Check for updates

Codon-optimization in gene therapy: promises, prospects and challenges

Anastasiia Iu Paremskaia[1], Anna A. Kogan[1],
Anastasiia Murashkina[1], Daria A. Naumova[1], Anakha Satish[1],
Ivan S. Abramov[1,2], Sofya G. Feoktistova[1], Olga N. Mityaeva[1],
Andrei A. Deviatkin[1*†] and Pavel Yu Volchkov[1,2*†]

[1]Federal Research Center for Innovator and Emerging Biomedical and Pharmaceutical Technologies,
Moscow, Russia, [2]The MCSC named after A. S. Loginov, Moscow, Russia

Codon optimization has evolved to enhance protein expression efficiency by exploiting the genetic code's redundancy, allowing for multiple codon options for a single amino acid. Initially observed in *E. coli*, optimal codon usage correlates with high gene expression, which has propelled applications expanding from basic research to biopharmaceuticals and vaccine development. The method is especially valuable for adjusting immune responses in gene therapies and has the potenial to create tissue-specific therapies. However, challenges persist, such as the risk of unintended effects on protein function and the complexity of evaluating optimization effectiveness. Despite these issues, codon optimization is crucial in advancing gene therapeutics. This study provides a comprehensive review of the current metrics for codon-optimization, and its practical usage in research and clinical applications, in the context of gene therapy.
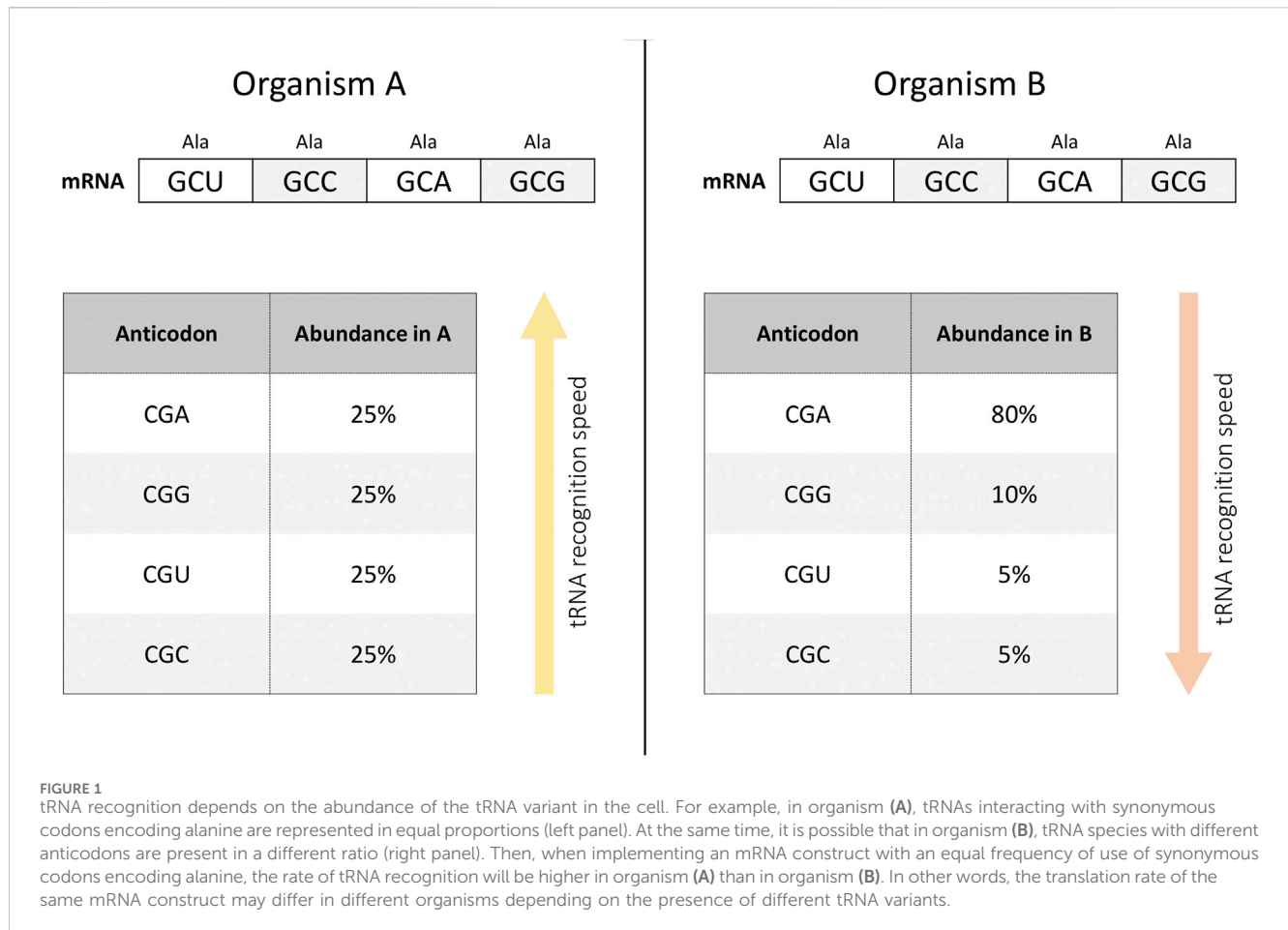
1 Introduction

Codon optimization first appeared due to the search for an approach to increase the efficiency of expression of target proteins in bacterial cultures. The known property of degeneracy of the genetic code allows mRNA to encode the same proteins in different ways since 20 proteinogenic amino acids can be encoded by 61 codons (Welch et al., 2009). This property formed the basis of the codon optimization method, when, with the advent of genetic sequencing, it became evident that the usage of codons is non-random. Bias in codon usage occurs between different organisms, tissues, and sometimes even between parts of the same gene (Athey et al., 2017; Pouyet et al., 2017). Thus, it became clear that the selection of the most common codons deemed suitable for an organism or cell line during genetic engineering research allows significantly changing approaches to conducting experiments.

Escherichia coli was the first organism with an analyzed codon usage system. Knowing the sequences of anticodons and the abundance of various tRNAs in the cell, the authors identified criteria for codon optimality (Ikemura, 1981). The first criterion was high codon recognition, the second was the highest abundance of tRNA. Highly expressed genes had a bias in frequency of use towards optimal codons, while genes with low expression were characterized by high randomness in the choice of codons (Gouy and Gautier, 1982).

FIGURE 1
tRNA recognition depends on the abundance of the tRNA variant in the cell. For example, in organism **(A)**, tRNAs interacting with synonymous codons encoding alanine are represented in equal proportions (left panel). At the same time, it is possible that in organism **(B)**, tRNA species with different anticodons are present in a different ratio (right panel). Then, when implementing an mRNA construct with an equal frequency of use of synonymous codons encoding alanine, the rate of tRNA recognition will be higher in organism **(A)** than in organism **(B)**. In other words, the translation rate of the same mRNA construct may differ in different organisms depending on the presence of different tRNA variants.

Currently, codon optimization has found application in a wide range of topics. In addition to fundamental research, control of the efficiency of protein expression through the selection of synonymous codons is also used for the development and production of biotherapies (Ayyar et al., 2017), most of which are based on the expression of recombinant proteins. The method has become indispensable for molecular pharming on plants, where the problem of low expression efficiency is most pressing (Perlak et al., 1991; Desai et al., 2010; Thomas and Walmsley, 2014).

Differentiated cells determine the formation of tissues of various types. This complicated process can be modulated at the cellular and molecular level (Simon et al., 2018). At the molecular level, this diversity is reflected in particular in differences in protein expression - proteins that are abundant in one tissue may be absent in another (Thul and Lindskog, 2018). Differences in protein abundance are, in turn, caused by differences in RNA expression. One of the possible factors affecting such patterns is the different frequency of use of synonymous codons encoding the same amino acid during translation (Kames et al., 2020) (Figure 1). Indeed, either the rarity of codon usage (Plotkin et al., 2004) or the frequency of tRNA variants (Dittmar et al., 2006; Gao et al., 2022) both vary between tissues. This can potentially be exploited for the construction of tissue-specific gene therapy. At the same time, to our knowledge, there is currently only one paper in peer-reviewed journals that has experimentally tested this hypothesis (Hernandez-

Alias et al., 2023). This study is evidence that tissue-specific codon usage can potentially be used to design tissue-specific transgenes. At the same time, this metric is only one additional tool in the gene design toolbox whose implementation needs to be further explored and cannot be considered in isolation from several other indicators discussed below (Hernandez-Alias et al., 2023).

One of the most relevant and important areas of codon optimization application is the development of vaccines. The current way to create non-live vaccines is the use of attenuated viruses. Several research groups have experimented with attenuating poliovirus by changing codon bias in the gene encoding the poliovirus capsid protein, which involved replacing more frequent codons with less frequent ones (Burns et al., 2006; Mueller et al., 2006). Moreover, increasing transgene expression in vaccines may improve the effectiveness of immunization and can be achieved through codon optimization (Chen et al., 2008; Bell et al., 2016). In addition, a new class of vaccines—mRNA vaccines—has recently been introduced into clinical practice in the context of the COVID-19 pandemic (Oliver et al., 2020). Currently, the possibility of a similar approach for the prevention of infectious diseases such as rabies (Wan et al., 2023), influenza virus (Lee et al., 2023), Zika virus (Bollman et al., 2023), Lassa virus (Ronk et al., 2023) is the subject of active research and development. Remarkably, codon optimization of mRNA vaccines can significantly improve their stability and immunogenicity (Zhang et al., 2023). Despite the benefits of codon optimization, it is

important to maintain a balance in the use of these techniques. Excessive interest in codon optimization can possibly lead to the accumulation of substances that are poorly excreted from the body, such as, for example, modified mRNA and the corresponding antigen (Bansal et al., 2021; Röltgen et al., 2022).

Currently, various approaches could be used for the development of gene therapeutics. Control of the immunogenicity of the administered drug is one of the most vital tasks not only in the preparation of vaccines but also for gene therapies. For the drug to work effectively, it is necessary to reduce the viral vector's immunogenicity. It has been shown that by varying synonymous codons in the transgene and vector, it is possible to increase the effectiveness of therapy by lowering immunogenicity (Athanasopoulos et al., 2011; Bell et al., 2016), which provides optimism for simplifying vector selection and expanding the application of this type of therapy.

Regrettably, codon optimization techniques, while widely employed in the development of gene therapies, are far from perfect and are fraught with several challenges. One prominent issue lies in the incomplete synonymy of substitutions. This drawback carries the potential to disrupt natural post-transcriptional modification sites or, alternatively, give rise to novel sites, leading to critical alterations in the final protein's structure, properties, and functions (Godfried Sie et al., 2012; Irimia et al., 2012). Furthermore, overlooking the existence of alternative translation initiation sites (Malarkannan et al., 1999; Matsuda and Mauro, 2010)can lead to the unintended production of new proteins, adding another layer of complexity to the process. Beyond these intrinsic challenges, the selection of an appropriate numerical method for evaluating the effectiveness of codon optimization poses an additional obstacle. The abundance of metrics available complicates the task, requiring careful consideration to ensure a meaningful assessment. Despite the above difficulties, codon optimization approaches are actively used in clinical trials around the world and, furthermore, COVID-19 mRNA vaccines Pfizer/BioNTech and Moderna employ codon optimization.

Codon optimization can be carried out in many different ways today. It is often not clear which of these approaches is best suited to fulfill a particular task. The purpose of this review is to cover the current state of this problem and future directions for codon optimization approaches for gene therapies.

2 The quantitative assessment of codon usage and optimization

2.1 Measures of codon usage

The codon usage bias (CUB), also known as codon usage preferences (CUP), is influenced by a combination of factors that vary among species. Such factors include mutation frequency (Pizzo et al., 2015), selection for translation efficiency (Navon and Pilpel, 2011), and the presence of transfer RNA (tRNA) molecules that recognize specific codons (Buchan, 2006; Wei et al., 2019), ribosome binding efficiency (Shi et al., 2020), and translation speed and co-translational protein folding (Mitarai et al., 2008; Liu, 2020).

Based on the non-random usage of codons in the genomes of different species and the previously demonstrated positive correlation between codon bias and gene expression efficiency, Sharp and Li developed the relative synonymous codon usage (RSCU) scale (Sharp and Li, 1986). The RSCU value was calculated for a set of genes as the ratio of the observed codon frequency to the expected frequency, assuming equal usage of synonymous codons. This research has made a substantial contribution to the creation of various metrics, including but not limited to codon adaptation index (CAI) (Sharp and Li, 1987), average ratio of RSCU (ARSCU) (Chamani Mohasses et al., 2020), and genetic tRNA adaptation index (gtAI) (Anwar et al., 2023). CAI continues to be a widely employed metric in both commercial and academic applications. CAI reflects the level of species-specific codon adaptation and is calculated as the geometric mean of RCSU values for each codon in the gene relative to the value of the most frequently used triplet encoding a single amino acid.

To date, numerous metrics for quantitative assessment of the level sequence optimization have been developed. Table 1 offers concise descriptions of commonly used metrics. To give the readers an idea of the frequency of metric usage, we added the citation rate of the original sources. However, it is important to emphasize that this approach does not reflect the level of usage of optimization tools based on the mentioned metrics.

Numerous metrics can be easily calculated with a reference set of genes to obtain the codon usage frequency. For example, Fop is calculated as the ratio of optimal codons to the total number of codons, excluding stop codons and codons without alternatives for amino acids (methionine, tryptophan) (Ikemura, 1981; 1982). The index aids in gauging the prevalence of synonymous codon usage. Other metrics are grounded in the assumption that the usage of codons is non-random. The metrics quantify the difference in codon usage frequency from a uniform distribution within the coding sequence. When all codon variants for a specific amino acid are utilized with equal frequency, such difference is minimal. Conversely, the maximum is achieved when only one codon out of the possible ones is utilized. Examples of such indices include ENC, CDC, SCUO, and others.

