Molecular Vision 2021; 27:233-242 <http://www.molvis.org/molvis/v27/233>
Received 3 June 2020 | Accepted 6 May 2021 | Published 8 May 2021

An analysis of the Kozak consensus in retinal genes and its relevance to gene therapy

Michelle E. McClements, Anum Butt, Elena Piotter, Caroline F. Peddle, Robert E. MacLaren

Nuffield Laboratory of Ophthalmology, University of Oxford, Oxford, UK,

Purpose: The classic Kozak consensus is a critical genetic element included in gene therapy transgenes to encourage the translation of the therapeutic coding sequence. Despite optimizations of other transgene elements, the Kozak consensus has not yet been considered for potential tissue-specific sequence refinement. We screened the −9 to −1 region relative to the AUG start codon of retina-specific genes to identify whether a Kozak consensus that is different from the classic sequence may be more appropriate for inclusion in gene therapy transgenes that treat inherited retinal disease.
Methods: Sequences for 135 genes known to cause nonsyndromic inherited retinal disease were extracted from the NCBI database, and the −9 to −1 nucleotides were compared. This panel was then refined to 75 genes with specific retinal functions, for which the −9 to −1 nucleotides were placed in front of a GFP transcript sequence and RNAfold predictions performed. These were compared with a GFP sequence with the classic Kozak consensus (GCCGCCACC), and sequences from retinal genes with minimum free energy (MFE) predictions greater than the reference sequence were selected to generate an optimized Kozak consensus sequence. The original Kozak consensus and the refined retina Kozak consensus were placed upstream of the *Renilla* luciferase coding sequence, which were used to transfect retinoblastoma cell lines Y-79 and WERI-RB-1 and HEK 293T/17 cells.
Results: The nucleotide frequencies of the original panel of genes were determined to be comparable to the classic Kozak consensus. RNAfold analysis of a GFP transcript with the classic Kozak sequence in the 5′ untranslated region (UTR) generated an MFE prediction of −503.3 kcal/mol. RNAfold analysis was then performed with a GFP transcript containing each −9 to −1 Kozak sequence of 75 retinal genes. Thirty-eight of the 75 genes provided a greater MFE value than −503.3 kcal/mol and exhibited an absence of stable secondary structures before the AUG codon. The −9 to −1 nucleotide frequencies of these genes identified a Kozak consensus of ACCGAGACC, differing from the classic Kozak consensus at positions −9, −5, and −4. Applying this sequence to the GFP transcript increased the MFE prediction to −500.1 kcal/mol. The newly identified retina Kozak sequence was also applied to *Renilla* luciferase plus the *REP1* and *RPGR* transcripts used in current clinical trials. In all examples, the predicted transcript MFE score increased when compared with the current transcript sequences containing classic Kozak consensus sequences. In vitro transfections identified a 7%–9% increase in Renilla activity when incorporating the optimized Kozak sequence.
Conclusions: The Kozak consensus is a critical element of eukaryotic genes; therefore, it is a required feature of gene therapy transgenes. To date, the classic sequence of GCCRCC (−6 to −1) has typically been incorporated in gene therapy transgenes, but the analysis described here suggests that, for vectors targeting the retina, using a Kozak consensus derived from retinal genes can provide increased expression of the target product.

The Kozak consensus is named after the extensive investigations of Marilyn Kozak, who first identified the significance of the nucleotides immediately upstream of the start codon [1]. Critical for guiding ribosomes in their identification of where to begin translation from eukaryotic transcripts, a consensus sequence common to all vertebrate genes from positions −9 to −1 was determined to be GCCGCCRCC in 1987 [2]. However, despite being called the Kozak consensus, the extent of conservation in vertebrate genes was considered to be low, at about 0.2% [3]. Despite this, it was clear that the identified Kozak consensus provided good translation efficiency compared to other sequence versions [4]. The −6 to −1 sequence of GCCRCC was defined as a strong Kozak sequence, providing efficient translation, with other variants, such as UAAACC, considered to be weak and to provide less efficient translation.

A more recent larger analysis of 10,012 human genes confirmed the preferred human Kozak sequence as GCCGCCRMC [5]. This was later supported by an analysis of 32,526 human genes that determined a consensus of GCCGCCACC [6]. This study also showed that vertebrate species-specific variations in the Kozak consensus influenced expression efficiencies within a Zebrafish model. The researchers identified a Zebrafish-specific Kozak consensus that showed a twofold increase in translation efficiency over the classic Kozak sequence, despite differing by just two nucleotides. If such small changes are important between species, it is

Correspondence to: Robert E MacLaren, Nuffield Laboratory of Ophthalmology, Department of Clinical Neurosciences, University of Oxford, Oxford, UK; Phone: 01865 234796; FAX: 01865 228974; email: maclaren@eye.ox.ac.uk

plausible to consider that there may also be tissue-specific biases within species.

Detailed investigations have revealed that single-nucleotide variations can alter the strength of a given Kozak sequence [7]. It has been considered that weak Kozak sequences could be important for the regulation of translation, and indeed, it has been suggested that mRNAs with weak Kozak sequences are enriched for genes involved in neurobiology in *Drosophila* [8]. However, to date, we are unaware of any human gene tissue-specific analysis of Kozak consensus sequences having been performed.

Mutations in the human Kozak sequence have been identified and associated with disease [9-13]. Of 275,716 current mutation entries in the Human Gene Mutation Database (HGMD, accessed March 2020), 4,575 were in regulatory regions, and of these, 2,695 were upstream of the AUG initiation codon, with 84 presenting evidence of disease-associated influence at nucleotides −9 to −1.

