon pair preferences in open reading frames. Nucleic Acids Res. *34*, 1015–1027.

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., Rodnina, M.V., and Komar, A.A. (2016). Synonymous codons direct cotranslational folding toward different protein conformations. Mol. Cell *61*, 341–351.

Burkhardt, D.H., Rouskin, S., Zhang, Y., Li, G.W., Weissman, J.S., and Gross, C.A. (2017). Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. eLife *6*, e22037.

Burow, D.A., Martin, S., Quail, J.F., Alhusaini, N., Coller, J., and Cleary, M.D. (2018). Attenuated codon optimality contributes to neural-specific mRNA decay in *Drosophila*. Cell Rep. *24*, 1704–1712.

Buschauer, R., Matsuo, Y., Sugiyama, T., Chen, Y.-H., Alhusaini, N., Sweet, T., Ikeuchi, K., Cheng, J., Matsuki, Y., Nobuta, R., et al. (2020). The Ccr4-Not complex monitors the translating ribosome for codon optimality. Science *368*, eaay6912.

Cambray, G., Guimaraes, J.C., and Arkin, A.P. (2018). Evaluation of 244,000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. Nat. Biotechnol. *36*, 1005–1015.

Chaney, J.L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A.T., Ngo, K., Li, J., Emrich, S., and Clark, P.L. (2017). Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput. Biol. *13*, e1005531.

Charneski, C.A., and Hurst, L.D. (2013). Positively charged residues are the major determinants of ribosomal velocity. PLoS Biol. *11*, e1001508.

Chekulaeva, M., and Landthaler, M. (2016). Eyes on Translation. Mol. Cell *63*, 918–925.

Cheng, J.K., Morse, N.J., Wagner, J.M., Tucker, S.K., and Alper, H.S. (2019). Design and evaluation of synthetic terminators for regulating mammalian cell transgene expression. ACS Synth. Biol. *8*, 1263–1275.

Choi, J., Ieong, K.W., Demirci, H., Chen, J., Petrov, A., Prabhakar, A., O'Leary, S.E., Dominissini, D., Rechavi, G., Soltis, S.M., et al. (2016). N^6-methyladenosine in mRNA disrupts tRNA selection and translation-elongation dynamics. Nat. Struct. Mol. Biol. *23*, 110–115.

Choi, J., Indrisiunaite, G., DeMirci, H., Ieong, K.W., Wang, J., Petrov, A., Prabhakar, A., Rechavi, G., Dominissini, D., He, C., et al. (2018). 2′-O-methylation in mRNA disrupts tRNA decoding during translation elongation. Nat. Struct. Mol. Biol. *25*, 208–216.

Chou, H.-J., Donnard, E., Gustafsson, H.T., Garber, M., and Rando, O.J. (2017). Transcriptome-wide Analysis of roles for tRNA modifications in translational regulation. Mol. Cell *68*, 978–992.e4.

Chu, D., Kazana, E., Bellanger, N., Singh, T., Tuite, M.F., and von der Haar, T. (2014). Translation elongation can control translation initiation on eukaryotic mRNAs. EMBO J. *33*, 21–34.

Claassens, N.J., Siliakus, M.F., Spaans, S.K., Creutzburg, S.C.A., Nijsse, B., Schaap, P.J., Quax, T.E.F., and van der Oost, J. (2017). Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS ONE *12*, e0184355.

Claassens, N.J., Finger-Bou, M., Scholten, B., Muis, F., de Groot, J.J., de Gier, J.W., de Vos, W.M., and van der Oost, J. (2019). Bicistronic design-based continuous and high-level membrane protein production in *Escherichia coli*. ACS Synth. Biol. *8*, 1685–1690.

Costello, A., Lao, N.T., Barron, N., and Clynes, M. (2020). Reinventing the wheel: synthetic circular RNAs for mammalian cell engineering. Trends Biotechnol. *38*, 217–230.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Cuperus, J.T., Groves, B., Kuchina, A., Rosenberg, A.B., Jojic, N., Fields, S., and Seelig, G. (2017). Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. Genome Res. *27*, 2015–2024.