2.2 Codon adaptation metrics for assessing mRNA properties

Codon optimization is a strategy aimed at increasing the efficiency of mRNA translation and overcoming protein expression limitations. The use of synonymous codons affects the stability of mRNA in human cells (Narula et al., 2019; Wu et al., 2019). The thermodynamic stability of mRNA within a cell significantly influences translation efficiency (Hanson and Coller, 2018; Diez et al., 2022). mRNA is inherently unstable and can undergo transient states and adopt multiple stable structures. One approach to selecting synonymous amino acids for the purpose of thermodynamic stabilization is aimed at minimizing the free energy ΔG (MFE) released during RNA folding (Zuker and Stiegler, 1981; Zuker, 1994). Ringner and Krogh demonstrated in *Saccharomyces cerevisiae* that the folding free energy in the vicinity of the 5′-UTR correlates positively with transcription efficiency and mRNA half-life (Ringnér and Krogh, 2005).

TABLE 1 Metrics for codon optimization with formal definition and description. The number of citations was retrieved from the Scopus database.

Index	Mathematical formula	Principle	Interpretation	Original source	Citation count
Frequency of optimal codons (Fop)	Fop = Nopt/Ntotal, where Nopt denotes number of optimal codons in a gene, Ntotal indicates number of total codons in sequence	The rationale behind Fop is that genes with a higher expression level tend to preferentially use certain codons, and this bias can be quantified using this ratio	Representing the occurrence of the optimal codon within a gene sequence. Values close to 1 indicate a strong bias toward optimal codons, while values closer to 0 suggest a more even or random usage of synonymous codons	Ikemura, 1981 (1982)	1724 and 768
Relative synonymous codon usage (RSCU)	$RSCU = \frac{Xij}{\frac{1}{ni}\sum_{j=1}^{ni}(Xij)}$ where xij denotes the number of occurrences of the j-th codon for the i-th amino acid, ni represents the degeneracy for the i-th amino acid	Calculated as the ratio of the observed codon frequency to the expected frequency assuming equal usage of synonymous codons	The codon with RSCU of 1 value indicates average (random or equally) synonymous codon usage; RSCU value greater than 1 indicates positive codon usage bias (overrepresented); RSCU value less than 1 indicates negative codon usage bias (underrepresented)	Sharp and Li (1986)	653
Codon adaptation index (CAI)	$CAI = exp(\frac{1}{L}\sum_{l=1}^{L} ln W_k)$, where L is the length of a gene measured in codons. Wk is the relative adaptiveness value for the k-th codon in the gene	Quantifies the geometric mean of the RSCU for each codon with respect to the codon usage of a reference set of highly expressed genes. Genes that possess higher scores are anticipated to demonstrate enhanced efficiency in translation and elevated levels of protein expression	A higher CAI score implies that the gene's codon usage is better aligned with the preferred codons identified in a highly expressed reference set. A CAI value of 1 suggests that the codon usage in the gene is perfectly adapted to the preferred codons observed in a highly expressed reference set; value of 0 indicates that the codon usage in the gene is not adapted	Sharp and Li (1987)	4091
Codon pair adaptation index (CPAI)	$CPAI = \prod_{i=1}^{N-1} w_{i,i+1}^{1/(N-1)}$, where N – 1 is the number of codon pairs in gene g and w-i, j is the relative adaptiveness of the codon pair (i, j). $w_{i,j}(G) = \frac{f_{i,j}(G)}{f_{max(i,j)}(G)}$, where fi, j is the frequency of codon pair (i, j) and $f(max(i,j))$ is the frequency of the codon pair most often used to code for the amino acid pair $(aa(i),aa(j))$ in a set of highly expressed genes G	The advantage of the index is the automatic weight selection algorithm Carbone et al., 2003. Combined use of CPAI and CAI have been shown to better predict gene expression	Similar to CAI.	Carbone et al. (2003), Friberg et al. (2004)	43 and 402
Codon bias index (CBI)	$CBI = \frac{(Nopt - Nrand))}{(Ntot - Nrand)}$, where $Nopt$ represents the number of preferred optimal codons, $Ntot$ is the total number of codons in the gene, and $Nrand$ denotes the expected number of preferred codons if random codon assignments were made for each amino acid	Codon bias refers to the unequal usage of synonymous codons encoding the same amino acid in a DNA sequence. The codon bias index is a measure that quantifies the extent of this bias	Values range from 0 to 1: 0 indicates random selection of codons, and 1 indicates a significant bias towards preferred codons	Bennetzen and Hall (1982)	1955
Effective number of codons (ENc or Nc)	$ENc = \frac{1}{Faa}$, where $Faa = n\sum_{i=1}^{k} \frac{p_i^2-1}{n-1}, n > 1$, where k is the degeneracy of the amino acid (number of synonymous codons), p_i is the fraction of each synonymous codon (i) out of the total codons (n) for that amino acid, $p = n_i/n$	The ENc is a measure designed to assess codon usage bias by evaluating how far the observed usage of synonymous codons deviates from an anticipated equal distribution	The ENc values range from 20 (maximum codon bias, only one codon used for each amino acid) to 61 (no bias, all synonymous codons used equally). If the ENc value is less than 35, it indicates a pronounced bias	Wright (1990)	2306

TABLE 1 (*Continued*) Metrics for codon optimization with formal definition and description. The number of citations was retrieved from the Scopus database.

Index	Mathematical formula	Principle	Interpretation	Original source	Citation count
Effective number of codon-pair (ENcp)	$ENcp = \sqrt{\sum m(\frac{k_m}{F_m})}$, where Fm is the average degeneracy of all amino acid pairs with m synonymous codon pairs, and km is the number of amino acid pairs with m synonymous representations	ENcp is a metric designed to assess the bias in the usage of codon pairs, defined analogously to ENc, with the addition of a square root	Similar to ENc	Alexaki et al. (2019b)	100
Codon usage similarity index (COUSIN)	$COUSIN_{18}^{a} = \frac{1}{N} \times \frac{\sum cek_a W_{c,a}^{que}}{\sum cek_a W_{c,a}^{ref}}$ $COUSIN_{59}^{a} = f_a^q \times \frac{\sum cek_a W_{c,a}^{que}}{\sum cek_a W_{c,a}^{ref}}$, where N is the number of amino acids present in both the query and the reference and k_a is the set of synonymous codons coding amino acid a. weight for each codon (Wc,a) in reference and test set f_a^q is the frequency of the amino acid a in the query	COUSIN evaluates the codon CUB of a given query in comparison to a reference, and it standardizes the results by applying a Null Hypothesis that assumes random codon usage. In COUSIN18, all 18 families of synonymous codons contribute equally to the overall index, whereas in COUSIN59, each family contributes proportionally based on the frequency of the corresponding amino acid in the query	A COUSIN score of 1 signifies that the CUB in the query closely resemble those in the reference dataset. A COUSIN score of 0 indicates that the CUB in the query align with those in the Null Hypothesis. For scores exceeding 1, the CUB in the query share similarities with those in the reference but on a larger scale. Scores falling between 0 and 1 suggest that the CUPrefs in the query are akin to those in the reference but with a smaller magnitude. A score below 0 implies that the CUPrefs in the query are opposite to those in the reference	Bourret et al. (2019)	51
Average ratio of RSCU (ARSCU)	$ARSCU = (\frac{\sum_{i=1}^{aa18} \frac{a}{b}}{18})$, where aa is amino acid, a is RSCU of GC end codons and b is RSCU of AT end codons (any a and b with a value of zero is arbitrarily assigned a value of 0.1)	Measures the ratio of RSCUs with GC-ending codons to the AT-ending codons for all amino acids in a gene	Genes exhibiting ARSCU values surpassing 13 (a subjective threshold) are anticipated to showcase heightened expression levels. Genes with ARSCU values within the range of 9–13 are predicted to have either high or intermediate expression, while genes possessing ARSCU values below 9 are expected to demonstrate low or intermediate expression	Chamani Mohasses et al. (2020)	21
Relative codon bias strength (RCBS)	$RCBS = (\prod_{l=1}^{L} (1 + d_{xyz}^i))^{\frac{1}{L}} - 1, d_{xyz} = \frac{f(x,y,z) - f_1(x)f_2(y)f_3(z)}{f_1(x)f_2(y)f_3(z)}$, where L is the length, in codons, of the gene, f (x,y,z) the observed frequency of codon xyz and f1(x), f2(y) and f3(z) the observed frequencies of bases x, y and z at, respectively, codon positions 1, 2 and 3	It calculates the observed frequency of specific codons relative to the expected frequency, considering biases in base composition at three codon sites, providing a reliable measure of codon preferences while accounting for sequence-specific features like GC content	An RCBS value near 0 implies an absence of codon usage bias, while a value exceeding 0.5 indicates a notable preference for specific codon usage	Roymondal et al. (2009)	109
Directional codon bias score (DCBS)	$DCBS = \frac{\sum_{1}^{L} d_{xyz}}{L}, d_{xyz} = max(\frac{f(x,y,z)}{f_1(x)f_2(y)f_3(z)}, \frac{f_1(x)f_2(y)f_3(z)}{f(x,y,z)})$,	DCBS is based on RCBS and allows the measurement of both positive and negative codon usage bias	Similar to RCBS.	Sabi and Tuller (2014)	116

(Continued on following page)

TABLE 1 (*Continued*) Metrics for codon optimization with formal definition and description. The number of citations was retrieved from the Scopus database.

Index	Mathematical formula	Principle	Interpretation	Original source	Citation count
Modified relative codon bias strength (MRCBS)	$MRCBS = \prod_{i=1}^{N} (MRCBS_{xyz})^{\frac{1}{N}}$ $MRCBS_{xyz} = \frac{RCBS_{xyz}}{RCBS_{aa,max}}$, where fxyz is the normalized codon frequency of a codon xyz and fn(m) is the normalized frequency of base m at codon position n in a gene. RCBSaa, max is the maximum value of RCBS of codon encoding the same amino acid aa in the same reference set, and N is the codon length of the query sequence	Ribosomal protein is used as a reference set of genes and the dependence on gene length is overcome	The score of the modified relative codon bias ranges from 0 to 1	Das (2017)	1
Relative codon adaptation (RCA)	$RCA_{xyz} = \frac{f(x,y,z)}{f_1(x)f_2(y)f_3(z)}$, $RCA = (\prod_{i=1}^{L} RSA_{xyz}(i))^{1/L}$, where fxyz is the observed relative frequency of codon xyz in any reference gene set, fi(m) is the observed relative frequency of base m at codon position i in the same reference set and L is the length of the query sequence	The RCA index calculates the anticipated frequency of a codon within a provided reference set by considering the positional base frequencies. Then, it assesses codon adaptation by comparing the observed codon frequency with the anticipated frequency	The score of the RCA ranges from 0 to 2	Fox and Erill (2010)	107
Codon deviation coefficient (CDC)	The calculation of the metric can be found in the original source	The index takes into account the nucleotide composition of the sequence, GC content, and purine content. CUB is estimated using the cosine distance between the expected and observed codon usage vectors. Assessing the statistical significance of results using bootstrap resampling	Values range between 0 and 1: value 0 - no bias, 1 - maximum bias	Zhang et al. (2012)	62
Index of Translation Elongation (ITE)	$I_{TE} = \frac{\sum_{i=1}^{N_s} F_i \ln w_i}{\sum_{i=1}^{N_s} F_i}, w_i = \frac{S_i}{Max(S_i)}$, where F_i is the frequency of codon i, N_s is the number of sense codons (excluding those in single-codon families)	CAI-like index, but it involves determining a weight for each codon by considering its frequency within the NNR and NNY codon subfamilies in the reference set. Subfamilies are distinguished due to different translation by different tRNAs and susceptibility to different mutational errors	ITE values range between 0 and 1. A greater score is assigned to genes containing codons that are more commonly found in highly expressed genes	[25]	136
Synonymous codon usage order (SCUO)	$SCUO = \sum_{i=1}^{n_i} (\frac{\sum_{j=1}^{n_i} x_{ij}}{\sum_{i=1}^{18} \sum_{j=1}^{n_i} x_{ij}}) SCUO_i$, $SCUO_i = \frac{H_i^{max} - H_i}{H_i^{max}}$ where j is the codon i-th amino acid. $SCUO_i$ is the SCUO for i-th amino acid in each sequence and Hi and $Hmaxi$ are the entropy and maximum for an i-th amino acid in a sequence	The SCUO index assesses how much a sequence deviates from a uniform distribution, using Shannon entropy as a basis. It involves the normalized difference between the maximum entropy and the observed entropy	SCUO the value varies between 0 and 1, and higher values indicate a stronger codon usage bias	Wan et al. (2006)	21

TABLE 2 Example representation of the 4-letter amino acid sequence ADGY (alanine-aspartic acid-glycine-tyrosine) via synonymous codons. Nucleotide sequence of wild-type GCC-GAT-GGT-TAT. There are 4 codon variants for the first and third amino acids, and 2 variants for the second and fourth amino acids. Total 64 possible variants of nucleotide presentation of this sequence.