It is clear that nucleotide variations in the Kozak sequence can influence the efficiency of translation. Transgenes for gene therapy have been optimized in various ways, including capsid and promoter selections [14] and the addition of untranslated regions that improve transcript stability, such as a Woodchuck posttranscriptional regulatory element (WPRE) [15]. Inclusion of introns between the transcriptional start site and Kozak sequence has also proven to be beneficial in improving expression levels from transgenes [16,17]; therefore, it seems rational to consider the Kozak consensus for refinement in therapeutic transgenes. The more efficient a transgene is at generating the required therapeutic product in the target cell type, the lower the viral load that should be required to achieve a therapeutic outcome. Researchers developing synthetic promoters have considered the potential for a Kozak sequence that is not necessarily the classic consensus. For example, in yeast, single point mutations to adenine at position −5 improved translational strength, whereas changes to guanine elsewhere reduced translational strength [18]. It was also identified that the Kozak sequence that provided the highest predicted minimum free energy (MFE) for a given transcript also provided the most protein synthesis. In this study, we systematically screened genes with known retinal functions to identify an optimized variant of the Kozak consensus that might provide translational benefits, and therefore, could be implemented as an enhanced element in transgenes for retinal gene therapy.

METHODS

Kozak sequence comparisons: Gene sequences were downloaded from the NCBI database using Geneious 10.2.6 in March 2020, and nucleotides −9 to −1 were extracted (relative to the "A" of the AUG start codon). Data were exported to Microsoft Excel, and nucleotide preferences at each position were determined for 135 genes linked to inherited retinal disease but not syndromic disorders (RetNet). Of this panel, 75 genes identified as having retina-specific isoform functions were extracted for nucleotide frequency comparisons and RNAfold analysis.

RNAfold comparisons: The RNAfold web server was used to analyze transcript sequences identical but for the Kozak −9 to −1 nucleotides. To directly compare the influence of different Kozak sequences on mRNA secondary structure predictions, a GFP transcript was used as the reference sequence. The −9 to −1 Kozak of 75 retinal genes were inserted in this GFP transcript, and the RNAfold predictions generated and compared.

Construct preparation: Primers were designed to create the original Kozak consensus (K SDM FW AAT ACG ACT CAC TAT AGG GCC GCC ACC ATG GCT TCC AAG G and K SDM RV TGG AAG CCA TGG TGG CGG CCC TAT AGT GAG TCG TAT TAA G) or the retina-specific Kozak consensus (rK SDM FW AAA CGA CTC ACT ATA GGA CCG AGA CCA TGG CTT CCA AGG and rK SDM RV TGG AAG CCA TGG TCT CGG TCC TAT AGT GAG TCG TAT TAA G) by site-directed mutagenesis (Agilent Technologies LDA UK Limited, Stockport, UK) in the Psi-CHECK2 dual-luciferase construct (Promega Corporation, Southampton, UK). Endotoxin-free preparations of plasmid were generated (QIAGEN Ltd, Manchester, UK), and both constructs were sequence verified. Prior to transfection, each sample was diluted to 200 ng/µl, the concentration of which was confirmed using a Nanodrop One (Labtech, Heathfield, UK).

In vitro transfections: Short tandem repeat (STR) profiling for cell line authentication of HEK 293T/17, Y-79, and WERI-RB-1 cell lines was achieved before performing the luciferase assays (Eurofins Genomics, Wolverhampton, UK). Corning white costar 96-well plates (Corning Optical Communications, Flintshire, UK) were coated with 0.2 mg/ml of poly-D-lysine hydrobromide (Merck Life Science UK Limited, Gillingham, UK) for 5 min at 37 °C, and the wells were rinsed with water before applying 0.05 mg/ml of human fibronectin (Merck Life Science) and incubating for 30 min at room temperature. After removal of the human fibronectin solution, plates were air dried for 2 h. HEK 293T/17 cells were seeded at 2E+05 cells/ml in no phenol red high glucose DMEM, pyruvate and FBS purchased from Life Technologies

Ltd, Paisley, UK, L-glutamine and pen&strep purchased from Merck Life Science, 10% fetal bovine serum (FBS), and 1% penicillin and streptomycin. Y-79 and WERI-RB-1 cells were seeded at 5E+05 cells/ml in RPMI 1640 (Life Technologies Ltd, Paisley, UK) 1640 with HEPES and $NHCO_3$ supplemented with 1% L-glutamine, 1% penicillin and streptomycin, and 10% FBS. A transfection mix of Opti-MEM (Life Technologies Ltd), ViaFect Transfection Reagent (Promega), and 400 ng of plasmid was prepared per well and applied immediately after cell seeding. Samples were incubated for 72 h at 37 °C and 5% CO_2.

Luciferase assays: The Dual-Glo luciferase kit (Promega) was used following the manufacturer guidelines. Briefly, 50 µl of media was removed from each well, and 60 µl of Dual-Glo Solution was added. Samples were left to incubate at room temperature for 30 min, after which the luminescence of the control luciferase (firefly) was determined. Then, 60 µl of Dual Stop&Glo Solution was added to each well, and the samples were left to incubate at room temperature for 30 min. The luminescence of the test luciferase (Renilla) was then determined. All luminescence readings were taken with a FLUOstar Omega device (BMG Labtech, Aylesbury, UK). Data are presented as levels of Renilla activity relative to firefly activity.

Statistics: Chi-square tests were performed to compare nucleotide frequencies using previously published human nucleotide frequency data from 32,526 human genes [6] as the reference values. Normal distribution of all luciferase assay data was confirmed (by Anderson-Darling, D'Agostino & Pearson, Shapiro–Wilk, and Kolmogorov–Smirnov tests, Appendix 1); a two-way analysis of variance (ANOVA) with multiple comparisons was performed on the dataset.

RESULTS

Nucleotide frequencies at positions −9 to −1 of genes that cause inherited retinal disease: Sequences for 135 genes with known retina-specific functions or causes of inherited retinal disease (but not syndromic disorders) as listed on RetNet were extracted for comparison of nucleotides at positions −9 to −1 (Appendix 2). The nucleotide frequencies at each position were determined (Table 1, Figure 1A) and compared with the preferences reported by analysis of 32,526 human gene sequences [6]. This analysis revealed comparable nucleotide frequencies for the 135 genes associated with inherited retinal disease and the classic Kozak consensus (Table 1). Only position −3 showed a significant variation in nucleotide preference between our 135 gene panel and the reference dataset (p<0.011). This did not reflect a change in the dominant nucleotide or the order of nucleotide preference of A>G>C>T at this position; instead, it reflected a difference in the frequency of the pattern of these nucleotides.