Curran, K.A., Morse, N.J., Markham, K.A., Wagman, A.M., Gupta, A., and Alper, H.S. (2015). Short synthetic terminators for improved heterologous gene expression in yeast. ACS Synth. Biol. *4*, 824–832.

de Boer, C.G., Vaishnav, E.D., Sadeh, R., Abeyta, E.L., Friedman, N., and Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. Nat. Biotechnol. *38*, 56–65.

de Freitas Nascimento, J., Kelly, S., Sunter, J., and Carrington, M. (2018). Codon choice directs constitutive mRNA levels in trypanosomes. eLife *7*, 1–26.

de Jongh, R.P.H., van Dijk, A.D.J., Julsing, M.K., Schaap, P.J., and de Ridder, D. (2020). Designing eukaryotic gene expression regulation using machine learning. Trends Biotechnol. *38*, 191–201.

De Nijs, Y., De Maeseneire, S.L., and Soetaert, W.K. (2020). 5′ Untranslated regions: the next regulatory sequence in yeast synthetic biology. Biol. Rev. Camb. Philos. Soc. *95*, 517–529.

Decoene, T., Peters, G., De Maeseneire, S., and De Mey, M. (2018). Toward predictable 5′UTRs in *Saccharomyces cerevisiae*: development of a yUTR calculator. ACS Synth. Biol. *7*, 622–634.

Ding, W., Cheng, J., Guo, D., Mao, L., Li, J., Lu, L., Zhang, Y., Yang, J., and Jiang, H. (2018). Engineering the 5′ UTR-Mediated regulation of protein abundance in yeast using nucleotide sequence activity relationships. ACS Synth. Biol. *7*, 2709–2714.

dos Reis, M., Wernisch, L., and Savva, R. (2003). Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res. *31*, 6976–6985.

Drummond, D.A., and Wilke, C.O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell *134*, 341–352.

Espah Borujeni, A., and Salis, H.M. (2016). Translation initiation is controlled by RNA folding kinetics via a ribosome drafting mechanism. J. Am. Chem. Soc. *138*, 7016–7023.

Espah Borujeni, A., Cetnar, D., Farasat, I., Smith, A., Lundgren, N., and Salis, H.M. (2017). Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. Nucleic Acids Res. *45*, 5437–5448.

Fleming, I., and Cavalcanti, A.R.O. (2019). Selection for tandem stop codons in ciliate species with reassigned stop codons. PLoS ONE *14*, e0225804.

Forrest, M.E., Pinkard, O., Martin, S., Sweet, T.J., Hanson, G., and Coller, J. (2020). Codon and amino acid content are associated with mRNA stability in mammalian cells. PLoS ONE *15*, e0228730.

Frumkin, I., Schirman, D., Rotman, A., Li, F., Zahavi, L., Mordret, E., Asraf, O., Wu, S., Levy, S.F., and Pilpel, Y. (2017). Gene architectures that minimize cost of gene expression. Mol. Cell *65*, 142–153.

Fu, J., Murphy, K.A., Zhou, M., Li, Y.H., Lam, V.H., Tabuloc, C.A., Chiu, J.C., and Liu, Y. (2016). Codon usage affects the structure and function of the *Drosophila* circadian clock protein PERIOD. Genes Dev. *30*, 1761–1775.

Fu, J., Dang, Y., Counter, C., and Liu, Y. (2018). Codon usage regulates human KRAS expression at both transcriptional and translational levels. J. Biol. Chem. *293*, 17929–17940.

Gamble, C.E., Brule, C.E., Dean, K.M., Fields, S., and Grayhack, E.J. (2016). Adjacent codons act in concert to modulate translation efficiency in yeast. Cell *166*, 679–690.

Gardin, J., Yeasmin, R., Yurovsky, A., Cai, Y., Skiena, S., and Futcher, B. (2014). Measurement of average decoding rates of the 61 sense codons in vivo. eLife *3*, 1–20.

