Nucleic Acids Research, 2026, **54**, gkag310
https://doi.org/10.1093/nar/gkag310
RNA and RNA-protein complexes

Kozak sequences regulate gene expression in *Trypanosoma brucei*

Philip Stettler*, Marina Cristodero, Norbert Polacek, André Schneider*

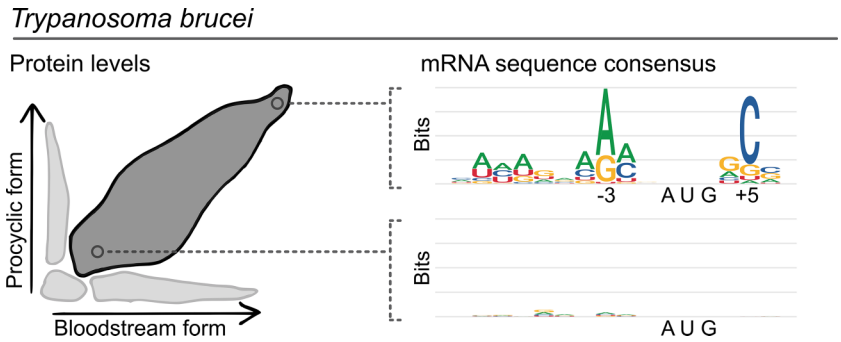
Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Bern 3012, Switzerland

*To whom correspondence should be addressed. Email: philip.stettler@sunrise.ch
Correspondence may also be addressed to André Schneider. Email: andre.schneider@unibe.ch

Abstract

The Kozak consensus sequence around the AUG start codon of mRNAs allows efficient cytosolic translation initiation in eukaryotes. It has not yet been investigated in the parasitic protozoan *Trypanosoma brucei*. mRNAs in *T. brucei* and other kinetoplastids are exclusively produced by polycistronic transcription. The levels of individual mRNAs can, therefore, not be regulated by transcription. Here, we show that in contrast to yeast and animals, no Kozak consensus sequence could be found in the total mRNA population of *T. brucei* or other kinetoplastids. However, when analyzing the subpopulation of constitutively expressed trypanosomal mRNAs encoding highly abundant proteins, a Kozak motif with a strong bias toward a +5C was detected. We tested how variants of the Kozak sequence influence the translation levels of a reporter protein. Using this *in vivo* approach, "weak" and "strong" Kozak sequences resulting in "low" and "high" translation levels could be defined. The "strong" sequences required a +5C and allowed initiation from CUG instead of AUG. We, therefore, suggest that in *T. brucei*, due to the lack of transcriptional control, Kozak sequences contribute to the regulation of protein levels. Moreover, we provide a new way to modulate protein abundance in transgenic trypanosomes in a predicted way.

Graphical abstract



Introduction

Cytosolic translation initiation in eukaryotes follows messenger RNA (mRNA) scanning by the 43S preinitiation complex consisting of the small ribosomal subunit and initiation factors. Starting at the 5′ cap of the mRNA, scanning stops at the AUG start codon, where additional factors bind to form the 48S initiation complex. Subsequently, the large ribosomal subunit is recruited and translation begins [1–6].

Already in the 1980s, it was shown that mRNAs of vertebrates and a few other organisms contain a short consensus sequence around the AUG start codon at the nucleotide positions −6 to +4 (start codon labelled as +1, +2, +3) [7–10]. Today, such sequence motifs are known as Kozak consensus sequences. They have since been identified in most eukaryotes and have been experimentally shown to increase translation efficiency in a few animals, fungi, and some plants [11–19].

Kozak consensus sequences of most species show an enrichment of purines (A or G) at the −3 position (−3A/G), and of an adenosine or a cytosine at the −2 position (−2A/C). Additionally, the position immediately following the AUG is in some groups enriched for guanosine (+4G) but may also display other lineage-specific bias or can be random [12]. Furthermore, an enrichment of a cytosine in position +5 (+5C) has been observed in some lineages and has been shown to correlate with enhanced translation efficiency in human [20–23].

In recent years, structures of the 43S and 48S initiation complexes confirmed that in addition to the eukaryotic translation initiation factors eIF1, eIF1A, and eIF2α, a few ribosomal proteins such as the rpS26e and rpuS19 (formerly rpS15)

Received: November 4, 2025. **Revised:** March 14, 2026. **Accepted:** March 19, 2026

directly interact with mRNAs at the −3, −2, +4, and +5 positions. This confirms the importance of the Kozak sequence for translation initiation [12, 24–29]. Additionally, analyses of the translation initiation kinetics showed that a sequence matching the Kozak consensus sequence accelerates the onset of translation by facilitating rapid transition from the 43S to the 48S initiation complex [30].