As the original panel contained genes with functions in other cell types, and our interest was investigating an appropriate Kozak consensus for use in retinal gene therapy

Table 1. Nucleotide frequencies at positions −9 to −1 relative to the start codon for 135 genes that cause inherited retinal disease.

	-9	-8	-7	-6	-5	-4	-3	-2	-1
Nucleotide frequencies for 135 genes causing non-syndromic inherited retinal disease									
A	25	16	20	16	16	24	50	29	18
G	32	26	26	42	30	30	40	19	27
C	21	38	36	27	37	35	8	45	49
T	22	21	18	14	18	10	2	7	6
Preference:	G	C	C	G	C	C	A	C	C
Nucleotide frequencies previously identified for 32,526 human genes									
A	20	20	21	21	19	24	46	29	19
G	35	29	28	39	30	26	37	20	28
C	27	33	32	23	32	38	10	38	45
T	18	18	19	17	19	12	7	13	8
Preference:	G	C	C	G	C	C	A	C	C
Chi-square:	0.254	0.457	0.847	0.404	0.674	0.757	0.011	0.066	0.769

The nucleotide frequencies for 135 genes listed in Appendix 1 that cause inherited retinal disease are compared to the nucleotide frequencies of 32,526 human genes previously published [6].

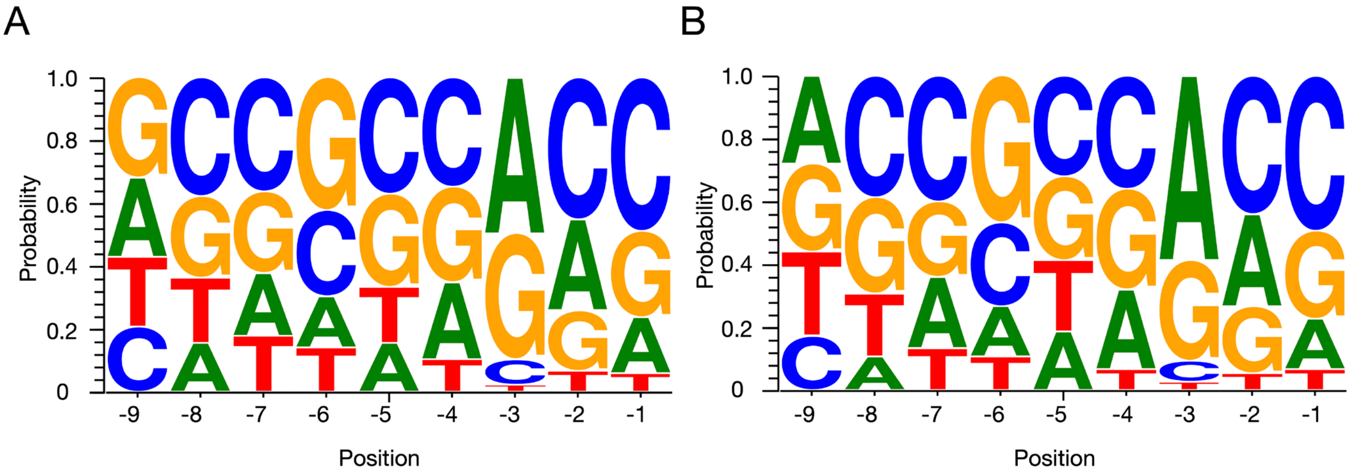

Figure 1. Representations using WebLogo 3.7.4 of nucleotide frequencies at positions −9 to −1 relative to the start codon. **A**: Following analysis of a 135 inherited retinal disease gene panel. **B**: Following analysis of a refined 75-gene panel with retina-specific roles.

vectors, the panel was refined to include only genes for which roles in other cell types are not known. This was achieved by investigating each gene in the original panel using the human gene database GeneCards. Genes for which only retinal roles (in any cell type) are currently known were selected for subgroup analysis, providing a 75-gene panel. Nucleotide frequency analysis of this refined gene panel maintained the Kozak consensus identified from the original panel of 135 genes (Figure 1B, Appendix 3). The only difference was at position −9; however, the nucleotide preference at this position was marginal in both panels.

RNAfold comparisons: It has previously been shown that mRNA transcripts with more efficient Kozak sequences provide greater MFE values [18,19]. To directly compare the influence of each Kozak sequence of the 75-gene panel with retina-specific roles on MFE, a reference mRNA transcript was generated. This was derived from a GFP reporter transgene (Figure 2), with the reference sequence containing the classic Kozak (GCCGCCACC) determined to have an MFE of −503.3 kcal/mol (Figure 3A). Each Kozak sequence from the 75-gene panel with retina-specific roles was applied in the GFP transcript, and the MFE was determined (Appendix 3).

Identification of a retina Kozak consensus: Of the 75 genes with retina-specific roles, 38 generated a transcript MFE greater than the reference sequence containing the classic Kozak consensus. Because transcripts with Kozak sequences that provide greater MFE values have been associated with more efficient translation rates, the −9 to −1 sequences for these 38 genes were extracted (Appendix 4). The nucleotide frequencies for this refined panel of genes were determined (Table 2, Figure 4). At all positions except −8 and −1, the nucleotide frequencies of the 38 genes with retina-specific roles were significantly different (p<0.05) from the classic Kozak consensus, and the nucleotide preference was changed at three of the nine positions. At the other six nucleotide positions, the strength of the nucleotide preference increased compared with the classic reference frequencies. For example, the differences in −6 nucleotide frequencies between our