Goodman, D.B., Church, G.M., and Kosuri, S. (2013). Causes and effects of N-terminal codon bias in bacterial genes. Science *342*, 475–479.

Gould, N., Hendy, O., and Papamichail, D. (2014). Computational tools and algorithms for designing customized synthetic genes. Front. Bioeng. Biotechnol. *2*, 41.

Groenke, N., Trimpert, J., Merz, S., Conradie, A.M., Wyler, E., Zhang, H., Hazapis, O.G., Rausch, S., Landthaler, M., Osterrieder, N., and Kunec, D. (2020). Mechanism of virus attenuation by codon pair deoptimization. Cell Rep. *31*, 107586.

Gutman, G.A., and Hatfield, G.W. (1989). Nonrandom utilization of codon pairs in Escherichia coli. Proc. Natl. Acad. Sci. USA *86*, 3699–3703.

Hanson, G., and Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. Nat. Rev. Mol. Cell Biol. *19*, 20–30.

Harigaya, Y., and Parker, R. (2016). Analysis of the association between codon optimality and mRNA stability in Schizosaccharomyces pombe. BMC Genomics *17*, 895.

Henderson, K.L., Felth, L.C., Molzahn, C.M., Shkel, I., Wang, S., Chhabra, M., Ruff, E.F., Bieter, L., Kraft, J.E., and Record, M.T., Jr. (2017). Mechanism of transcription initiation and promoter escape by *E. coli* RNA polymerase. Proc. Natl. Acad. Sci. USA *114*, E3032–E3040.

Hess, A.K., Saffert, P., Liebeton, K., and Ignatova, Z. (2015). Optimization of translation profiles enhances protein expression and solubility. PLoS ONE *10*, e0127039.

Hia, F., Yang, S.F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., Fukao, A., Fujiwara, T., Landthaler, M., Natsume, T., et al. (2019). Codon bias confers stability to human mRNAs. EMBO Rep. *20*, e48220.

Hockenberry, A.J., Jewett, M.C., Amaral, L.A.N., and Wilke, C.O. (2018). Within-gene shine-dalgarno sequences are not selected for function. Mol. Biol. Evol. *35*, 2487–2498.

Höllerer, S., Papaxanthos, L., Gumpinger, A.C., Fischer, K., Beisel, C., Borgwardt, K., Benenson, Y., and Jeschek, M. (2020). Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. Nat. Commun. *11*, 3551.

Huter, P., Arenz, S., Bock, L.V., Graf, M., Frister, J.O., Heuer, A., Peil, L., Starosta, A.L., Wohlgemuth, I., Peske, F., et al. (2017). Structural Basis for Polyproline-Mediated Ribosome Stalling and Rescue by the Translation Elongation Factor EF-P. Mol. Cell *68*, 515–527.e6.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. *2*, 13–34.

Ingolia, N.T., Ghaemmaghami, S., Newman, J.R.S., and Weissman, J.S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science *324*, 218–223.

Jeacock, L., Faria, J., and Horn, D. (2018). Codon usage bias controls mRNA and protein abundance in trypanosomatids. eLife *7*, 1–20.

Jeschek, M., Gerngross, D., and Panke, S. (2016). Rationally reduced libraries for combinatorial pathway optimization minimizing experimental effort. Nat. Commun. *7*, 11163.

Johnson, G.E., Lalanne, J.B., Peters, M.L., and Li, G.W. (2020). Functionally uncoupled transcription-translation in Bacillus subtilis. Nature *585*, 124–128.

Kelsic, E.D., Chung, H., Cohen, N., Park, J., Wang, H.H., and Kishony, R. (2016). RNA Structural Determinants of Optimal Codons Revealed by MAGE-Seq. Cell Syst. *3*, 563–571.e6.

Ki, M.R., and Pack, S.P. (2020). Fusion tags to enhance heterologous protein expression. Appl. Microbiol. Biotechnol. *104*, 2411–2425.

Kim, S.J., Yoon, J.S., Shishido, H., Yang, Z., Rooney, L.A.A., Barral, J.M., and Skach, W.R. (2015). Protein folding: translational tuning optimizes nascent protein folding in cells. Science *348*, 444–448.