Synonymous codon variants	1st	2nd	3rd	4th
Amino acid				
A	GCT	GCC	GCA	GCG
D	GAT	GAC		
G	GGT	GGC	GGA	GGG
Y	TAT	TAC		

TABLE 3 An example of how the LCO method works to optimize the four codons of the mRNA encoding ADGY (see Table 2). A probability is calculated for all possible codons for a particular amino acid at a particular position. The most probable codons are marked in bold. Accordingly: GCC-GAT-GGT-TAT (wild-type nucleotide sequence)—would be optimized to GCT-GAT-GGA-TAC (final LCO-optimised sequence).

codon	Position 1	Position 2	Position 3	Position 4
	Probability of finding a codon at a given position			
GCT	**0.4**			
GCC	0.17			
GCA	0.24			
GCG	0.19			
GAT		**0.68**		
GAC		0.32		
GGT			0.26	
GGC			0.22	
GGA			**0.36**	
GGG			0.16	
TAT				0.39
TAC				**0.61**

An alternative approach suggests that the optimal structure will possess the maximum number of chemical bonds (Wayment-Steele et al., 2021). The AUP (Average Unpaired Probability) and SUP (Sum of Unpaired Probabilities) metrics, employed to assess RNA stability against hydrolytic degradation, operate under the premise that structures formed by paired bases exhibit lower susceptibility to hydrolysis.

Cluster analysis discovered that different mRNAs preferentially use different types of codons. Some mRNAs predominantly use optimal codons, while others prefer non-optimal codons. Furthermore, they observed that mRNAs with a higher proportion of optimal codons tend to be more stable, while those with a lower proportion of optimal codons are more unstable. Based on conducted experimental research, a metric called the codon stability coefficient (CSC) has been proposed. It is calculated as

the Pearson correlation coefficient between the frequency of each codon and mRNA half-lives (Presnyak et al., 2015).

In the standard genetic code, the first two positions of a codon play a decisive role in coding an amino acid, while the third position is variable for one amino acid. Collection of metrics developed GC1, GC2, and GC3 represents the frequency of G + C usage at the first, second, and third positions, respectively (Stenico et al., 1994). Another evaluation derived from RSCU is the Average RSCU Ratio (ARSCU) (Chamani Mohasses et al., 2020). Its noteworthy feature involves considering the base at the third position of the codon. The optimization of protein expression often involves the frequent usage of GC content. The model of post-transcriptional mRNA regulation involving P-bodies, 5′-3′ exonuclease XRN1, RNA helicase DDX6, and enhancer of decapping PAT1B shows that GC-rich coding sequences (CDS) result in higher protein production compared to AU-rich ones, and are controlled by a mechanism involving degradation factors DDX6 and XRN1 (Courel et al., 2019). On the contrary, reducing the GC content in the 5′-UTR leads to an increase in free energy and also enhances protein yield, presumably due to mRNA destabilization in the translation initiation region and greater accessibility of the ribosome binding site (Dewi and Fuad, 2020). The GC3 content varies depending on the type of tissue but is not an exhaustive characteristic for tissue-specific gene separation (Plotkin et al., 2004). GC3 codons are also associated with a longer half-life of mRNA (Kudla et al., 2006; Hia et al., 2019).

2.3 Metrics for adaptation to tRNA pool

Codon usage bias is closely linked to translational selection, which is the process of selecting codons that match abundant tRNAs, the molecules responsible for carrying amino acids during protein synthesis. Highly expressed genes tend to use such preferred codons, resulting in enhanced translation rates and accuracy. Dittmar et al., 2006 showed that the expression levels of nuclear and mitochondrial tRNAs vary between human tissues, indicating tissue-specific translational selection. However, minor differences in mouse mitochondrial RNA have only been detected for cardiac tissue, while significant differences between the central nervous system and other tissues have been demonstrated at the level of tRNA isodecoders, i.e., transcripts with the same anticodon but encoded by numerous different genes (Pinkard et al., 2020). It is important to note that the strength of translational selection varies across different organisms based on their genome sizes and genomic tRNA content (Reis, 2004).

To account for the role of intracellular tRNA content in translation efficiency, the following indices have been developed: P2index (Gouy and Gautier, 1982) and tRNA adaptation index (tAI) (dos Reis, 2003).

Initially, tAI was only applicable to *S. cerevisiae*, but its subsequent modifications, stAI (Sabi et al., 2017) and gtAI (Anwar et al., 2023)—overcome this limitation by incorporating species-specific weights through algorithmic approaches to find extrema. gtAI demonstrated greater efficiency by employing a genetic algorithm to identify the optimal set of weights. In its calculation, indices ENc and RSCU are also incorporated. gtAI

ranges from 0 to 1, where a higher value implies better adaptation of the codon to the tRNA pool.

The P2 Index is a metric used for the quantitative assessment of the efficiency of interactions between codons and their corresponding anticodons during the translation process. Based on the frequency of specific types of codons, values exceeding 0.5 indicate the presence of translational selection influencing the coding sequence.

2.4 Algorithmic approaches and tools for codon optimization

Currently, various optimization algorithms are utilized, such as the genetic algorithm (Błażej et al., 2018), multi-objective artificial bee colony (Gonzalez-Sanchez et al., 2019), Ribotree Monte Carlo (Leppek et al., 2022), and dynamic programming (Pham et al., 2004; Taneda and Asai, 2020), to identify codon combinations with desired characteristics. In several studies, the use of recurrent neural networks for codon optimization in heterologous protein expression has been presented in Chinese hamster (*Gricetulus griseus*) ovary cells (Goulet et al., 2023) and *E. coli* (Jain et al., 2023). The Bidirectional Long Short-Term Memory (LSTM) deep learning model has also been trained for *E. coli* (Fu et al., 2020).

Other studies applied machine learning methods for mRNA stabilization, such as integrated deep learning-based mRNA optimization (iDRO) (Jain et al., 2023), which provides a two-step optimization for the open reading frame and the untranslated regions. S. Castillo-Hair and G. Seelig trained a model on the 5′UTR polysome profile dataset to predict ribosome loading and protein expression (Castillo-Hair and Seelig, 2022). The predictive power of such models strongly depends on the quantity and quality of the training datasets. At the same time, the accumulation of experimentally verified data sets is often not as fast as the development of machine learning methods. For example, to date (February 2024) only 6,142, of which 1,416 are human, experimentally validated RNA structures have been deposited in the Protein Data Bank (Berman, 2000). This indicates that the high-precision prediction of RNA 3D structures using machine learning methods may be accurate for training data, but not for new data (Sato and Hamada, 2023).

Several software tools that utilize statistical and algorithmic solutions are available for commercial and free use. Here, we present some current tools that can be used for various tasks, including those related to gene therapy: ATGme (Daniel et al., 2015), OPTIMIZER (Puigbo et al., 2007), CHARMING (Wright et al., 2022), %MinMax (Rodriguez et al., 2018), JCat (Grote et al., 2005), Optipyzer (LeRoy and Roleck, 2023), IDT (Owczarzy et al., 2008), gtAI (Anwar et al., 2023).

3 Codon optimization for gene therapy vectors

Above, the elucidation of metrics and principles related to codon optimization has been expounded. At the same time, it should be noted that the resources required to test the functionality of *in silico* predicted RNA variants significantly exceed the cost of the prediction itself. For this reason, studies often mainly present

unconfirmed hypotheses in *in vitro* or *in vivo* experiments. Nevertheless, we present below some examples where codon optimization has been successfully applied *in vitro*. Proceeding to *in vitro* studies, it should be noted that gene therapeutics consist of a delivery vector and a therapeutic gene. Currently many types of vectors are used as a transgene vehicle (e.g., lipoplexes (Chen et al., 2016), polyplexes (Hayat et al., 2019), virus-like particles (Pitoiset et al., 2017)).
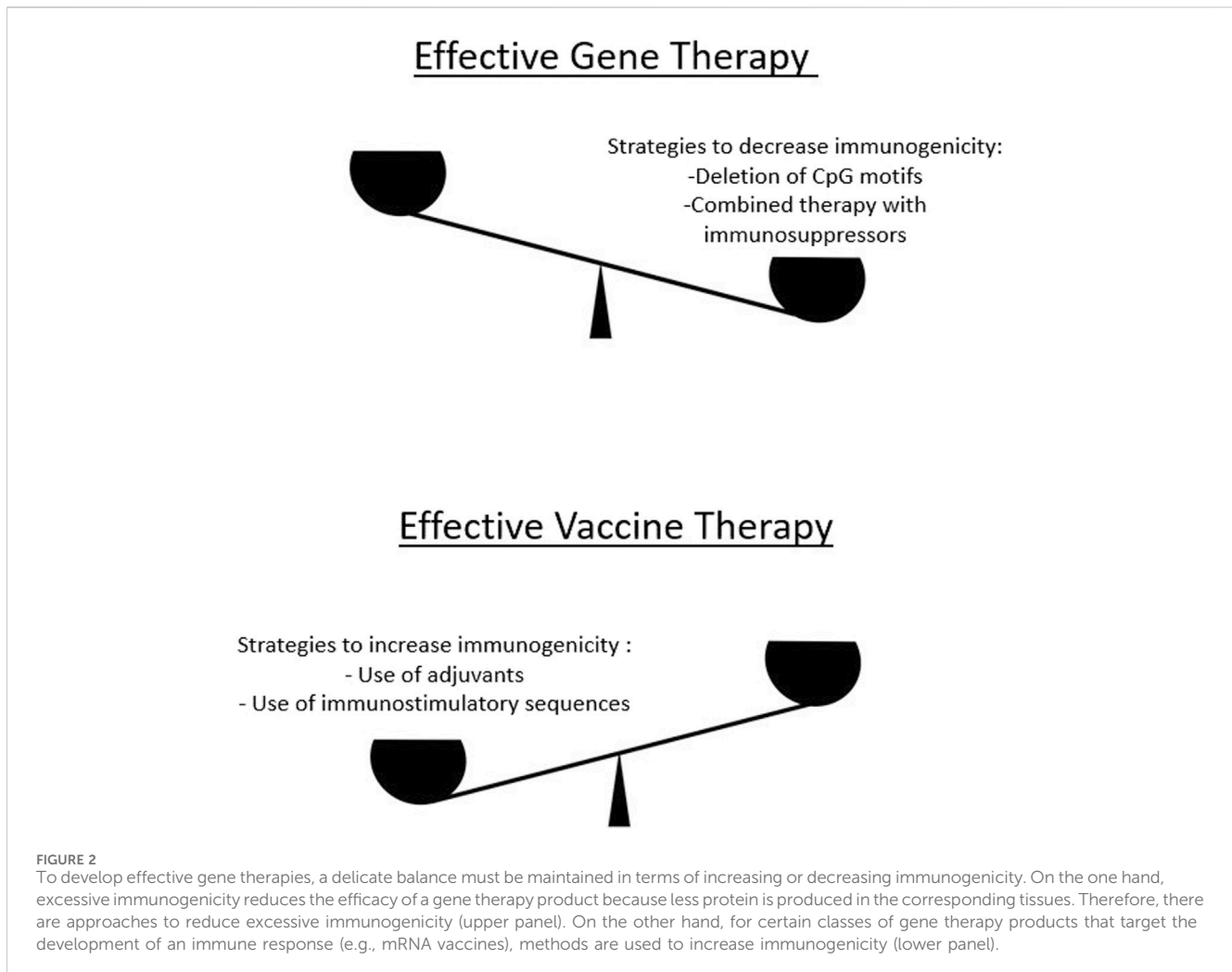
Some of these vectors are a cassette with the selected viral genes, others do not contain nucleic acids. In some cases, wild-type viral genes in the gene therapy vector are not optimized for efficient application (Bainbridge et al., 2008). At the same time, codon-optimized variants of these sequences increase the efficacy of gene therapy, although they may lead to unfavorable results such as undesirable conformational changes and subsequently alterations in protein activity and function. Examples of codon optimization of adenoviral (Coughlan, 2020), retroviral and lentiviral vectors (Breckpot et al., 2010) are discussed below.

Since adeno-associated vectors have recently become the most widely used platform for gene transfer (Mendell et al., 2021) and adenoviruses have long been successfully used to deliver genes (Bulcha et al., 2021), we will consider the application of optimizations on their example.

It has been shown that in adenoviruses, the genes responsible for highly abundant late structural proteins tend to use codons frequently used in humans (optimal codons), while early regulatory use less optimal codons (Villanueva et al., 2016). However, the adenoviral fiber protein specifically uses suboptimal codons for efficient viral replication. Surprisingly, analysis of transgenes expressed in oncolytic adenoviruses, that are used for the oncoselective expression of a wide range of therapeutic molecules (de Sostoa et al., 2019; Huang et al., 2019) shows that most transgenes also use suboptimal codons. This contradicts the recommendation to use optimal host codons in transgenes to maximize gene expression. The study investigates the impact of transgene codon usage on viral fitness and finds that transgenes with higher GC3 content (optimal codon usage) have higher gene expression and viral replication, while those with lower GC3 content have lower expression and replication (Núñez-Manchón et al., 2021). By tuning the codon usage of transgenes, it is possible to achieve better transgene expression without compromising viral replication, thus optimizing the therapeutic outcome.

In the development of gene therapies, the problem arises of achieving high titers and a high ratio of empty to full capsids in viral vectors. One of the solutions to this obstacle is codon optimization of viral genomes encoding capsid proteins and assembly proteins. Thus, not only transgenes but also the coding sequences of the viral vector itself are subjected to codon optimization. For AAV-based (adeno-associated virus) vectors a novel codon optimization method was presented (Localized Codon-Optimization or LCO) (Cabanes-Creus et al., 2019).

This method aims to preserve functional elements of the capsid genes and improve capsid shuffling efficiency for AAV engineering. The LCO algorithm performs localized optimization of codons at each position independently, based on the usage frequency of codons observed in the input variants of AAV sequences. A codon usage frequency table is generated for each amino acid

FIGURE 2
To develop effective gene therapies, a delicate balance must be maintained in terms of increasing or decreasing immunogenicity. On the one hand, excessive immunogenicity reduces the efficacy of a gene therapy product because less protein is produced in the corresponding tissues. Therefore, there are approaches to reduce excessive immunogenicity (upper panel). On the other hand, for certain classes of gene therapy products that target the development of an immune response (e.g., mRNA vaccines), methods are used to increase immunogenicity (lower panel).

position, and this table is used to optimize individual sequences (Table 3). The LCO-modified capsid genes showed increased sequence identity between parental AAV capsids and novel AAV capsid variants.