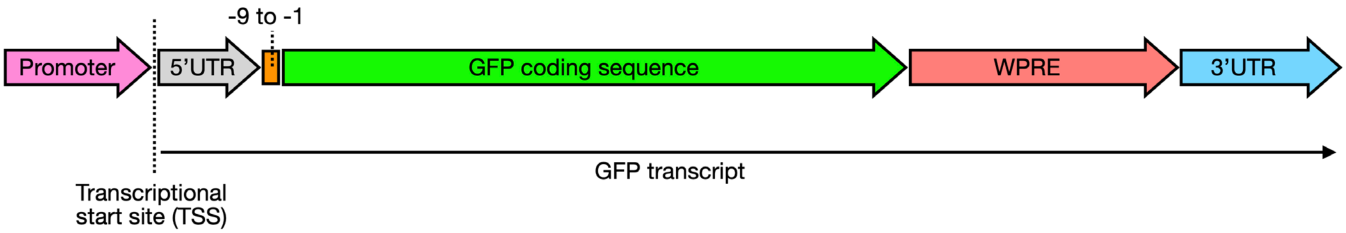

Figure 2. Depiction of the reference GFP transcript. A standard GFP transcript sequence was used to compare the influence of the -9 to -1 nucleotides of 75 genes with known retina-specific functions on mRNA secondary structure minimum free energy (MFE) predictions using RNAfold (listed in Appendix 2). UTR = untranslated region; GFP = green fluorescent protein; WPRE = woodchuck hepatitis virus post-transcriptional regulatory element.

data set and the reference panel were highly significant (p<0.00001), yet the preference for guanine did not change; rather, the strength of the preference increased while the use of thymine at this position decreased in our refined panel.

Our analysis identified a potential optimized retina Kozak consensus of ACCGAGACC (Figure 4). When this −9 to −1 consensus was applied to the GFP transcript sequence, the MFE increased from −503.3 kcal/mol containing the classic Kozak sequence, to −500.1 kcal/mol and was predicted to remove a stem-loop structure immediately upstream of the AUG start codon (Figure 3B).

Our research group has reported on two ongoing clinical trials delivering coding sequences for Rab escort protein 1 (REP1) [20] and retinitis pigmentosa GTPase regulator (RPGR) [21], and the transgenes for these two vectors contain different versions of the classic Kozak consensus. The *REP1* transgene uses the original *REP1 cDNA* [22] with the −9 to −1 sequence GGCGGCACC, and the predicted transcript has an MFE of −819.1 kcal/mol. When the Kozak sequence was changed to the optimized retina Kozak of ACCGAGACC, the MFE for this transcript increased to −812.1 kcal/mol. In contrast, the *RPGR* clinical trial transgene contains the Kozak sequence GGGGCCACC [23], and the MFE for the predicted transcript was determined to be −1090.37 kcal/mol. When the retina Kozak identified above was applied in the RPGR transcript, the MFE was also increased to −1086.67 kcal/mol. Given the sizes of these clinical trial transcript sequences, it was interesting to observe that changing the sequence of −9 to −1 nucleotides in the 5′ untranslated region (UTR) was predicted to change the MFE in both cases, although by a limited amount. When considering the differences in Kozak consensus between these vectors, it should be noted that REP1 is expressed ubiquitously, whereas RPGR is expressed in rod and cone photoreceptors only.

In vitro assessment of the optimized Kozak consensus: As the identification of an optimized Kozak consensus from genes with specific roles in the retina was derived from human sequences, testing of this sequence would be most relevant