Kimura, S., Srisuknimit, V., and Waldor, M.K. (2020). Probing the diversity and regulation of tRNA modifications. Curr. Opin. Microbiol. *57*, 41–48.

Komarova, E.S., Chervontseva, Z.S., Osterman, I.A., Evfratov, S.A., Rubtsova, M.P., Zatsepin, T.S., Semashko, T.A., Kostryukova, E.S., Bogdanov, A.A., Gelfand, M.S., et al. (2020). Influence of the spacer region between the Shine-Dalgarno box and the start codon for fine-tuning of the translation efficiency in Escherichia coli. Microb. Biotechnol. *13*, 1254–1261.

Kozak, M. (1981). Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res. *9*, 5233–5252.

Kudla, G., Murray, A.W., Tollervey, D., and Plotkin, J.B. (2009). Coding-sequence determinants of expression in *Escherichia coli*. Science *324*, 255–258.

Kunec, D., and Osterrieder, N. (2016). Codon pair bias is a direct consequence of dinucleotide bias. Cell Rep. *14*, 55–67.

Lahtvee, P.J., Sánchez, B.J., Smialowska, A., Kasvandik, S., Elsemman, I.E., Gatto, F., and Nielsen, J. (2017). Absolute quantification of protein and mrna abundances demonstrate variability in gene-specific translation efficiency in yeast. Cell Syst. *4*, 495–504.e5.

Lee, J., and Borukhov, S. (2016). Bacterial RNA polymerase-DNA interaction—the driving force of gene expression and the target for drug action. Front. Mol. Biosci. *3*, 73.

Lenstra, T.L., Rodriguez, J., Chen, H., and Larson, D.R. (2016). Transcription dynamics in living cells. Annu. Rev. Biophys. *45*, 25–47.

Leppek, K., Das, R., and Barna, M. (2018). Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. Nat. Rev. Mol. Cell Biol. *19*, 158–174.

Levo, M., Avnit-Sagi, T., Lotan-Pompan, M., Kalma, Y., Weinberger, A., Yakhini, Z., and Segal, E. (2017). Systematic investigation of transcription factor activity in the context of chromatin using massively parallel binding and expression assays. Mol. Cell *65*, 604–617.e6.

Li, G.W., Oh, E., and Weissman, J.S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. Nature *484*, 538–541.

Li, G.W., Burkhardt, D., Gross, C., and Weissman, J.S. (2014). Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. Cell *157*, 624–635.

Menendez-Gil, P., Caballero, C.J., Catalan-Moreno, A., Irurzun, N., Barrio-Hernandez, I., Caldelari, I., and Toledo-Arana, A. (2020). Differential evolution in 3’UTRs leads to specific gene expression in Staphylococcus. Nucleic Acids Res. *48*, 2544–2563.

Mishima, Y., and Tomari, Y. (2016). Codon usage and 3′ UTR length determine maternal mRNA stability in zebrafish. Mol. Cell *61*, 874–885.

Mittal, P., Brindle, J., Stephen, J., Plotkin, J.B., and Kudla, G. (2018). Codon usage influences fitness through RNA toxicity. Proc. Natl. Acad. Sci. USA *115*, 8639–8644.

Mohammad, F., Woolstenhulme, C.J., Green, R., and Buskirk, A.R. (2016). Clarifying the translational pausing landscape in bacteria by ribosome profiling. Cell Rep. *14*, 686–694.

Mohanty, B.K., and Kushner, S.R. (2016). Regulation of mRNA decay in bacteria. Annu. Rev. Microbiol. *70*, 25–44.

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G.D., Cox, J., Geiger, T., Lindner, A.B., and Pilpel, Y. (2019). Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. Mol. Cell *75*, 427–441.e5.

Mugridge, J.S., Coller, J., and Gross, J.D. (2018). Structural and molecular mechanisms for the control of eukaryotic 5′-3′ mRNA decay. Nat. Struct. Mol. Biol. *25*, 1077–1085.