Kozak consensus sequences have mainly been studied in animals and fungi, which belong to the Opisthokonta in the supergroup Amorphea, and in land plants which belong to the supergroup Archeaplastida [11, 12]. Only few, mostly *in silico*, analyses have been done in protists representing other supergroups [12, 23].

Among protists, the Kinetoplastida of the supergroup Discoba are of particular interest [31]. Unlike all other known eukaryotes, Kinetoplastida species have organized most nuclear-encoded protein coding genes in polycistrons. During transcription by RNA polymerase II monocistronic mRNAs of the approximately 10 000 protein coding genes are generated by 5′ *trans*-splicing and 3′ polyadenylation [32, 33]. Consequently, gene expression in kinetoplastid species is not regulated on the transcriptional level but must rely mainly on post-transcriptional mechanisms.

In this study, we have focused on the best studied kinetoplastid, *Trypanosoma brucei*, a dixenous parasite with an originally sub-Saharan distribution [34]. *T. brucei* is transmitted by blood-feeding tsetse flies (*Glossina* spp.) and causes devastating disease, such as sleeping sickness in humans and nagana in some domestic mammals [35]. Previous research has shown that translation in *T. brucei* is regulated predominantly through mRNA features, including length and sequence of the 3′ untranslated region (3′UTR) and to a lesser extent also the 5′UTR, codon usage preferences, as well as possibly upstream open reading frames (uORFs) [36–41]. As a parasite, *T. brucei* also requires gene expression regulation during the lifecycle, where it progresses through several insect and mammalian lifecycle stages and undergoes massive transcriptome and proteome remodeling [39, 42–45].

In this study, we compiled genome-wide Kozak consensus sequences for ten publicly available genera of kinetoplastids. Intrigued by the absence of Kozak consensus sequences in most mRNAs in these organisms, we focused on *T. brucei* and found that only a subset of mRNAs, coding for highly abundant and lifecycle stage nonspecifically expressed proteins, exhibits a canonical Kozak consensus sequence.

Materials and methods

Kozak consensus sequences, sequence logos, and Kozak similarity scores

Genomic sequences were retrieved from **VEuPathDB** databases [46]: **FungiDB** (*S. cerevisiae* S288C), **HostDB** (*H. sapiens* REF), **TriTrypDB** (*A. deanei* strain Cavalho ATCC PRA-265, *B. ayalai* B08-376, *B. saltans* strain Lake Konstanz (quality: draft genome), *C. fasciculata* strain Cf-Cl, *E. monterogeii* strain LV88, *L. donovani* BPK282A1, *L. major* strain Friedlin, *L. pyrrhocoris* H10, *P. confusum* CUL13, *P. hertigi* MCOE/PA/1965/C119, *T. brucei forma brucei* TREU927, *T. cruzi* CL Brener Esmeraldo-like), and **VectorBase** (*D. melanogaster* iso-1). For Kozak consensus sequence logos, each one transcript per annotated gene was included if it fulfilled the criteria (i) of a single AUG start codon (ii) a known sequence from −10 nucleotides upstream to +10 nucleotides downstream of the start codon. The number of sequences per species fulfilling these criteria are indicated in sequence logos.

Sequence logos were created with R (version 4.2.1) using the packages "Biostrings" (version 2.64.1) [47] from Bioconductor (version 3.17), ggplot2 (version 3.4.0) [48], and ggseqlogo (version 0.2) [49]. Default settings were used where applicable.

Total information contents of sequence logos were calculated as the sum of information contents in every position of a sequence logo. The information contents were calculated from the Shannon entropy as described previously [50, 51].

Kozak similarity scores were calculated as defined by Gleason *et al.* [52]. Briefly, scores have a value range of 0.0–1.0. The maximal value 1.0 is reached if a sequence of interest matches the most frequent nucleotide in every position of a certain sequence consensus. Minimal values ≥0 are obtained when the given sequence matches the least frequent nucleotide in every position of a given sequence consensus. In our analysis, the positions −6 to +6 in sequences of interest were in all cases compared to the Kozak consensus sequence logo created from the genes of the top 100 ranked group 1 proteins.

Multiple sequence alignments

Sequences of the eIF1, eIF1A, eIF2α, rpS26e, and rpuS19 (formerly rpS15) were retrieved from reference genomes of the respective species deposited in VEuPathDB databases (see above).

Multiple sequence alignments were calculated in R using the msa package (version 1.30.1) [53] and visualized with ggmsa (version 1.4.0) [54]. Default settings were used where applicable.