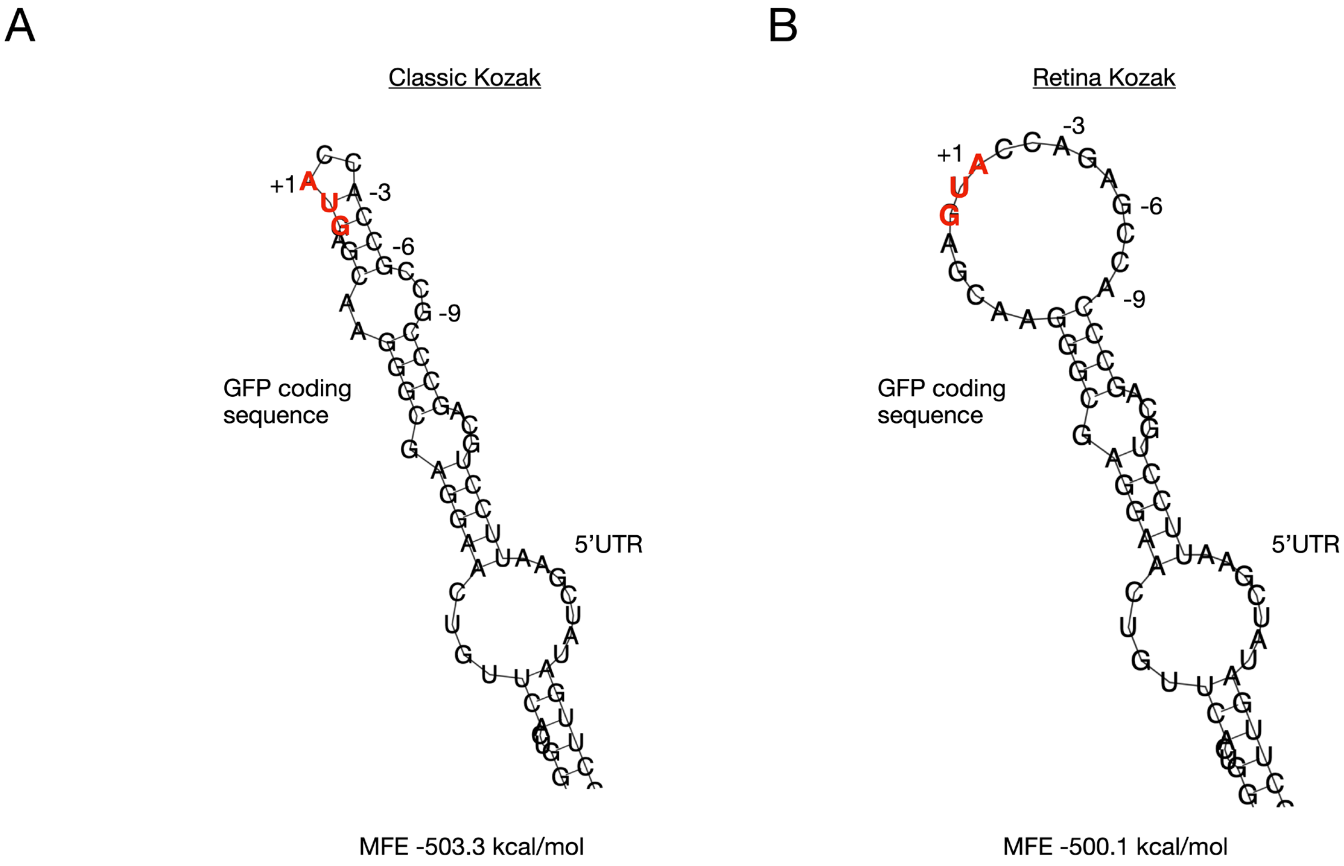

Figure 3. Secondary structure predictions at the nucleotides immediately surrounding the start codon of a GFP transcript. RNAfold assessments of a standard GFP transcript were performed as follows: **A**: with the classic Kozak consensus, and **B**: with the retina-derived Kozak consensus.

Table 2. Nucleotide frequencies at positions −9 to −1 relative to the start codon for 38 genes with retina-specific functions.

Nucleotide frequencies for 38 retina-specific genes following MFE predictions									
	-9	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**
A	34	18	32	29	37	24	74	37	18
G	21	21	16	42	18	39	18	13	29
C	21	42	39	26	24	29	5	42	47
T	24	18	13	3	21	8	3	8	5
Preference:	A	C	C	G	A	G	A	C	C
Nucleotide frequencies previously identified for 32,526 human genes									
A	20	20	21	21	19	24	46	29	19
G	35	29	28	39	30	26	37	20	28
C	27	33	32	23	32	38	10	38	45
T	18	18	19	17	19	12	7	13	8
Preference:	G	C	C	G	C	C	A	C	C
Chi-square:	0.0004	0.1636	0.0007	<0.00001	0.0002	0.0226	<0.00001	0.0303	0.6615

The nucleotide frequencies for 38 genes listed in Appendix 3 with known retina-specific functions compared to the nucleotide frequencies of 32,526 human genes previously published [6].

in human cells. To this end, the retinoblastoma cell lines Y-79 and WERI-RB-1 were selected for an in vitro comparison of the original Kozak consensus with the optimized retina Kozak consensus, along with HEK 293T/17 cells.

STR profiling confirmed the identity of these cell lines, and to provide the most sensitive output, a dual-luciferase assay was designed using the Psi-CHECK2 construct. In this plasmid, firefly luciferase was driven by an identical expression cassette in both constructs, and the *Renilla* luciferase expression cassette differed only in the −9 to −1 Kozak consensus (Figure 5A). RNAfold predictions of the original Kozak-Renilla and retina Kozak-Renilla transcripts provided an MFE of −472.20 kcal/mol and −471.80 kcal/mol, respectively, an increase in MFE that aligned with previous transcript predictions described above.

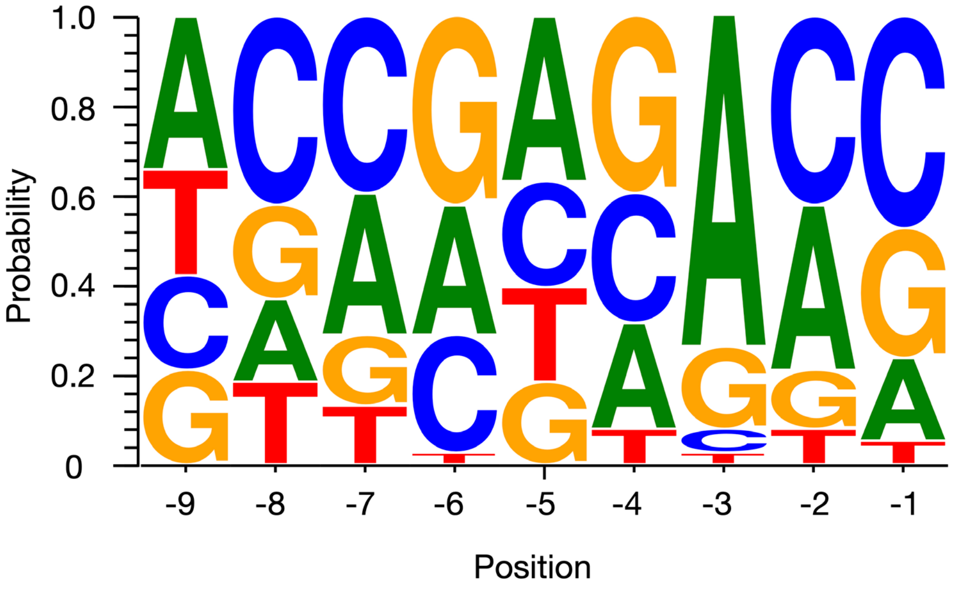

Figure 4. The optimised Kozak consensus achieved after refining the retina gene panel. Nucleotide frequencies using WebLogo 3.7.4 were achieved for positions -9 to -1 relative to the start codon. A standard GFP transcript sequence was used to compare the influence of the -9 to -1 nucleotides of 38 genes with known retina-specific roles. The identified preferential -9 to -1 nucleotide consensus of ACCGAGACC is in contrast to the classic consensus of GCCGCCACC (as in Figure 1).