Mustoe, A.M., Busan, S., Rice, G.M., Hajdin, C.E., Peterson, B.K., Ruda, V.M., Kubica, N., Nutiu, R., Baryza, J.L., and Weeks, K.M. (2018). Pervasive regulatory functions of mRNA structure revealed by high-resolution SHAPE probing. Cell *173*, 181–195.e18.

Mutalik, V.K., Guimaraes, J.C., Cambray, G., Lam, C., Christoffersen, M.J., Mai, Q.A., Tran, A.B., Paull, M., Keasling, J.D., Arkin, A.P., and Endy, D. (2013). Precise and reliable gene expression via standard transcription and translation initiation elements. Nat. Methods *10*, 354–360.

Narula, A., Ellis, J., Taliaferro, J.M., and Rissland, O.S. (2019). Coding regions affect mRNA stability in human cells. RNA *25*, 1751–1764.

Navon, S.P., Kornberg, G., Chen, J., Schwartzman, T., Tsai, A., Puglisi, E.V., Puglisi, J.D., and Adir, N. (2016). Amino acid sequence repertoire of the bacterial proteome and the occurrence of untranslatable sequences. Proc. Natl. Acad. Sci. USA *113*, 7166–7170.

Nedialkova, D.D., and Leidel, S.A. (2015). Optimization of codon translation rates via tRNA Modifications Maintains Proteome Integrity. Cell *161*, 1606–1618.

Newman, Z.R., Young, J.M., Ingolia, N.T., and Barton, G.M. (2016). Differences in codon bias and GC content contribute to the balanced expression of TLR7 and TLR9. Proc. Natl. Acad. Sci. USA *113*, E1362–E1371.

Nieuwkoop, T., Claassens, N.J., and van der Oost, J. (2019). Improved protein production and codon optimization analyses in Escherichia coli by bicistronic design. Microb. Biotechnol. *12*, 173–179.

O’Reilly, F.J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., Blötz, C., Singh, N., Hagen, W.J.H., Cramer, P., et al. (2020). In-cell architecture of an actively transcribing-translating expressome. Science *369*, 554–557.

Parret, A.H., Besir, H., and Meijers, R. (2016). Critical reflections on synthetic gene design for recombinant protein expression. Curr. Opin. Struct. Biol. *38*, 155–162.

Pechmann, S., and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nat. Struct. Mol. Biol. *20*, 237–243.

Perriman, R., and Ares, M., Jr. (1998). Circular mRNA can direct translation of extremely long repeating-sequence proteins in vivo. RNA *4*, 1047–1054.

Petersen, S.D., Zhang, J., Lee, J.S., Jakociunas, T., Grav, L.M., Kildegaard, H.F., Keasling, J.D., and Jensen, M.K. (2018). Modular 5′-UTR hexamers for context-independent tuning of protein expression in eukaryotes. Nucleic Acids Res. *46*, e127.

Presnyak, V., Alhusaini, N., Chen, Y.H., Martin, S., Morris, N., Kline, N., Olson, S., Weinberg, D., Baker, K.E., Graveley, B.R., and Coller, J. (2015). Codon optimality is a major determinant of mRNA stability. Cell *160*, 1111–1124.

Quax, T.E.F., Wolf, Y.I., Koehorst, J.J., Wurtzel, O., van der Oost, R., Ran, W., Blombach, F., Makarova, K.S., Brouns, S.J., Forster, A.C., et al. (2013). Differential translation tunes uneven production of operon-encoded proteins. Cell Rep. *4*, 938–944.

Quax, T.E.F., Claassens, N.J., Söll, D., and van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. Mol. Cell *59*, 149–161.

Radhakrishnan, A., Chen, Y.H., Martin, S., Alhusaini, N., Green, R., and Coller, J. (2016). The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. Cell *167*, 122–132.e9.