Functionality tests demonstrated that the optimized capsids retained their function, and transduction efficiency was similar to unoptimized counterparts. The LCO method also improved the efficiency of capsid shuffling, resulting in a highly shuffled library with increased complexity and reduced size of donor sequence segments. The shuffled clones generated using LCO-encoded capsids demonstrated successful transduction, indicating the effectiveness of LCO in generating novel AAV variants.

Ironically, the extensive use of codon optimization occurred simultaneously with abundant research findings that revealed the impact of synonymous mutations on protein function. This has been shown on a variety of proteins (Buhr et al., 2016; Kirchner et al., 2017).

The mechanism being discussed involves the comparison between codon-optimized (CO) and wild-type (WT) variants of a protein named FIX (coagulation factor IX). The results highlight that the CO and WT FIX variants exhibit distinct conformations, suggesting that the codon optimization process has influenced the protein's structure. Ribosome profiling analyses uncover altered ribosomal distribution patterns and local translational kinetics in the CO variant when compared to the WT

variant. Notably, these differences are unique to the CO FIX variant, as control genes demonstrate comparable ribosome distribution profiles (Alexaki et al., 2019a).

Despite the observed differences in translational kinetics, the overall efficiency of protein synthesis between the CO and WT variants remained similar. This finding is consistent with previous studies conducted *in vitro* (outside of a living organism) and suggests that the rate of protein synthesis is comparable between the two variants. The researchers propose that differences in translational kinetics within these domains may contribute to the observed conformational differences between the CO and WT FIX variants.

Codon optimization can be approached not only by a global view of codon usage in general, but also by a local optimization for each individual position in a particular amino acid. Moreover, it is also important to check that the functions of the essential elements and the optimized protein of interest remain unchanged.

4 The effect of codon optimization on immunogenicity

The immune response to an administered foreign substance or molecule can be defined as immunogenicity. It should be noted that

higher immunogenicity increases the efficacy of the drug in some cases, but decreases it in others (Figure 2). For example, the purpose of immunization is to generate an immune response against a pathogen. In this case, methods should be used to increase the immunogenicity of the drug. It should be noted that in the development of mRNA vaccines, an excessive overreaction of the immune system is undesirable due to possible damage to the human organism (Igyártó and Qin, 2024) and should be taken into account during codon optimization. On the other hand, if a transgene introduced into the organism is intended to lead to the production of the corresponding protein, any degree of immunogenicity will reduce the effectiveness of the therapy. The innate and adaptive immune response to gene therapy may vary depending on the source of immunogenicity. These may be factors related to the capsid of the virion or to the viral genome. In relation to the capsid, binding of TLR2 or TLR9 can potentially activate the innate immune response and initiate the MyD88 signaling cascade, which in turn stimulates the production of proinflammatory cytokines such as TNF-alpha or induces the synthesis of IFN-gamma (Yang et al., 2022). Depending on the composition of the viral vector, the innate immune response can lead to enhanced adaptive immune responses. For example, AAVs, which are often used as gene therapy vectors, circulate naturally between humans. As a result, most people develop antibodies against natural AAV serotypes due to previous exposure. These antibodies are even known to cross-react with engineered vectors (Boutin et al., 2010). As a result, these antibodies can lead to either complement activation or neutralization of the capsid. The adaptive immune response is characterized by the degradation of the capsid protein by the proteasome and peptide presentation on MHC class I molecules. CD8[+] cytotoxic T-cell lymphocytes can bind to the MHC, which leads to cell death (Martino et al., 2013). Peptide presentation on MHC class II molecules after phagocytosis and proteolysis can be recognized by CD4[+] T lymphocytes, which can then stimulate the proliferation of B cells and the production of capsid-specific antibodies (Li et al., 2013). Studies have shown that plasmacytoid dendritic cells (pDCs) and conventional dendritic cells (cDCs) co-operate to achieve cross-priming of CD8[+] T cells (Rogers et al., 2017). pDCs recognize the AAV genome via TLR9, while cDCs present the antigen on MHC I. The binding of cytokine-produced IFN to its receptor on cDCs is necessary for this process, indicating a direct relationship between pDC-produced cytokines and the activation of cDCs. Cross-priming of CD8[+] T cells against AAV capsids requires CD40−CD40L co-stimulation, which is performed in addition to T1 IFN from CD4[+] Th cells (Shirley et al., 2020b).
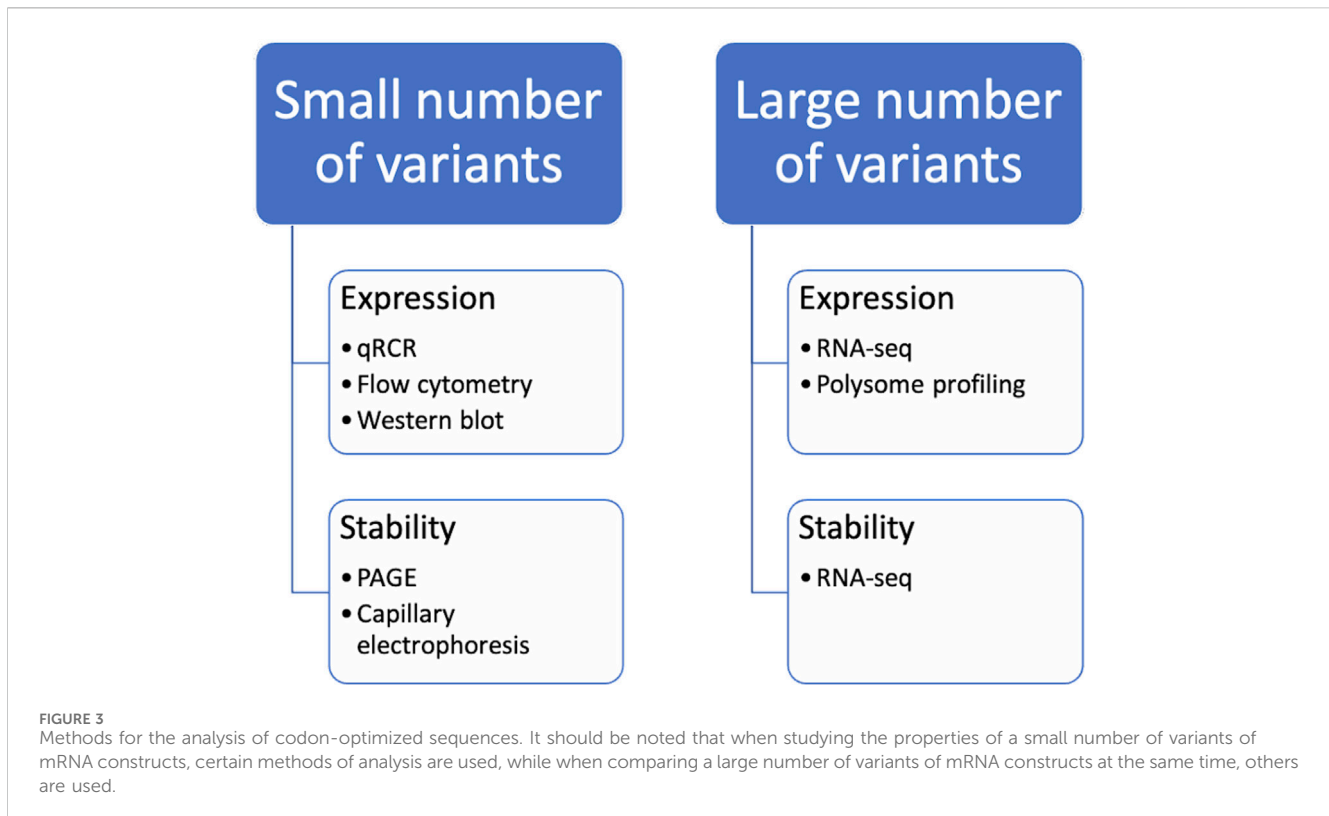
After viral uncoating, TLR9 receptors can recognize unmethylated CpG motifs in the released single-stranded DNA, which also leads to activation of the innate immune system and stimulates cytokine production. The humoral and cellular innate immune responses described above for AAV capsids also occur for the transgene protein. The adaptive immune response can depend on various factors such as the target tissue, vector design and dose. Depending on the specificity of the promoter, there is a potential risk of immunogenicity (Shirley et al., 2020a). For example, a ubiquitous promoter can increase the risk of an adaptive cellular immune response of target and non-target cells (Sun et al., 2005).

It should be noted that the appearance of a foreign protein in the human organism is associated with the development of autoimmune diseases due to the similarity of individual epitopes of foreign and self proteins (Rojas et al., 2018). For example, it was recently shown that the same antibodies cross-react with the Epstein-Barr virus protein and the human alpha-crystallin B protein (Thomas et al., 2023). This phenomenon of molecular mimicry could be associated with the development of multiple sclerosis. The possibility of molecular mimicry of proteins resulting from the translation of the nucleic acids used must therefore be taken into account in the development of gene therapeutics. As already mentioned, codon optimization of the RNA can influence the structure of the translated protein (Alexaki et al., 2019a). As a result, depending on the different variants of the synonymous substitutions, the presentation of different epitopes of the same protein is possible.

It is of interest to reduce these CpG motifs to circumvent the possible human immune response, which can be achieved by codon optimization. For example, various elements of an AAV vector such as the CMV enhancer and promoter, ITR regions, UTR regions and the therapeutic transgene itself may contain CpG motifs. The CpGs within the promoter sequence can be removed, but with unpredictable effects on the activity and specificity of the promoter. For example, the authors have shown that the removal of CpGs within the CMV promoter gene significantly reduces its activity (Yew and Cheng, 2004). Although CpGs can be removed from the expression cassette, as in the case of human coagulation factor IX (hFIX) (Bertolini et al., 2021), this does not always increase efficiency—CpG elimination had only reduced antibody formation against the transgene and not against the capsid itself. There are several studies in which this strategy was used, but mostly with a modification of the transgene. They have shown that the elimination of CpG motifs may lead to a significant reduction in the CD8[+] T cell response (Yew and Cheng, 2004; Faust et al., 2013; Herzog et al., 2019; Wright, 2020; Bertolini et al., 2021; Konkle et al., 2021).

Several codon optimization strategies, including the chemical modification of nucleosides (Karikó et al., 2005) and the incorporation of pseudouridine (Karikó et al., 2008; Anderson et al., 2010; Thess et al., 2015), have been shown to improve translation and reduce the immune response to mRNAs. pDCs exposed to such modified RNA exhibit a significant reduction in cytokines and activation markers. Nucleoside modification at a single position in a chemically synthesized oligoribonucleotide (ORN) is sufficient to abrogate TLR activation. In addition, the incorporation of pseudouridine in particular has been shown to facilitate evasion of recognition by Toll-like receptors (Karikó et al., 2005), although the molecular differences contributing to this mechanism has not yet been elucidated. Although the implementation of pseudouridine increases the stability of the mRNA and its translational capacity, it is important to note the disadvantages of replacing uridine with pseudouridine (Xia, 2021; Mueller, 2023). A recent study has shown that the presence of pseudouridine in IVT mRNA increases ribosomal + 1 frameshifting during mRNA translation. In addition, new peptides were generated that triggered an immune response (Mulroney et al., 2024). The presence of pseudouridine in the stop codon region suppresses translation termination and allows non-canonical base pairing, which is particularly detrimental for *in vitro* transcribed mRNAs (Loomis et al., 2016). The negative effects of pseudouridine synthases have been associated with various cancers (Xue et al.,

FIGURE 3
Methods for the analysis of codon-optimized sequences. It should be noted that when studying the properties of a small number of variants of mRNA constructs, certain methods of analysis are used, while when comparing a large number of variants of mRNA constructs at the same time, others are used.

2022) and autoimmune diseases (Festen et al., 2011). This strongly suggests that the influence of codon optimization and pseudouridine incorporation on mRNA expression needs to be further investigated. A limitation of the present review is that it does not focus on a detailed description of the specific effects of codon optimization on the mRNA vaccines against COVID-19 *per se* that have been introduced into clinical practice (reviewed in Xia, 2021), but aims to discuss the advantages and disadvantages of the different options for the use of codon optimization in gene therapy in general.

To summarize, a common strategy to avoid immunogenicity is to eliminate redundant CpG motifs, implement chemical modifications of ORNs and replace uridine with pseudouridine. However, it should be noted that the implementation of codon optimization to eliminate CpG motifs and pseudouridine modification must be performed strategically to avoid the negative consequences of both approaches. Given the various unresolved factors leading to potential immunogenicity as a consequence of gene therapy, developing metrics for prediction is a complicated task. Nevertheless, a recent report (Wright, 2020) proposed a metric for prediction focusing exclusively on CpG motifs and their potential immunogenicity. Three formulas were developed that take into account the amount of unmethylated CpG motifs in the vector sequence. Known immunostimulatory sequences commonly used in DNA vaccines were also considered in the development of the formulae (Bode et al., 2011). Although these formulae still need to be improved for full validation and accurate prediction, they reflect the beginning of a deeper understanding of how codon optimization can contribute to the reduction of immunogenicity.