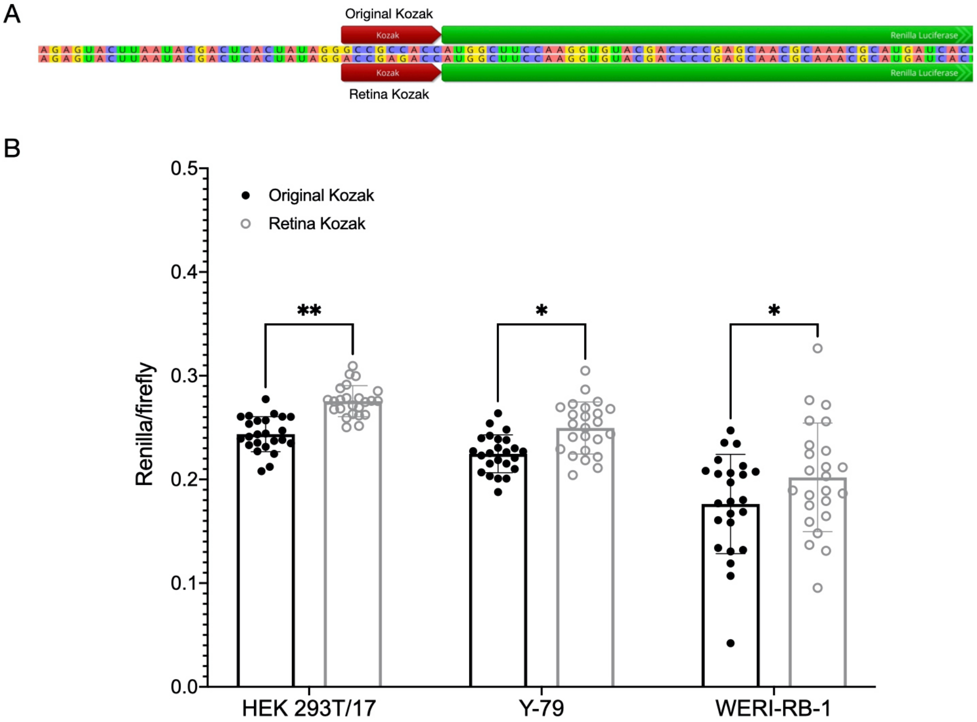

Figure 5. Improved luciferase activity when using an optimized Kozak consensus. **A**: HEK 293T/17, Y-79, and WERI-RB-1 cells were transfected with dual-luciferase constructs that were identical but for the Kozak consensus of the Renilla luciferase. One construct contained the original Kozak consensus (GCCGCCACC) and the other the retina-derived Kozak consensus (ACCGAGACC) at positions −9 to −1 of the Renilla coding sequence. The remaining construct sequence, including the firefly expression cassette, was identical for both plasmids. **B**: Luciferase activity was assessed 48 h post-transfection with Renilla luciferase activity presented relative to firefly activity. Data were confirmed to be normally distributed (Appendix 1), and a two-way analysis of variance (ANOVA) identified a significant influence of the Kozak sequence on Renilla luciferase activity ($p<0.0001$, n = 24 from four independent biological replicates with six technical replicates of each). **$p<0.004$, *$p<0.03$. Error bars represent the standard error of the mean (SEM).

Despite the small change in MFE prediction for the Renilla transcripts, the Kozak consensus change influenced the levels of Renilla activity in vitro ($p<0.0001$, Figure 5B). The optimized retina Kozak consensus provided approximately 7%–9% more Renilla activity relative to firefly in all three cell lines (9% in Y-79, $p = 0.0289$; 6.9% in WERI-RB-1, $p = 0.0234$; 7.6% in HEK 293T/17, $p = 0.0032$). It is worth noting that the transfection efficiencies were not saturated in any of the cell lines (Appendix 5). The retinoblastoma cell lines did not transfect as efficiently as HEK 293T/17 cells did, where the latter consistently achieved transfection rates of 80%–90%. This was compared with 50%–60% of Y-79 cells and 20%–30% of WERI-RB-1 cells. Furthermore, the transfection rates of each cell line were consistent between replicates, and luciferase expression levels were detectable with the Dual-Glo assay. Hence, when tested in vitro using a reliable plasmid assay, the slightly modified Kozak consensus derived from retina-specific genes led to small but consistent increases in translation in three independent human cell lines of retinal and neuronal lineage.

DISCUSSION

The Kozak sequence has been known as an important feature for achieving efficient translation for decades following the pioneering work of Marilyn Kozak [4]. A Kozak consensus for vertebrate genes was identified in 1987 and has since been confirmed as being GCCGCCRCC for positions −9 to −1 relative to the AUG initiation codon in human genes [2,5,6]. Variations in this sequence have been shown to influence the efficiency of translation [7] and cause human disease [9-13].

It has been identified in *Drosophila* that transcripts with weak Kozak sequences are enriched in neuron-related genes, indicating an important role for a neuron-related Kozak consensus that varies from other genes [8]. That study indicated that the impact of the Kozak sequence on translation efficiency is related to the elongation and initiation rates of specific cell types, suggesting that, for a given cell or tissue type, the preferred Kozak arrangement may differ from the classic Kozak consensus. Given the inclusion of a Kozak sequence in all gene therapy transgenes, it seems unusual not to consider the potential optimization of this sequence when designing tissue-specific vectors. Critical features of retinal gene therapy vectors are that they transduce the cell type of interest and efficiently produce the desired therapeutic product [14]. The more optimized and efficient a vector is, the lower the dose required to achieve a therapeutic outcome while avoiding the potential for toxic responses from an increased number of vector particles [24].