Rennig, M., Martinez, V., Mirzadeh, K., Dunas, F., Röjsäter, B., Daley, D.O., and Nørholm, M.H.H. (2018). TARSyn: tunable antibiotic resistance devices enabling bacterial synthetic evolution and protein production. ACS Synth. Biol. *7*, 432–442.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., and Weissman, J.S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. Nature *505*, 701–705.

Saito, K., Green, R., and Buskirk, A.R. (2020). Translational initiation in *E. coli* occurs at the correct sites genome-wide in the absence of mRNA-rRNA base-pairing. eLife *9*, 1–19.

Salis, H.M., Mirsky, E.A., and Voigt, C.A. (2009). Automated design of synthetic ribosome binding sites to control protein expression. Nat. Biotechnol. *27*, 946–950.

Schmid, M., and Jensen, T.H. (2018). Controlling nuclear RNA levels. Nat. Rev. Genet. *19*, 518–529.

Sharp, P.M., and Li, W.H. (1987). The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. *15*, 1281–1295.

Shine, J., and Dalgarno, L. (1974). The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. Proc. Natl. Acad. Sci. USA *71*, 1342–1346.

Siegfried, N.A., Busan, S., Rice, G.M., Nelson, J.A.E., and Weeks, K.M. (2014). RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat. Methods *11*, 959–965.

Sun, M., Gao, X., Zhao, Z., Li, A., Wang, Y., Yang, Y., Liu, X., and Bai, Z. (2020). Enhanced production of recombinant proteins in *Corynebacterium glutamicum* by constructing a bicistronic gene expression system. Microb. Cell Fact. *19*, 113.

Takyar, S., Hickerson, R.P., and Noller, H.F. (2005). mRNA helicase activity of the ribosome. Cell *120*, 49–58.

Taniguchi, Y., Choi, P.J., Li, G.-W., Chen, H., Babu, M., Hearn, J., Emili, A., and Xie, X.S. (2010). Quantifying *E. coli* proteome and transcriptome with single-molecule sensitivity in single cells. Science *329*, 533–538.

Tesina, P., Lessen, L.N., Buschauer, R., Cheng, J., Wu, C.C., Berninghausen, O., Buskirk, A.R., Becker, T., Beckmann, R., and Green, R. (2020). Molecular mechanism of translational stalling by inhibitory codon combinations and poly(A) tracts. EMBO J. *39*, e103365.

Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30.

Traverse, C.C., and Ochman, H. (2016). Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. Proc. Natl. Acad. Sci. USA *113*, 3311–3316.

Tuller, T., and Zur, H. (2015). Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Res. *43*, 13–28.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpel, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell *141*, 344–354.

Urtecho, G., Tripp, A.D., Insigne, K.D., Kim, H., and Kosuri, S. (2019). Systematic dissection of sequence elements controlling σ70 promoters using a genomically encoded multiplexed reporter assay in *Escherichia coli*. Biochemistry *58*, 1539–1551.

Urtecho, G., Insigne, K., Tripp, A.D., Brinck, M., Lubock, N.B., Kim, H., Chan, T., and Kosuri, S. (2020). Genome-wide functional characterization of *Escherichia coli* promoters and regulatory elements responsible for their function. BioRxiv, 2020.01.04.894907.

Vasquez, K.A., Hatridge, T.A., Curtis, N.C., and Contreras, L.M. (2016). Slowing translation between protein domains by increasing affinity between mRNAs and the ribosomal anti-Shine-Dalgarno sequence improves solubility. ACS Synth. Biol. *5*, 133–145.

Verma, M., Choi, J., Cottrell, K.A., Lavagnino, Z., Thomas, E.N., Pavlovic-Djuranovic, S., Szczesny, P., Piston, D.W., Zaher, H.S., Puglisi, J.D., and Djuranovic, S. (2019). A short translational ramp determines the efficiency of protein synthesis. Nat. Commun. *10*, 5774.

Wade, J.T., and Struhl, K. (2008). The transition from transcriptional initiation to elongation. Curr. Opin. Genet. Dev. *18*, 130–136.