5 Experimental testing of codon optimized sequences

There are numerous strategies for optimizing codons in nucleic acids. The methods mentioned above enable the creation of numerous optimized sequence variants. However, experimental verification of properties such as mRNA stability and protein expression levels is necessary before further experimentation can be conducted. Depending on the goals and available resources, it may be possible to select the best candidates based on chosen criteria from the range of design variants. These candidates can then be examined using routine laboratory methods. Alternatively, a pool of hundreds of sequences can be studied, in which case high-throughput protocols must be developed (Figure 3).

When studying a small number of variants, it is possible to determine the expression level separately for each construct after transfecting the cells. To quantify transgene expression in this case, the most common method is to use target-specific primers with cDNA obtained from RNA by reverse transcription as a matrix and perform qPCR (Leppek et al., 2022). Expression can be quantified at both the transcriptional and translational levels. The latter involves the analysis of synthesized proteins and can be performed using antibodies specific to the target protein. For instance, Zhang (Zhang et al., 2023) described the properties of the optimized structure of the SARS-CoV-2 virus S protein using flow cytometry. A possible alternative method for determining protein concentrations is to use SDS-PAGE gels for Western blot analysis, along with specific antibodies (Raab et al., 2010; Fath et al., 2011).

Although codon optimization of the target sequence can provide certain benefits, it may also result in reduced mRNA stability in

solution, which impairs its functionality. Therefore, it is necessary to experimentally confirm the stability of the structure of optimized nucleic acids. The stability of mRNA molecules is inversely proportional to their degradation rate in solution. To determine the degradation rate, mRNAs are incubated in PBS buffer containing Mg2+ ions. Samples are collected at various time intervals of 1–2 h, and the number of fragments produced is estimated using capillary electrophoresis (Zhang et al., 2023) or polyacrylamide gel electrophoresis with urea. Therefore, the RNA is less stable if it degrades more quickly after being incubated in solution.

However, the laboratory approaches described above are time-consuming when testing multiple variants of codon-optimized sequences. In light of this, there is a great need to create high-throughput methods for studying many sequences simultaneously.

Most methods that allow mass screening of sequences follow a general principle: a unique barcode, a sequence of several nucleotides, is inserted into each variant. All the sequences to be tested can then be pooled and processed in a multiplex format. The presence of the barcode makes it possible to identify a variant using high-throughput sequencing platforms after all the necessary protocol steps have been completed.

Massively parallel variant analysis requires the synthesis of a library of DNA templates. The next steps in the study can be performed in two ways. The first involves transcription and modification (3′ polyA tail and 5′ m7G capping) *in vitro*, followed by transfection of the resulting mRNA pool into cells for further experiments. The "PERSIST-seq" method was developed based on this approach. It enables the simultaneous evaluation of stability and translation efficiency of over 200 mRNA molecules, making it a convenient tool for messenger RNA development (Leppek et al., 2022). In this case, the design of the DNA must take into account the presence of a promoter in the initial sequence. The second approach involves creating a vector library with cassettes that contain the sequence under study and regions of homology. The cells are then transfected with the library, and the sequences are integrated into the genome using CRISPR/Cas. This process enables the direct synthesis of mRNA within the cells. A study of the motifs that cause ribosome slowdown in a yeast model system describes a similar approach (Chen et al., 2023). The next steps for experimental validation in both cases involve isolating RNA from cell culture, analyzing it through high-throughput sequencing, and quantifying the results. To identify inserts in the pool of isolated nucleic acids, unique barcodes are introduced into the library construct, which is a common aspect of the described strategies.

The presence of unique barcodes in the original DNA matrices allows quantitative assessment of the expression level for each individual variant using high-throughput RNA sequencing.

Translation of sequence variants has been demonstrated to be a crucial determinant in mammalian gene expression (Burke et al., 2022). However, genomic expression profiling alone cannot reveal the precise regulation provided by post-transcriptional mechanisms, such as 5′ capping, splicing, polyadenylation, nuclear export, translation, and decay. To overcome this limitation, a polysome profiling method can be used to isolate ribosome-free and polysome-associated RNAs for further independent analysis (Pereira et al., 2018) This method involves separating mRNA in a sucrose gradient into two fractions: polysome-bound and polysome-free. The mRNA

is then isolated from both fractions and sequenced using one of the available high-throughput platforms.

When studying multiple variants, stability assessment is also important. To identify full-length molecules that have not degraded, it is necessary to amplify the cDNA that was reverse transcribed from the RNA and then sequence it to quantify the amount of intact mRNA at each time point. This method can evaluate mRNA stability in both solution and cells. The solution replicates the conditions in which the molecules may be present during therapy, typically high pH and positively charged media. It is important to note that the outcomes obtained after incubation in solution differ significantly from those obtained after isolation from cells. This is likely due to cellular mechanisms of RNA degradation (Leppek et al., 2022).

Therefore, there are approaches that allow for the evaluation of the efficiency and stability of nucleic acid sequences obtained during codon optimization. The choice of a particular method depends on the number of variants to be analyzed. If there are only a few variants, it is possible to describe the properties of each variant separately, providing a fairly accurate understanding of its characteristics. When dealing with hundreds or thousands of variants, high-throughput methods are necessary. This allows for a pool of samples to be tested instead of individual samples, greatly increasing the productivity of experimental work. It is important to note that massively parallel sequencing methods provide high accuracy analysis, while polysome profiling can offer additional insights into the impact of codon optimization on the final product's quality.

6 Future directions

Currently, there are some gene therapies that use different codon optimization metrics and are approved by the FDA (FDA, 2024). To analyse other therapies that are in clinical trials and where codon optimization has been used, we conducted a thorough examination of the data available on ClinicalTrials.gov (ClinicalTrials.gov, 2024) until December 2023. A systematic search strategy was devised using the keyword "gene therapy" in the Condition/disease field. In addition to the specified search criteria, it is important to note that the term "vector" was included in the "Other terms" considered in the search. The algorithm did not include any specified values for the "Intervention/treatment" and "Location" categories in the search process. After searching, the algorithm automatically incorporated synonyms for the given query: gene: "Genes," gene therapy: "Gene transfer"; "Gene Transfer Procedure,", therapy: "treatment"; "Therapeutic"; "therapeutics".

Furthermore, a comprehensive search was conducted using the specific only Condition/disease of "codon optimized" and excluded any specified values for the "Other terms," "Intervention/treatment" and "Location" categories in the search process. However, it is crucial to mention that studies explicitly referring to monoclonal antibodies and enzymes as drugs in the Study URL and Brief Summary columns were manually excluded from the sample. This careful exclusion strategy ensured that the selected studies focused specifically on codon optimization. The search was conducted over a period of 20 years to capture an extensive range of relevant clinical studies.
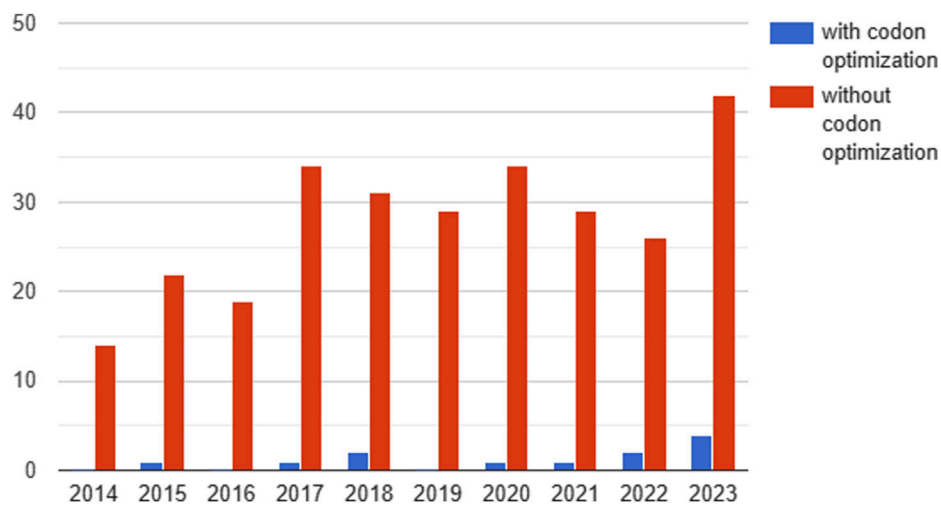
FIGURE 4
Dynamics of the number of studies reported on clinicaltrails.gov testing gene therapeutics with and without codon optimization by year (2014-2023). Since 2020, a trend towards an increase in the proportion of studies with codon optimization can be observed.
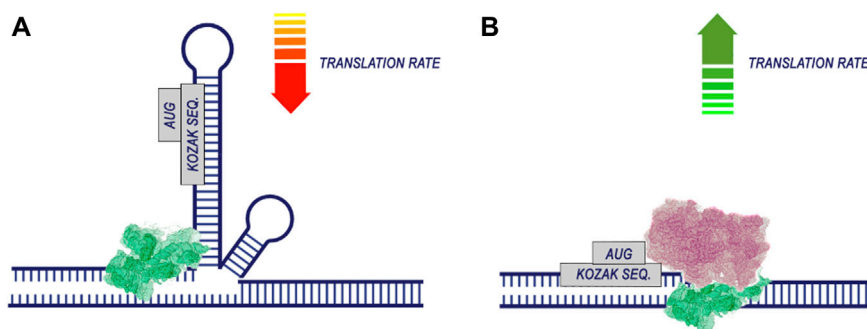

FIGURE 5
The secondary structure of RNA reduces the efficiency of translation. The process of translation initiation is completed by the recognition of the start codon by the 43S preinitiation complex and the assembly of the ribosome. If the region of the start codon is hidden in the secondary structure of the RNA **(A)**, translation is likely to be less efficient. At the same time, if there are no pronounced secondary RNA structures in the region of the start codon **(B)**, the probability of translation initiation increases.

Of the 395 clinical studies analyzed, only 12 contained information on codon optimization (Figure 4).

Prior to experimental testing of codon-optimized sequences using any of the aforementioned methods, it is essential to synthesize these sequences, often in large quantities. The most widely used method currently is phosphoramidite synthesis, which involves the interaction of nucleotide phosphoramidite monomers protected by acid-labile groups with an activating agent, binding to the growing oligonucleotide (Sinyakov et al., 2021). There are two main types of implementation for this approach, depending on the equipment used: synthesis on columns or on microarrays. The former option allows for the synthesis of oligonucleotides at a relatively low cost and with an error rate of 1 per 600 base pairs or less on average. However, it does not provide sufficient throughput for mass synthesis of oligonucleotides (Ma et al., 2012). Furthermore, if the sequence of interest exceeds 200 base pairs (some estimates suggest 300 (Palluk et al., 2018)), an additional assembly step via molecular cloning is required (Casini et al., 2015). These factors significantly limit the speed of testing and represent the primary bottleneck in experimental design.

This problem can be solved by integrating higher-throughput oligonucleotide microarray synthesisers into laboratory practice (Song et al., 2021). Commercially available technologies are also based on phosphoramidite synthesis, albeit with slight modifications. Although microarray-based nucleotide synthesis is more error-prone due to heterogeneity and edge effects, it enables the synthesis of oligonucleotide pools and also reduces the cost per nucleotide by 2–4 orders of magnitude compared to column synthesis (Kosuri and Church, 2014). This suggests that advances in *de novo* DNA synthesis and experimental verification of codon-optimized sequences are likely to be associated with the microarray approach.

Since 2020, a trend towards an increase in the proportion of codon-optimized studies has been observed. In 2020, 1 in 34 (2.9%) clinical trials used codon optimization, compared to 4 in 42 (9.5%)

in the first 11 months of 2023 (Figure 4). The main aim of codon optimization was to increase the level of transgene expression and the stability of the mRNA. In addition, a study using codon optimization to reduce immunogenicity was reported in 2021.

To effectively achieve the goals of codon optimization in research, it is important to follow established metrics. However, today there is no single generally accepted standard for codon optimization. Therefore, it is possible to use a large number of combinations of the methods described above to create optimal RNA variants. Some of these approaches significantly increase the efficacy of gene therapeutics. Therefore, several drug options have been registered in clinical trials, for example.

Codon optimization has played an important role in the development of RNA-based COVID-19 vaccines. Current research efforts are focused on further advancing the field of codon optimization for COVID-19 vaccines to address new strains of the coronavirus (Wu et al., 2023). Unfortunately, it was not possible to provide here the specific metrics used for codon optimization in the above-mentioned studies for commercial product development. This limitation results from the intellectual property of the original codon-optimized constructs. In this article, we have explored various metrics for assessing codon usage, based on both the composition of the coding sequence and the composition of a reference set of genes. One widely used metric is the Codon Adaptation Index (CAI). Although these measures provide useful information about adaptation to the host organism, they do not necessarily indicate an increase in translational efficiency due to selection pressure (Rahman et al., 2018; Feng et al., 2022). Furthermore, CAI is also interpreted as an indicator of the speed of translational elongation (Kudla et al., 2009). In turn, an increase in translation speed may not necessarily result in the production of a protein with similar properties in greater quantities.

Apparently, during translation, the most important regions for codon optimization are the areas around the start codon. This is supported by work demonstrating the contribution of the CDS position near the start codon (Höllerer and Jeschek, 2023; Nieuwkoop et al., 2023) and the 5′UTR sequence region (Capell et al., 2014). The efficiency of translation is significantly dependent on the energy of mRNA folding, particularly in the vicinity of the start codon (Gu et al., 2010). This is associated with the fact that unfolding more stable RNA secondary structures require greater energy before the initiation of translation (Figure 5). Additionally, the presence of hairpin, stem-loop, and pseudoknot structures in mRNA can hinder ribosome translocation and tRNA binding, thus impeding translation elongation (Kozak, 2005; Bao et al., 2020).

Thus, advancements in gene therapy could be directed towards a more comprehensive exploration of the impact of codon optimization on the characteristics and secondary structure of mRNA. Also, it is possible to apply optimization metrics locally to the start region, but there are limitations since many of them are based on codon usage frequency without taking into account the features of untranslated regions.