Many aspects of transgene design have already been considered and optimized, such as refinement of capsid choices; cell-specific promoters [14], including synthetic promoters [25]; and refined regulatory elements that enhance transcript stability and translation efficiency [15,16]. Given that vectors for gene therapy have undergone many types of optimization to ensure they are as efficient as possible, in this study, we considered the Kozak sequence for the first time. To further investigate transgene design for retinal gene therapy, we screened retinal genes to identify whether a different Kozak consensus should be considered for incorporation into transgenes for the treatment of inherited retinal disease. Using in silico tools, the data presented here identified a prediction for an efficient Kozak consensus derived from genes with retina-specific roles. The sequence we propose is based on RNAfold predictions and the link previously shown between greater MFE values and translation efficiency [18,19]. It has previously been suggested that such secondary structures in the 5′UTR impede translation initiation [26,27]. Our secondary structure predictions indicated that our refined retina Kozak sequence removed stable secondary structure formation in the 5′UTR, although it should be noted that the MFE may not always represent the native conformation [28], and this form of prediction is limited [29]. Indeed, some studies have shown no significant relationship between predicted mRNA folding energy and translation efficiency [30]. However, in this study, we showed that implementation of an optimized Kozak consensus could generate more luciferase activity compared with an identical construct containing the original Kozak consensus. The optimized Kozak consensus was derived from genes with retina-specific roles with the intention of enhancing transgenes for retinal gene therapy, and improvements from this consensus were evident in human retinoblastoma cell lines. However, the RNAfold predictions identified an increase in MFE that could provide a translational benefit regardless of the cell type, which proved to be the case when we found that an improvement was also achieved in HEK 293T/17 cells. It may be that, were the assessment process described here applied to other genes with tissue-specific functions, a similar optimized Kozak consensus could also be achieved.

In this study, for the first time, we considered the potential for implementing a retina-derived Kozak consensus for gene therapy vectors used in the treatment of inherited retinal disease. We identified a new sequence that differs in three of the nine nucleotide positions, providing an enhancement over the classic Kozak consensus currently used in retinal gene therapy vectors.

APPENDIX 1. NORMALITY ASSESSMENTS OF THE LUCIFERASE ASSAY DATA.

To access the data, click or select the words "Appendix 1." All samples from the luciferase assay experiments were confirmed to be normally distributed using four tests (Anderson-Darling, D-Agostino & Pearson, Shapiro-Wilk and Kolmogorov–Smirnov).

APPENDIX 2. 135 GENES ASSOCIATED WITH INHERITED RETINAL DISEASE.

To access the data, click or select the words "Appendix 2" The −9 to −1 nucleotides for 135 genes associated with inherited retinal diseases are shown.

APPENDIX 3. −9 TO −1 NUCLEOTIDE SEQUENCES OF 75 GENES WITH RETINA-SPECIFIC FUNCTIONS.

To access the data, click or select the words "Appendix 3." The −9 to −1 nucleotides for each gene were inserted in a reference GFP transcript from which the minimum free energy (MFE) was determined using RNAFold. The classic −9 to −1 consensus achieved an MFE of −503.3 kcal/mol.

APPENDIX 4. −9 TO −1 NUCLEOTIDE SEQUENCES OF 38 GENES WITH RETINA-SPECIFIC FUNCTIONS.

To access the data, click or select the words "Appendix 4." The −9 to −1 nucleotides for each gene were inserted in a reference GFP transcript and the minimum free energy (MFE) determined using RNAFold for these genes was greater than that of the classic −9 to −1 consensus.

APPENDIX 5. EXAMPLE HEK 293T/17, Y-79 AND WERI-RB-1 CELL IMAGES.

To access the data, click or select the words "Appendix 5." Cells were healthy and confluent 72 h post-transfection. A pCAG.GFP plasmid was used as a positive transfection control, revealing consistent transfection efficiencies of 80%–90% for HEK 293T/17 cells, 50%–60% for Y-79 cells and 20%–30% for WERI-RB-1 cells. These were consistent between technical and biological replicates.

ACKNOWLEDGMENTS

Funding was provided by the NIHR Oxford Biomedical Research Centre, Fight for Sight UK, Retina UK, Wellcome Trust Medical and Life Sciences Translational Fund and the Royal College of Surgeons of Edinburgh. Author contributions: conceptualisation, M.E.M, R.E.M; methodology, M.E.M; formal analysis, M.E.M; investigation, M.E.M, A.B, E.P, C.P; writing - original draft, M.E.M; writing - review & editing, M.E.M, A.B, E.P, C.P, R.E.M; supervision, R.E.M; funding acquisition, R.E.M.