Wan, J., Gao, X., Mao, Y., Zhang, X., and Qian, S.-B. (2018). A coding sequence-embedded principle governs translational reading frame fidelity. Research (Wash. D.C.) *2018*, 7089174.

Webster, M.W., Chen, Y.H., Stowell, J.A.W., Alhusaini, N., Sweet, T., Graveley, B.R., Coller, J., and Passmore, L.A. (2018). mRNA deadenylation is coupled to translation rates by the differential activities of Ccr4-not nucleases. Mol. Cell *70*, 1089–1100.e8.

Weenink, T., van der Hilst, J., McKiernan, R.M., and Ellis, T. (2018). Design of RNA hairpin modules that predictably tune translation in yeast. Synth. Biol. *3*, ysy019.

Weinberg, D.E., Shah, P., Eichhorn, S.W., Hussmann, J.A., Plotkin, J.B., and Bartel, D.P. (2016). Improved ribosome-footprint and mRNA measurements provide insights into dynamics and regulation of yeast translation. Cell Rep. *14*, 1787–1799.

Wesselhoeft, R.A., Kowalski, P.S., and Anderson, D.G. (2018). Engineering circular RNA for potent and stable translation in eukaryotic cells. Nat. Commun. *9*, 2629.

Wilson, D.N., Arenz, S., and Beckmann, R. (2016). Translation regulation via nascent polypeptide-mediated ribosome stalling. Curr. Opin. Struct. Biol. *37*, 123–133.

Winkelman, J.T., Vvedenskaya, I.O., Zhang, Y., Zhang, Y., Bird, J.G., Taylor, D.M., Gourse, R.L., Ebright, R.H., and Nickels, B.E. (2016). Multiplexed protein-DNA cross-linking: scrunching in transcription start site selection. Science *351*, 1090–1093.

Wu, Q., Medina, S.G., Kushawah, G., DeVore, M.L., Castellano, L.A., Hand, J.M., Wright, M., and Bazzini, A.A. (2019). Translation affects mRNA stability in a codon-dependent manner in human cells. eLife *8*, 1–22.

Yan, X., Hoek, T.A., Vale, R.D., and Tanenbaum, M.E. (2016). Dynamics of translation of single mRNA molecules *in vivo*. Cell *165*, 976–989.

Yang, Q., Yu, C.-H., Zhao, F., Dang, Y., Wu, C., Xie, P., Sachs, M.S., and Liu, Y. (2019). eRF1 mediates codon usage effects on mRNA translation efficiency through premature termination at rare codons. Nucleic Acids Res. *47*, 9243–9258.

Yona, A.H., Alm, E.J., and Gore, J. (2018). Random sequences rapidly evolve into de novo promoters. Nat. Commun. *9*, 1530.

Yu, C.H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M.S., and Liu, Y. (2015). Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. Mol. Cell *59*, 744–754.

Zhao, B.S., Roundtree, I.A., and He, C. (2017a). Post-transcriptional gene regulation by mRNA modifications. Nat. Rev. Mol. Cell Biol. *18*, 31–42.

Zhao, F., Yu, C.H., and Liu, Y. (2017b). Codon usage regulates protein structure and function by affecting translation elongation speed in *Drosophila* cells. Nucleic Acids Res. *45*, 8484–8492.

Zhao, J.P., Zhu, H., Guo, X.P., and Sun, Y.C. (2018). AU-rich long 3′ untranslated region regulates gene expression in bacteria. Front. Microbiol. *9*, 3080.

Zhou, M., Guo, J., Cha, J., Chae, M., Chen, S., Barral, J.M., Sachs, M.S., and Liu, Y. (2013). Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature *495*, 111–115.

Zhou, M., Wang, T., Fu, J., Xiao, G., and Liu, Y. (2015). Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Mol. Microbiol. *97*, 974–987.

Zhou, Z., Dang, Y., Zhou, M., Li, L., Yu, C.H., Fu, J., Chen, S., and Liu, Y. (2016). Codon usage is an important determinant of gene expression levels largely through its effects on transcription. Proc. Natl. Acad. Sci. USA *113*, E6117–E6125.