In addition, consideration of local codon optimization is a critical aspect that must be taken into account during codon optimization for a particular protein of interest. Furthermore, essential protein functions may change due to the possible influence of codon optimization on the conformation of the resulting protein, which should also be taken into account.

Author contributions

AP: Writing–original draft, Writing–review and editing. AK: Writing–original draft, Writing–review and editing. AM: Writing–original draft, Writing–review and editing. DN: Writing–original draft, Writing–review and editing. AS: Writing–original draft, Writing–review and editing. IA: Writing–original draft, Writing–review and editing. SF: Writing–original draft, Writing–review and editing. OM: Writing–original draft, Writing–review and editing. AD: Writing–original draft, Writing–review and editing. PV: Writing–original draft.

Funding

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alexaki, A., Hettiarachchi, G. K., Athey, J. C., Katneni, U. K., Simhadri, V., Hamasaki-Katagiri, N., et al. (2019a). Effects of codon optimization on coagulation factor IX translation and structure: implications for protein and gene therapies. *Sci. Rep.* 9, 15449. doi:10.1038/s41598-019-51984-2

Alexaki, A., Kames, J., Holcomb, D. D., Athey, J., Santana-Quintero, L. V., Lam, P. V. N., et al. (2019b). Codon and codon-pair usage tables (CoCoPUTs): facilitating genetic variation analyses and recombinant gene design. *J. Mol. Biol.* 431, 2434–2441. doi:10.1016/j.jmb.2019.04.021

Anderson, B. R., Muramatsu, H., Nallagatla, S. R., Bevilacqua, P. C., Sansing, L. H., Weissman, D., et al. (2010). Incorporation of pseudouridine into mRNA enhances translation by diminishing PKR activation. *Nucleic Acids Res.* 38, 5884–5892. doi:10.1093/nar/gkq347

Anwar, A. M., Khodary, S. M., Ahmed, E. A., Osama, A., Ezzeldin, S., Tanios, A., et al. (2023). gtAI: an improved species-specific tRNA adaptation index using the genetic algorithm. *Front. Mol. Biosci.* 10, 1218518. doi:10.3389/fmolb.2023.1218518

Athanasopoulos, T., Foster, H., Foster, K., and Dickson, G. (2011). Codon optimization of the microdystrophin gene for Duchene muscular dystrophy gene therapy. *Gene Ther.* 709, 21–37. doi:10.1007/978-1-61737-982-6_2

Athey, J., Alexaki, A., Osipova, E., Rostovtsev, A., Santana-Quintero, L. V., Katneni, U., et al. (2017). A new and updated resource for codon usage tables. *BMC Bioinforma.* 18, 391. doi:10.1186/s12859-017-1793-7

Ayyar, B. V., Arora, S., and Ravi, S. S. (2017). Optimizing antibody expression: the nuts and bolts. *Methods* 116, 51–62. doi:10.1016/j.ymeth.2017.01.009

Bainbridge, J. W. B., Smith, A. J., Barker, S. S., Robbie, S., Henderson, R., Balaggan, K., et al. (2008). Effect of gene therapy on visual function in leber's congenital amaurosis. *N. Engl. J. Med.* 358, 2231–2239. doi:10.1056/NEJMoa0802268

Bansal, S., Perincheri, S., Fleming, T., Poulson, C., Tiffany, B., Bremner, R. M., et al. (2021). Cutting edge: circulating exosomes with covid spike protein are induced by BNT162b2 (Pfizer–BioNTech) vaccination prior to development of antibodies: a novel mechanism for immune activation by mRNA vaccines. *J. Immunol.* 207, 2405–2410. doi:10.4049/jimmunol.2100637

Bao, C., Loerch, S., Ling, C., Korostelev, A. A., Grigorieff, N., and Ermolenko, D. N. (2020). mRNA stem-loops can pause the ribosome by hindering A-site tRNA binding. *Elife* 9, e55799. doi:10.7554/eLife.55799

Bell, P., Wang, L., Chen, S.-J., Yu, H., Zhu, Y., Nayal, M., et al. (2016). Effects of self-complementarity, codon optimization, transgene, and dose on liver transduction with AAV8. *Hum. Gene Ther. Methods* 27, 228–237. doi:10.1089/hgtb.2016.039

Bennetzen, J. L., and Hall, B. D. (1982). Codon selection in yeast. *J. Biol. Chem.* 257, 3026–3031. doi:10.1016/S0021-9258(19)81068-2

Berman, H. M. (2000). The protein Data Bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Bertolini, T. B., Shirley, J. L., Zolotukhin, I., Li, X., Kaisho, T., Xiao, W., et al. (2021). Effect of CpG depletion of vector genome on CD8+ T cell responses in AAV gene therapy. *Front. Immunol.* 12, 672449. doi:10.3389/fimmu.2021.672449

Błażej, P., Wnętrzak, M., Mackiewicz, D., and Mackiewicz, P. (2018). Optimization of the standard genetic code according to three codon positions using an evolutionary algorithm. *PLoS One* 13, e0201715. doi:10.1371/journal.pone.0201715

Bode, C., Zhao, G., Steinhagen, F., Kinjo, T., and Klinman, D. M. (2011). CpG DNA as a vaccine adjuvant. *Expert Rev. Vaccines* 10, 499–511. doi:10.1586/erv.10.174

Bollman, B., Nunna, N., Bahl, K., Hsiao, C. J., Bennett, H., Butler, S., et al. (2023). An optimized messenger RNA vaccine candidate protects non-human primates from Zika virus infection. *npj Vaccines* 8, 58. doi:10.1038/s41541-023-00656-4

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. doi:10.1093/gbe/evz262

Boutin, S., Monteilhet, V., Veron, P., Leborgne, C., Benveniste, O., Montus, M. F., et al. (2010). Prevalence of serum IgG and neutralizing factors against adeno-associated virus (AAV) types 1, 2, 5, 6, 8, and 9 in the healthy population: implications for gene therapy using AAV vectors. *Hum. Gene Ther.* 21, 704–712. doi:10.1089/hum.2009.182

Breckpot, K., Escors, D., Arce, F., Lopes, L., Karwacz, K., Van Lint, S., et al. (2010). HIV-1 lentiviral vector immunogenicity is mediated by toll-like receptor 3 (TLR3) and TLR7. *J. Virol.* 84, 5627–5636. doi:10.1128/JVI.00014-10

Buchan, J. R. (2006). tRNA properties help shape codon pair preferences in open reading frames. *Nucleic Acids Res.* 34, 1015–1027. doi:10.1093/nar/gkj488

Buhr, F., Jha, S., Thommen, M., Mittelstaet, J., Kutz, F., Schwalbe, H., et al. (2016). Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell* 61, 341–351. doi:10.1016/j.molcel.2016.01.008

Bulcha, J. T., Wang, Y., Ma, H., Tai, P. W. L., and Gao, G. (2021). Viral vector platforms within the gene therapy landscape. *Signal Transduct. Target. Ther.* 6, 53. doi:10.1038/s41392-021-00487-6

Burke, P. C., Park, H., and Subramaniam, A. R. (2022). A nascent peptide code for translational control of mRNA stability in human cells. *Nat. Commun.* 13, 6829. doi:10.1038/s41467-022-34664-0

Burns, C. C., Shaw, J., Campagnoli, R., Jorba, J., Vincent, A., Quay, J., et al. (2006). Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region. *J. Virol.* 80, 3259–3272. doi:10.1128/JVI.80.7.3259-3272.2006

Cabanes-Creus, M., Ginn, S. L., Amaya, A. K., Liao, S. H. Y., Westhaus, A., Hallwirth, C. V., et al. (2019). Codon-optimization of wild-type adeno-associated virus capsid sequences enhances DNA family shuffling while conserving functionality. *Mol. Ther. - Methods Clin. Dev.* 12, 71–84. doi:10.1016/j.omtm.2018.10.016

Capell, A., Fellerer, K., and Haass, C. (2014). Progranulin transcripts with Short and long 5′ untranslated regions (UTRs) are differentially expressed via posttranscriptional and translational repression. *J. Biol. Chem.* 289, 25879–25889. doi:10.1074/jbc.M114.560128

Carbone, A., Zinovyev, A., and Képès, F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics* 19, 2005–2015. doi:10.1093/bioinformatics/btg272

Casini, A., Storch, M., Baldwin, G. S., and Ellis, T. (2015). Bricks and blueprints: methods and standards for DNA assembly. *Nat. Rev. Mol. Cell Biol.* 16, 568–576. doi:10.1038/nrm4014

Castillo-Hair, S. M., and Seelig, G. (2022). Machine learning for designing next-generation mRNA therapeutics. *Acc. Chem. Res.* 55, 24–34. doi:10.1021/acs.accounts.1c00621

Chamani Mohasses, F., Solouki, M., Ghareyazie, B., Fahmideh, L., and Mohsenpour, M. (2020). Correlation between gene expression levels under drought stress and synonymous codon usage in rice plant by in-silico study. *PLoS One* 15, e0237334. doi:10.1371/journal.pone.0237334

Chen, K. Y., Park, H., and Subramaniam, A. R. (2023). Massively parallel identification of sequence motifs triggering ribosome-associated mRNA quality control. bioRxiv, 2023.09.27.2023.09.27.559793. doi:10.1101/2023.09.27.559793

Chen, M.-W., Cheng, T.-J. R., Huang, Y., Jan, J.-T., Ma, S.-H., Yu, A. L., et al. (2008). A consensus–hemagglutinin-based DNA vaccine that protects mice against divergent H5N1 influenza viruses. *Proc. Natl. Acad. Sci.* 105, 13538–13543. doi:10.1073/pnas.0806901105

Chen, W., Li, H., Liu, Z., and Yuan, W. (2016). Lipopolyplex for therapeutic gene delivery and its application for the treatment of Parkinson's disease. *Front. Aging Neurosci.* 8, 68. doi:10.3389/fnagi.2016.00068

ClinicalTrials.gov ClinicalTrials.gov (2024).

Coughlan, L. (2020). Factors which contribute to the immunogenicity of non-replicating adenoviral vectored vaccines. *Front. Immunol.* 11, 909. doi:10.3389/fimmu.2020.00909

Courel, M., Clément, Y., Bossevain, C., Foretek, D., Vidal Cruchez, O., Yi, Z., et al. (2019). GC content shapes mRNA storage and decay in human cells. *Elife* 8, e49708. doi:10.7554/eLife.49708

Daniel, E., Onwukwe, G. U., Wierenga, R. K., Quaggin, S. E., Vainio, S. J., and Krause, M. (2015). ATGme: open-source web application for rare codon identification and custom DNA sequence optimization. *BMC Bioinforma.* 16, 303. doi:10.1186/s12859-015-0743-5

Das, S. (2017). Analysis of gene expression using modified relative codon bias strength in nanoarchaeum equitans. *Biosci. Biotechnol. Res. Asia* 14, 793–799. doi:10.13005/bbra/2510

Desai, P. N., Shrivastava, N., and Padh, H. (2010). Production of heterologous proteins in plants: strategies for optimal expression. *Biotechnol. Adv.* 28, 427–435. doi:10.1016/j.biotechadv.2010.01.005

de Sostoa, J., Fajardo, C. A., Moreno, R., Ramos, M. D., Farrera-Sal, M., and Alemany, R. (2019). Targeting the tumor stroma with an oncolytic adenovirus secreting a fibroblast activation protein-targeted bispecific T-cell engager. *J. Immunother. Cancer* 7, 19. doi:10.1186/s40425-019-0505-4

Dewi, K. S., and Fuad, A. M. (2020). Improving the expression of human granulocyte colony stimulating factor in *Escherichia coli* by reducing the GC-content and increasing mRNA folding free energy at 5'-terminal end. *Adv. Pharm. Bull.* 10, 610–616. doi:10.34172/apb.2020.073

Diez, M., Medina-Muñoz, S. G., Castellano, L. A., da Silva Pescador, G., Wu, Q., and Bazzini, A. A. (2022). iCodon customizes gene expression based on the codon composition. *Sci. Rep.* 12, 12126. doi:10.1038/s41598-022-15526-7

Dittmar, K. A., Goodenbour, J. M., and Pan, T. (2006). Tissue-specific differences in human transfer RNA expression. *PLoS Genet.* 2, e221. doi:10.1371/journal.pgen.0020221

dos Reis, M. (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. doi:10.1093/nar/gkg897

Fath, S., Bauer, A. P., Liss, M., Spriestersbach, A., Maertens, B., Hahn, P., et al. (2011). Multiparameter RNA and codon optimization: a standardized tool to assess and enhance autologous mammalian gene expression. *PLoS One* 6, e17596. doi:10.1371/journal.pone.0017596

Faust, S. M., Bell, P., Cutler, B. J., Ashley, S. N., Zhu, Y., Rabinowitz, J. E., et al. (2013). CpG-depleted adeno-associated virus vectors evade immune detection. *J. Clin. Invest.* 123, 2994–3001. doi:10.1172/JCI68205

FDA (2024).