REFERENCES

1. Kozak M. Point mutations close to the AUG initiator codon affect the efficiency of translation of rat preproinsulin in vivo. Nature. Nature Publishing Group 1984; 308:241-6. [PMID: 6700727].
2. Kozak M. An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res 1987; 15:8125-48. [PMID: 3313277].
3. Cavener DR, Ray SC. Eukaryotic start and stop translation sites. Nucleic Acids Res 1991; 19:3185-92. [PMID: 1905801].
4. Kozak M. At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. J Mol Biol 1987; 196:947-50. [PMID: 3681984].
5. Nakagawa S, Niimura Y, Gojobori T, Tanaka H, Miura K-I. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. Nucleic Acids Res 2008; 36:861-71. [PMID: 18086709].
6. Grzegorski SJ, Chiari EF, Robbins A, Kish PE, Kahana A. Natural Variability of Kozak Sequences Correlates with Function in a Zebrafish Model. Neuhauss SC, editor. PLoS ONE. 2014 Sep 23;9:1–6.
7. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell 1986; 44:283-92. [PMID: 3943125].
8. Acevedo JM, Hoermann B, Schlimbach T, Teleman AA. Changes in global translation elongation or initiation rates shape the proteome via the Kozak sequence. Sci Rep. Nature Publishing Group 2018; 8:1-12. .
9. Afshar-Kharghan V, Li CQ, Khoshnevis-Asl M, López JA. Kozak sequence polymorphism of the glycoprotein (GP) Ibalpha gene is a major determinant of the plasma membrane levels of the platelet GP Ib-IX-V complex. Blood 1999; 94:186-91. [PMID: 10381512].
10. Kozak M. Emerging links between initiation of translation and human diseases. Mamm Genome 2002; 8:401-10. .
11. Kanaji T, Okamura T, Osaki K, Kurowia M, Shimoda K, Hamasaki N, Niho Y. A Common Genetic Polymorphism (46 C to T Substitution) in the 5′-Untranslated Region of the Coagulation Factor XII Gene is Associated With Low Translation Efficiency and Decrease in Plasma Factor XII Level. Blood 1998; 91:2010-4. [PMID: 9490684].
12. Jacobson EM, Concepcion E, Oashi T, Tomer Y. Graves' Disease-Associated Kozak Sequence Single-Nucleotide Polymorphism Enhances the Efficiency of CD40 Gene Translation: A Case for Translational Pathophysiology | Endocrinology | Oxford Academic. Endocrinology 2005; 146:2684-91. [PMID: 15731360].
13. Wolf A, Caliebe A, Thomas NST, Ball EV, Mort M, Stenson PD, Krawczak M, Cooper DN. Single base-pair substitutions at the translation initiation sites of human genes as a cause of inherited disease. Hum Mutat 2011; 32:1137-43. .
14. McClements ME, MacLaren RE. Gene therapy for retinal disease. Transl Res 2013; 161:241-54. [PMID: 23305707].
15. Patrício MI, Barnard AR, Orlans HO, McClements ME, MacLaren RE. Inclusion of the Woodchuck Hepatitis Virus Posttranscriptional Regulatory Element Enhances AAV2-Driven Transduction of Mouse and Human Retina. Mol Ther Nucleic Acids 2017; 6:198-208. [PMID: 28325286].
16. Nott A, Meislin SH, Moore MJ. A quantitative analysis of intron effects on mammalian gene expression. RNA 2003; 9:607-17. [PMID: 12702819].
17. McClements ME, Barnard AR, Singh MS, Charbel Issa P, Jiang Z, Radu RA, MacLaren RE. An AAV dual vector strategy ameliorates the Stargardt phenotype in adult Abca4–/– mice. Hum Gene Ther 2019; 30:156-600. [PMID: 30381971].
18. Li J, Liang Q, Song W, Marchisio MA. Nucleotides upstream of the Kozak sequence strongly influence gene expression in the yeast S. cerevisiae. J Biol Eng. BioMed Central 2017; 11:1-14. .
19. Dvir S, Velten L, Sharon E, Zeevi D, Carey LB, Weinberger A, Segal E. Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. Proc Natl Acad Sci USA 2013; 110:2792-801. .
20. Xue K, Jolly JK, Barnard AR, Rudenko A, Salvetti AP, Patrício MI, Edwards TL, Groppe M, Orlans HO, Tolmachova T, Black GC, Webster AR, Lotery AJ, Holder GE, Downes SM, Seabra MC, MacLaren RE. Beneficial effects on vision in patients undergoing retinal gene therapy for choroideremia. Nat Med. Nature Publishing Group 2018; 24:1507-12. [PMID: 30297895].
21. Cehajic-Kapetanovic J, Xue K. la Camara de CM-F, Nanda A, Davies A, Wood LJ, Salvetti AP, Fischer MD, Aylward JW, Barnard AR, Jolly JK, Luo E, Lujan BJ, Ong T, Girach A, Black GC, Gregori NZ, Davis JL, Rosa PR, Lotery AJ, Lam BL, Stanga PE, MacLaren RE. Initial results from a first-in-human gene therapy trial on X-linked retinitis pigmentosa caused by mutations in RPGR. Nat Med. Nature Publishing Group 2020; 26:354-9. .
22. Cremers FPM, van de Pol DJR, van Kerkhoff LPM, Wieringa B, Ropers HH. Cloning of a gene that is rearranged in patients with choroideraemia. Nature. Nature Publishing Group 1990; 347:674-7. [PMID: 2215697].
23. Fischer MD, McClements ME, Martinez-Fernandez de la Camara C, Bellingrath J-S, Dauletbekov D, Ramsden SC, Hickey DG, Barnard AR, MacLaren RE. Codon-Optimized RPGR Improves Stability and Efficacy of AAV8 Gene

Therapy in Two Mouse Models of X–Linked Retinitis Pigmentosa. Mol Ther 2017; 25:1854-65. [PMID: 28549772].

24. Xiong W, Wu DM, Xue Y, Wang SK, Chung MJ, Ji X, Rana P, Zhao SR, Mai S, Cepko C. AAV cis-regulatory sequences are correlated with ocular toxicity. Proc Natl Acad Sci USA 2019; 116:5785-94. .

25. Juttner J, Szabo A, Gross-Scherf B, Morikawa RK, Rompani SB, Hantz P, Szikra T, Esposti F, Cowan CS, Bharioke A, Patino-Alvarez CP, Keles O, Kusnyerik A, Azoulay T, Harti D, Krebs AR, Schübeler D, Hajdu RI, Lukats A, Nemeth J, Nagy ZZ, Wu KC, Wu RH, Xiang L, Fang XL, Jin ZB, Goldblum D, Hasler PW, Scholl HPN, Krol J, Roska B. Targeting neuronal and glial cell types with synthetic promoter AAVs in mice, non-human primates and humans. Nat Neurosci. Nature Publishing Group 2019; 22:1345-56. [PMID: 31285614].

26. Kozak M. Structural features in eukaryotic mRNAs that modulate the initiation of translation. J Biol Chem 1991; 266:19867-70. [PMID: 1939050].

27. Pickering BM, Willis AE. The implications of structured 5′ untranslated regions on translation and disease. Semin Cell Dev Biol 2005; 16:39-47. .

28. Reeder J, Höchsmann M, Rehmsmeier M, Voss B, Giegerich R. Beyond Mfold: Recent advances in RNA bioinformatics. Journal of Biotechnology. Elsevier 2006; 124:41-55. [PMID: 16530285].

29. Corley M, Solem A, Phillips G, Lackey L, Ziehr B, Vincent HA, Mustoe AM, Ramos SBV, Weeks KM, Moorman NJ, Laederach A. An RNA structure-mediated, posttranscriptional model of human α-1-antitrypsin expression. Proc Natl Acad Sci USA 2017; 114:10244-53. .

30. Noderer WL, Flockhart RJ, Bhaduri A, de Arce AJD, Zhang J, Khavari PA, Wang CL. Quantitative analysis of mammalian translation initiation sites by FACS-seq. Molecular Systems Biology. John Wiley & Sons Ltd 2014; 10:1-14. .