Feng, H., Segalés, J., Wang, F., Jin, Q., Wang, A., Zhang, G., et al. (2022). Comprehensive analysis of codon usage patterns in Chinese porcine circoviruses based on their major protein-coding sequences. *Viruses* 14, 81. doi:10.3390/v14010081

Festen, E. A. M., Goyette, P., Green, T., Boucher, G., Beauchamp, C., Trynka, G., et al. (2011). A meta-analysis of genome-wide association scans identifies IL18RAP, PTPN2, TAGAP, and PUS10 as shared risk loci for crohn's disease and celiac disease. *PLoS Genet.* 7, e1001283. doi:10.1371/journal.pgen.1001283

Fox, J. M., and Erill, I. (2010). Relative codon adaptation: a generic codon bias index for prediction of gene expression. *DNA Res.* 17, 185–196. doi:10.1093/dnares/dsq012

Friberg, M., von Rohr, P., and Gonnet, G. (2004). Limitations of codon adaptation index and other coding DNA-based features for prediction of protein expression in *Saccharomyces cerevisiae*. *Yeast* 21, 1083–1093. doi:10.1002/yea.1150

Fu, H., Liang, Y., Zhong, X., Pan, Z., Huang, L., Zhang, H., et al. (2020). Codon optimization with deep learning to enhance protein expression. *Sci. Rep.* 10, 17617–17619. doi:10.1038/s41598-020-74091-z

Gao, W., Gallardo-Dodd, C. J., and Kutter, C. (2022). Cell type–specific analysis by single-cell profiling identifies a stable mammalian tRNA–mRNA interface and increased translation efficiency in neurons. *Genome Res.* 32, 97–110. doi:10.1101/gr.275944.121

Godfried Sie, C., Hesler, S., Maas, S., and Kuchka, M. (2012). IGFBP7's susceptibility to proteolysis is altered by A-to-I RNA editing of its transcript. *FEBS Lett.* 586, 2313–2317. doi:10.1016/j.febslet.2012.06.037

Gonzalez-Sanchez, B., Vega-Rodríguez, M. A., Santander-Jiménez, S., and Granado-Criado, J. M. (2019). Multi-Objective Artificial Bee Colony for designing multiple genes encoding the same protein. *Appl. Soft Comput.* 74, 90–98. doi:10.1016/j.asoc.2018.10.023

Goulet, D. R., Yan, Y., Agrawal, P., Waight, A. B., Mak, A. N., and Zhu, Y. (2023). Codon optimization using a recurrent neural network. *J. Comput. Biol.* 30, 70–81. doi:10.1089/cmb.2021.0458

Gouy, M., and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10, 7055–7074. doi:10.1093/nar/10.22.7055

Grote, A., Hiller, K., Scheer, M., Munch, R., Nortemann, B., Hempel, D. C., et al. (2005). JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.* 33, W526–W531. doi:10.1093/nar/gki376

Gu, W., Zhou, T., and Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Comput. Biol.* 6, e1000664. doi:10.1371/journal.pcbi.1000664

Hanson, G., and Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nat. Rev. Mol. Cell Biol.* 19, 20–30. doi:10.1038/nrm.2017.91

Hayat, S. M. G., Farahani, N., Safdarian, E., Roointan, A., and Sahebkar, A. (2019). Gene delivery using lipoplexes and polyplexes: principles, limitations and solutions. *Crit. Rev. Eukaryot. Gene Expr.* 29, 29–36. doi:10.1615/CritRevEukaryotGeneExpr.2018025132

Hernandez-Alias, X., Benisty, H., Radusky, L. G., Serrano, L., and Schaefer, M. H. (2023). Using protein-per-mRNA differences among human tissues in codon optimization. *Genome Biol.* 24, 34–20. doi:10.1186/s13059-023-02868-2

Herzog, R. W., Cooper, M., Perrin, G. Q., Biswas, M., Martino, A. T., Morel, L., et al. (2019). Regulatory T cells and TLR9 activation shape antibody formation to a secreted transgene product in AAV muscle gene transfer. *Cell. Immunol.* 342, 103682. doi:10.1016/j.cellimm.2017.07.012

Hia, F., Yang, S. F., Shichino, Y., Yoshinaga, M., Murakawa, Y., Vandenbon, A., et al. (2019). Codon bias confers stability to human mRNA s. *EMBO Rep.* 20, e48220. doi:10.15252/embr.201948220

Höllerer, S., and Jeschek, M. (2023). Ultradeep characterisation of translational sequence determinants refutes rare-codon hypothesis and unveils quadruplet base pairing of initiator tRNA and transcript. *Nucleic Acids Res.* 51, 2377–2396. doi:10.1093/nar/gkad040

Huang, H., Liu, Y., Liao, W., Cao, Y., Liu, Q., Guo, Y., et al. (2019). Oncolytic adenovirus programmed by synthetic gene circuit for cancer immunotherapy. *Nat. Commun.* 10, 4801. doi:10.1038/s41467-019-12794-2

Igyártó, B. Z., and Qin, Z. (2024). The mRNA-LNP vaccines – the good, the bad and the ugly? *Front. Immunol.* 15, 1336906. doi:10.3389/fimmu.2024.1336906

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.* 151, 389–409. doi:10.1016/0022-2836(81)90003-6

Ikemura, T. (1982). Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. *J. Mol. Biol.* 158, 573–597. doi:10.1016/0022-2836(82)90250-9

Irimia, M., Denuc, A., Ferran, J. L., Pernaute, B., Puelles, L., Roy, S. W., et al. (2012). Evolutionarily conserved A-to-I editing increases protein stability of the alternative splicing factor Nova1. *RNA Biol.* 9, 12–21. doi:10.4161/rna.9.1.18387

Jain, R., Jain, A., Mauro, E., LeShane, K., and Densmore, D. (2023). ICOR: improving codon optimization with recurrent neural networks. *BMC Bioinforma.* 24, 132. doi:10.1186/s12859-023-05246-8

Kames, J., Alexaki, A., Holcomb, D. D., Santana-Quintero, L. V., Athey, J. C., Hamasaki-Katagiri, N., et al. (2020). TissueCoCoPUTs: novel human tissue-specific codon and codon-pair usage tables based on differential tissue gene expression. *J. Mol. Biol.* 432, 3369–3378. doi:10.1016/j.jmb.2020.01.011

Karikó, K., Buckstein, M., Ni, H., and Weissman, D. (2005). Suppression of RNA recognition by toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. *Immunity* 23, 165–175. doi:10.1016/j.immuni.2005.06.008

Karikó, K., Muramatsu, H., Welsh, F. A., Ludwig, J., Kato, H., Akira, S., et al. (2008). Incorporation of pseudouridine into mRNA yields superior nonimmunogenic vector with increased translational capacity and biological stability. *Mol. Ther.* 16, 1833–1840. doi:10.1038/mt.2008.200

Kirchner, S., Cai, Z., Rauscher, R., Kastelic, N., Anding, M., Czech, A., et al. (2017). Alteration of protein function by a silent polymorphism linked to tRNA abundance. *PLOS Biol.* 15, e2000779. doi:10.1371/journal.pbio.2000779

Konkle, B. A., Walsh, C. E., Escobar, M. A., Josephson, N. C., Young, G., von Drygalski, A., et al. (2021). BAX 335 hemophilia B gene therapy clinical trial results: potential impact of CpG sequences on gene expression. *Blood* 137, 763–774. doi:10.1182/blood.2019004625

Kosuri, S., and Church, G. M. (2014). Large-scale *de novo* DNA synthesis: technologies and applications. *Nat. Methods* 11, 499–507. doi:10.1038/nmeth.2918

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene* 361, 13–37. doi:10.1016/j.gene.2005.06.037

Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. (2006). High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* 4, e180. doi:10.1371/journal.pbio.0040180

Kudla, G., Murray, A. W., Tollervey, D., and Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324, 255–258. doi:10.1126/science.1170160

Lee, I. T., Nachbagauer, R., Ensz, D., Schwartz, H., Carmona, L., Schaefers, K., et al. (2023). Safety and immunogenicity of a phase 1/2 randomized clinical trial of a quadrivalent, mRNA-based seasonal influenza vaccine (mRNA-1010) in healthy adults: interim analysis. *Nat. Commun.* 14, 3631. doi:10.1038/s41467-023-39376-7

Leppek, K., Byeon, G. W., Kladwang, W., Wayment-Steele, H. K., Kerr, C. H., Xu, A. F., et al. (2022). Combinatorial optimization of mRNA structure, stability, and translation for RNA-based therapeutics. *Nat. Commun.* 13, 1536. doi:10.1038/s41467-022-28776-w

LeRoy, N., and Roleck, C. (2023). Optipyzer: a fast and flexible multi-species codon optimization server. bioRxiv, 2023.05.22.541759. doi:10.1101/2023.05.22.541759

Li, C., He, Y., Nicolson, S., Hirsch, M., Weinberg, M. S., Zhang, P., et al. (2013). Adeno-associated virus capsid antigen presentation is dependent on endosomal escape. *J. Clin. Invest.* 123, 1390–1401. doi:10.1172/JCI66611

Liu, Y. (2020). A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun. Signal.* 18, 145. doi:10.1186/s12964-020-00642-6

Loomis, K. H., Kirschman, J. L., Bhosle, S., Bellamkonda, R. V., and Santangelo, P. J. (2016). Strategies for modulating innate immune activation and protein production of *in vitro* transcribed mRNAs. *J. Mater. Chem. B* 4, 1619–1632. doi:10.1039/C5TB01753J

Ma, S., Tang, N., and Tian, J. (2012). DNA synthesis, assembly and applications in synthetic biology. *Curr. Opin. Chem. Biol.* 16, 260–267. doi:10.1016/j.cbpa.2012.05.001

Malarkannan, S., Horng, T., Shih, P. P., Schwab, S., and Shastri, N. (1999). Presentation of out-of-frame peptide/MHC class I complexes by a novel translation initiation mechanism. *Immunity* 10, 681–690. doi:10.1016/S1074-7613(00)80067-9

Martino, A. T., Basner-Tschakarjan, E., Markusic, D. M., Finn, J. D., Hinderer, C., Zhou, S., et al. (2013). Engineered AAV vector minimizes *in vivo* targeting of transduced hepatocytes by capsid-specific CD8+ T cells. *Blood* 121, 2224–2233. doi:10.1182/blood-2012-10-460733

Matsuda, D., and Mauro, V. P. (2010). Determinants of initiation codon selection during translation in mammalian cells. *PLoS One* 5, e15057. doi:10.1371/journal.pone.0015057

Mendell, J. R., Al-Zaidy, S. A., Rodino-Klapac, L. R., Goodspeed, K., Gray, S. J., Kay, C. N., et al. (2021). Current clinical applications of *in vivo* gene therapy with AAVs. *Mol. Ther.* 29, 464–488. doi:10.1016/j.ymthe.2020.12.007

Mitarai, N., Sneppen, K., and Pedersen, S. (2008). Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J. Mol. Biol.* 382, 236–245. doi:10.1016/j.jmb.2008.06.068

Mueller, S. (2023). *Challenges and opportunities of mRNA vaccines against SARS-CoV-2*. Cham: Springer International Publishing. doi:10.1007/978-3-031-18903-6

Mueller, S., Papamichail, D., Coleman, J. R., Skiena, S., and Wimmer, E. (2006). Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol.* 80, 9687–9696. doi:10.1128/JVI.00738-06

Mulroney, T. E., Pöyry, T., Yam-Puc, J. C., Rust, M., Harvey, R. F., Kalmar, L., et al. (2024). N1-methylpseudouridylation of mRNA causes +1 ribosomal frameshifting. *Nature* 625, 189–194. doi:10.1038/s41586-023-06800-3

Narula, A., Ellis, J., Taliaferro, J. M., and Rissland, O. S. (2019). Coding regions affect mRNA stability in human cells. *RNA* 25, 1751–1764. doi:10.1261/rna.073239.119

Navon, S., and Pilpel, Y. (2011). The role of codon selection in regulation of translation efficiency deduced from synthetic libraries. *Genome Biol.* 12, R12. doi:10.1186/gb-2011-12-2-r12

Nieuwkoop, T., Terlouw, B. R., Stevens, K. G., Scheltema, R. A., de Ridder, D., van der Oost, J., et al. (2023). Revealing determinants of translation efficiency via whole-gene codon randomization and machine learning. *Nucleic Acids Res.* 51, 2363–2376. doi:10.1093/nar/gkad035

Núñez-Manchón, E., Farrera-Sal, M., Otero-Mateo, M., Castellano, G., Moreno, R., Medel, D., et al. (2021). Transgene codon usage drives viral fitness and therapeutic efficacy in oncolytic adenoviruses. *Nar. Cancer* 3, zcab015. doi:10.1093/narcan/zcab015

Oliver, S. E., Gargano, J. W., Marin, M., Wallace, M., Curran, K. G., Chamberland, M., et al. (2020). The advisory committee on immunization practices' interim recommendation for use of pfizer-bioNTech COVID-19 vaccine — United States, december 2020. *MMWR. Morb. Mortal. Wkly. Rep.* 69, 1922–1924. doi:10.15585/mmwr.mm6950e2

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., McQuisten, K. A., Almabrazi, H. G., et al. (2008). IDT SciTools: a suite for analysis and design of nucleic acid oligomers. *Nucleic Acids Res.* 36, W163–W169. doi:10.1093/nar/gkn198

Palluk, S., Arlow, D. H., de Rond, T., Barthel, S., Kang, J. S., Bector, R., et al. (2018). *De novo* DNA synthesis using polymerase-nucleotide conjugates. *Nat. Biotechnol.* 36, 645–650. doi:10.1038/nbt.4173

Pereira, I. T., Spangenberg, L., Robert, A. W., Amorín, R., Stimamiglio, M. A., Naya, H., et al. (2018). Polysome profiling followed by RNA-seq of cardiac differentiation stages in hESCs. *Sci. Data* 5, 180287. doi:10.1038/sdata.2018.287

Perlak, F. J., Fuchs, R. L., Dean, D. A., McPherson, S. L., and Fischhoff, D. A. (1991). Modification of the coding sequence enhances plant expression of insect control protein genes. *Proc. Natl. Acad. Sci.* 88, 3324–3328. doi:10.1073/pnas.88.8.3324

Pham, T. D., O'Connell, J., and Crane, D. I. (2004). "Constrained codon optimization by dynamic programming," in Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004 (IEEE), 153–156. doi:10.1109/ISIMP.2004.1434023

Pinkard, O., McFarland, S., Sweet, T., and Coller, J. (2020). Quantitative tRNA-sequencing uncovers metazoan tissue-specific tRNA regulation. *Nat. Commun.* 11, 4104. doi:10.1038/s41467-020-17879-x

Pitoiset, F., Vazquez, T., Levacher, B., Nehar-Belaid, D., Dérian, N., Vigneron, J., et al. (2017). Retrovirus-based virus-like particle immunogenicity and its modulation by toll-like receptor activation. *J. Virol.* 91, e01230-17. doi:10.1128/JVI.01230-17

Pizzo, L., Iriarte, A., Alvarez-Valin, F., and Marín, M. (2015). Conservation of CFTR codon frequency through primates suggests synonymous mutations could have a functional effect. *Mutat. Res. Mol. Mech. Mutagen.* 775, 19–25. doi:10.1016/j.mrfmmm.2015.03.005

Plotkin, J. B., Robins, H., and Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proc. Natl. Acad. Sci.* 101, 12588–12591. doi:10.1073/pnas.0404957101

Pouyet, F., Mouchiroud, D., Duret, L., and Sémon, M. (2017). Recombination, meiotic expression and human codon usage. *Elife* 6, e27344. doi:10.7554/eLife.27344

Presnyak, V., Alhusaini, N., Chen, Y.-H., Martin, S., Morris, N., Kline, N., et al. (2015). Codon optimality is a major determinant of mRNA stability. *Cell* 160, 1111–1124. doi:10.1016/j.cell.2015.02.029

Puigbo, P., Guzman, E., Romeu, A., and Garcia-Vallve, S. (2007). OPTIMIZER: a web server for optimizing the codon usage of DNA sequences. *Nucleic Acids Res.* 35, W126–W131. doi:10.1093/nar/gkm219

Raab, A. M., Gebhardt, G., Bolotina, N., Weuster-Botz, D., and Lang, C. (2010). Metabolic engineering of *Saccharomyces cerevisiae* for the biotechnological production of succinic acid. *Metab. Eng.* 12, 518–525. doi:10.1016/j.ymben.2010.08.005

Rahman, S. U., Yao, X., Li, X., Chen, D., and Tao, S. (2018). Analysis of codon usage bias of Crimean-Congo hemorrhagic fever virus and its adaptation to hosts. *Infect. Genet. Evol.* 58, 1–16. doi:10.1016/j.meegid.2017.11.027

Reis, M. d. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32, 5036–5044. doi:10.1093/nar/gkh834

Ringnér, M., and Krogh, M. (2005). Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.* 1, e72. doi:10.1371/journal.pcbi.0010072

Rodriguez, A., Wright, G., Emrich, S., and Clark, P. L. (2018). %MinMax: a versatile tool for calculating and comparing synonymous codon usage and its impact on protein folding. *Protein Sci.* 27, 356–362. doi:10.1002/pro.3336

Rogers, G. L., Shirley, J. L., Zolotukhin, I., Kumar, S. R. P., Sherman, A., Perrin, G. Q., et al. (2017). Plasmacytoid and conventional dendritic cells cooperate in crosspriming AAV capsid-specific CD8+ T cells. *Blood* 129, 3184–3195. doi:10.1182/blood-2016-11-751040

Rojas, M., Restrepo-Jiménez, P., Monsalve, D. M., Pacheco, Y., Acosta-Ampudia, Y., Ramírez-Santana, C., et al. (2018). Molecular mimicry and autoimmunity. *J. Autoimmun.* 95, 100–123. doi:10.1016/j.jaut.2018.10.012

Röltgen, K., Nielsen, S. C. A., Silva, O., Younes, S. F., Zaslavsky, M., Costales, C., et al. (2022). Immune imprinting, breadth of variant recognition, and germinal center response in human SARS-CoV-2 infection and vaccination. *Cell* 185, 1025–1040.e14. doi:10.1016/j.cell.2022.01.018

Ronk, A. J., Lloyd, N. M., Zhang, M., Atyeo, C., Perrett, H. R., Mire, C. E., et al. (2023). A Lassa virus mRNA vaccine confers protection but does not require neutralizing antibody in a Guinea pig model of infection. *Nat. Commun.* 14, 5603. doi:10.1038/s41467-023-41376-6

Roymondal, U., Das, S., and Sahoo, S. (2009). Predicting gene expression level from relative codon usage bias: an application to *Escherichia coli* genome. *DNA Res.* 16, 13–30. doi:10.1093/dnares/dsn029

Sabi, R., and Tuller, T. (2014). Modelling the efficiency of codon–tRNA interactions based on codon usage bias. *DNA Res.* 21, 511–526. doi:10.1093/dnares/dsu017

Sabi, R., Volvovitch Daniel, R., and Tuller, T. (2017). stAIcalc: tRNA adaptation index calculator based on species-specific weights. *Bioinformatics* 33, 589–591. doi:10.1093/bioinformatics/btw647

Sato, K., and Hamada, M. (2023). Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. *Brief. Bioinform.* 24, bbad186. doi:10.1093/bib/bbad186

Sharp, P., and Li, W.-H. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res.* 14, 7737–7749. doi:10.1093/nar/14.19.7737

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. doi:10.1093/nar/15.3.1281

Shi, F., Fan, Z., Zhang, S., Wang, Y., Tan, S., and Li, Y. (2020). Optimization of ribosomal binding site sequences for gene expression and 4-hydroxyisoleucine biosynthesis in recombinant corynebacterium glutamicum. *Enzyme Microb. Technol.* 140, 109622. doi:10.1016/j.enzmictec.2020.109622

Shirley, J. L., de Jong, Y. P., Terhorst, C., and Herzog, R. W. (2020a). Immune responses to viral gene therapy vectors. *Mol. Ther.* 28, 709–722. doi:10.1016/j.ymthe.2020.01.001

Shirley, J. L., Keeler, G. D., Sherman, A., Zolotukhin, I., Markusic, D. M., Hoffman, B. E., et al. (2020b). Type I IFN sensing by cDCs and CD4+ T cell help are both requisite for cross-priming of AAV capsid-specific CD8+ T cells. *Mol. Ther.* 28, 758–770. doi:10.1016/j.ymthe.2019.11.011

Simon, C. S., Hadjantonakis, A., and Schröter, C. (2018). Making lineage decisions with biological noise: lessons from the early mouse embryo. *WIREs Dev. Biol.* 7, e319. doi:10.1002/wdev.319

Sinyakov, A. N., Ryabinin, V. A., and Kostina, E. V. (2021). Application of array-based oligonucleotides for synthesis of genetic designs. *Mol. Biol.* 55, 487–500. doi:10.1134/S0026893321030109

Song, L.-F., Deng, Z.-H., Gong, Z.-Y., Li, L.-L., and Li, B.-Z. (2021). Large-scale *de novo* oligonucleotide synthesis for whole-genome synthesis and data storage: challenges and opportunities. *Front. Bioeng. Biotechnol.* 9, 689797. doi:10.3389/fbioe.2021.689797

Stenico, M., Lloyd, A. T., and Sharp, P. M. (1994). Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucleic Acids Res.* 22, 2437–2446. doi:10.1093/nar/22.13.2437

Sun, B., Zhang, H., Franco, L. M., Brown, T., Bird, A., Schneider, A., et al. (2005). Correction of glycogen storage disease type II by an adeno-associated virus vector containing a muscle-specific promoter. *Mol. Ther.* 11, 889–898. doi:10.1016/j.ymthe.2005.01.012

Taneda, A., and Asai, K. (2020). COSMO: a dynamic programming algorithm for multicriteria codon optimization. *Comput. Struct. Biotechnol. J.* 18, 1811–1818. doi:10.1016/j.csbj.2020.06.035

Thess, A., Grund, S., Mui, B. L., Hope, M. J., Baumhof, P., Fotin-Mleczek, M., et al. (2015). Sequence-engineered mRNA without chemical nucleoside modifications enables an effective protein therapy in large animals. *Mol. Ther.* 23, 1456–1464. doi:10.1038/mt.2015.103

Thomas, D. R., and Walmsley, A. M. (2014). Improved expression of recombinant plant-made hEGF. *Plant Cell Rep.* 33, 1801–1814. doi:10.1007/s00299-014-1658-8

Thomas, O. G., Bronge, M., Tengvall, K., Akpinar, B., Nilsson, O. B., Holmgren, E., et al. (2023). Cross-reactive EBNA1 immunity targets alpha-crystallin B and is associated with multiple sclerosis. *Sci. Adv.* 9, eadg3032–14. doi:10.1126/sciadv.adg3032

Thul, P. J., and Lindskog, C. (2018). The human protein atlas: a spatial map of the human proteome. *Protein Sci.* 27, 233–244. doi:10.1002/pro.3307

Villanueva, E., Martí-Solano, M., and Fillat, C. (2016). Codon optimization of the adenoviral fiber negatively impacts structural protein expression and viral fitness. *Sci. Rep.* 6, 27546. doi:10.1038/srep27546

Wan, J., Yang, J., Wang, Z., Shen, R., Zhang, C., Wu, Y., et al. (2023). A single immunization with core–shell structured lipopolyplex mRNA vaccine against rabies induces potent humoral immunity in mice and dogs. *Emerg. Microbes Infect.* 12, 2270081. doi:10.1080/22221751.2023.2270081

Wan, X.-F., Zhou, J., and Xu, D. (2006). CodonO: a new informatics method for measuring synonymous codon usage bias within and across genomes. *Int. J. Gen. Syst.* 35, 109–125. doi:10.1080/03081070500502967

Wayment-Steele, H. K., Kim, D. S., Choe, C. A., Nicol, J. J., Wellington-Oguri, R., Watkins, A. M., et al. (2021). Theoretical basis for stabilizing messenger RNA through secondary structure design. *Nucleic Acids Res.* 49, 10604–10617. doi:10.1093/nar/gkab764

Wei, Y., Silke, J. R., and Xia, X. (2019). An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Sci. Rep.* 9, 3184. doi:10.1038/s41598-019-39369-x

Welch, M., Villalobos, A., Gustafsson, C., and Minshull, J. (2009). You're one in a googol: optimizing genes for protein expression. *J. R. Soc. Interface* 6, S467–S476. doi:10.1098/rsif.2008.0520.focus

Wright, F. (1990). The 'effective number of codons' used in a gene. *Gene* 87, 23–29. doi:10.1016/0378-1119(90)90491-9

Wright, G., Rodriguez, A., Li, J., Milenkovic, T., Emrich, S. J., and Clark, P. L. (2022). CHARMING: harmonizing synonymous codon usage to replicate a desired codon usage pattern. *Protein Sci.* 31, 221–231. doi:10.1002/pro.4223

Wright, J. F. (2020). Quantification of CpG motifs in rAAV genomes: avoiding the Toll. *Mol. Ther.* 28, 1756–1758. doi:10.1016/j.ymthe.2020.07.006

Wu, Q., Medina, S. G., Kushawah, G., DeVore, M. L., Castellano, L. A., Hand, J. M., et al. (2019). Translation affects mRNA stability in a codon-dependent manner in human cells. *Elife* 8, e45396. doi:10.7554/eLife.45396

Wu, X., Shan, K., Zan, F., Tang, X., Qian, Z., and Lu, J. (2023). Optimization and deoptimization of codons in SARS-CoV-2 and related implications for vaccine development. *Adv. Sci.* 10, e2205445. doi:10.1002/advs.202205445

Xia, X. (2015). A major controversy in codon-anticodon adaptation resolved by a new codon usage index. *Genetics* 199, 573–579. doi:10.1534/genetics.114.172106

Xia, X. (2021). Detailed dissection and critical evaluation of the pfizer/BioNTech and Moderna mRNA vaccines. *Vaccines* 9, 734. doi:10.3390/vaccines9070734

Xue, C., Chu, Q., Zheng, Q., Jiang, S., Bao, Z., Su, Y., et al. (2022). Role of main RNA modifications in cancer: N6-methyladenosine, 5-methylcytosine, and pseudouridine. *Signal Transduct. Target. Ther.* 7, 142. doi:10.1038/s41392-022-01003-0

Yang, T. yuan, Braun, M., Lembke, W., McBlane, F., Kamerud, J., DeWall, S., et al. (2022). Immunogenicity assessment of AAV-based gene therapies: an IQ consortium industry white paper. *Mol. Ther. - Methods Clin. Dev.* 26, 471–494. doi:10.1016/j.omtm.2022.07.018

Yew, N. S., and Cheng, S. H. (2004). Reducing the immunostimulatory activity of CpG-containing plasmid DNA vectors for non-viral gene therapy. *Expert Opin. Drug Deliv.* 1, 115–125. doi:10.1517/17425247.1.1.115

Zhang, H., Zhang, L., Lin, A., Xu, C., Li, Z., Liu, K., et al. (2023). Algorithm for optimized mRNA design improves stability and immunogenicity. *Nature* 621, 396–403. doi:10.1038/s41586-023-06127-z

Zhang, Z., Li, J., Cui, P., Ding, F., Li, A., Townsend, J. P., et al. (2012). Codon Deviation Coefficient: a novel measure for estimating codon usage bias and its statistical significance. *BMC Bioinforma.* 13, 43. doi:10.1186/1471-2105-13-43

Zuker, M. (1994). "Prediction of RNA secondary structure by energy minimization," in *Computer analysis of sequence data* (Totowa, NJ: Humana Press), 267–294. doi:10.1385/0-89603-276-0:267

Zuker, M., and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* 9, 133–148. doi:10.1093/nar/9.1.133