Reanalysis of proteomics and ribosome profiling data

Quantitative proteomic data from procyclic form (PCF) and long slender bloodstream form (BSF) *T. brucei* were taken from Tinti et *al.* [55]. Proteins were divided into four groups after the following criteria: Group 1, iBAQ PCF > 4.3 & iBAQ BSF > 4.0; Group 2, iBAQ PCF > 4.3 & iBAQ BSF ≤ 4.0; Group 3, iBAQ PCF ≤ 4.3 & iBAQ BSF > 4.0; Group 4, iBAQ PCF ≤ 4.3 & iBAQ BSF ≤ 4.0. Within groups, proteins were assigned ranks where the lowest rank (rank 1) was assigned to the highest intensity-based absolute protein quantification (iBAQ) value (i.e. the most abundant protein). Group 2 and 3 proteins were ordered by ranks in the iBAQ PCF and BSF measurements, respectively. Group 1 and 4 proteins were ordered by the rank sum of the iBAQ PCF and iBAQ BSF ranks, and new ranks were assigned. The total information content or the nucleotide frequencies were calculated from sliding windows with a width of 100 ranks, i.e. windows were sequences of ranks 1:100, 2:101, 3:102, …, $(n-99)$:n (n: group size).

Translation efficiency data for group 1 genes were taken from ribosome profiling data published by Vasquez *et al.* [45]. We exclusively considered genes for which the translation efficiency in PCF and BSF was estimated and ordered genes by the rank sum of the PCF and BSF translation efficiencies. Nucleotide frequencies were calculated as described above.

Transgenic cell lines and reporter gene design

Cell lines are derivatives of a single marker PCF *Trypanosoma brucei forma brucei* 427 strain [56]. Cells were cultivated in SDM-79 supplemented with 5% v/v fetal calf serum at 27°C.

The reporter gene consisted of an eGFP gene fused 5′ to the coding sequence for three hemagglutinin (HA) tags. To systematically modify the Kozak sequence, nucleotides +10 to +15 of the eGFP gene were replaced by a HindIII restriction site, resulting in a G5L substitution for all constructs. The reporter gene was cloned into a modified pLEW100 vector carrying a Puromycin resistance cassette for selection [57], placing the expression of the reporter gene under the control of a tetracycline repressor and a procyclin promotor. Vectors were linearized by a NotI restriction enzyme digest for stable integration into ribosomal DNA (rDNA) loci. For each final construct, five individual cell lines were selected and initially screened for reporter gene translation efficiency. The cell line exhibiting the median eGFP fluorescence 12 h post induction with 1.0 μg/ml tetracycline was selected for further experiments.

Uniform expression of eGFP-3xHA in selected cell lines was investigated by immunofluorescence microscopy (IFA). Briefly, one million cells induced with tetracycline for 12 h were allowed to settle on glass slides before fixation for 10 min with 4% w/v paraformaldehyde in phosphate-buffered saline (PBS, 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, and 1.8 mM KH_2PO_4, pH 7.4). Fixed cells were permeabilized for 4 min with 0.2% v/v Triton X-100 in PBS. After permeabilization, cell were washed with chilled PBS, and blocked with PBS containing 2% w/v bovine serum albumin (BSA) before decoration with mouse anti-HA (Sigma, H9658; diluted 1:1000 in 2% BSA w/v in PBS) followed by goat anti-mouse Alexa Fluor 596 antibodies (Thermo Scientific, diluted 1:1000 in 2% BSA w/v in PBS). Pictures were acquired on a DMI6000B microscope equipped with a DFC360 FX monochrome camera operated with the LAS X software (Leica Microsystems). Images were processed using Fiji software.

Reporter protein expression measurements

In-cell eGFP fluorescence was measured to compare reporter gene expression over a 12-h period post induction with tetracycline (1.0 μg/ml). Exponentially growing cultures were diluted to a final concentration of ten million cells per ml. Induced (with tetracycline) and uninduced control cultures were prepared in 96-well plates (black wells, clear bottom). Wells containing only SDM-79 media served as baselines. Measurements were performed on a Tecan Infinite M1000 PRO plate reader maintained at 27°C. Measurement cycles were repeated for 12 h and consisted of the following steps: shaking for 10 min at 186 rpm (orbital movement, 3.5 mm amplitude), followed by optical density measurement at 600 nm and fluorescence measurements with an excitation at 488 nm and emission detection at 510 nm.

Immunoblotting

Samples for immune blotting were taken from exponentially growing cultures 6 h post induction with tetracycline. Cells were collected by centrifugation at 2700 × *g* for one minute and washed with PBS before lysis and denaturation in SDS–PAGE sample buffer. Two million cell equivalents were loaded onto 12% polyacrylamide gels. Gels were transferred to nitrocellulose membranes which were decorated with following antibodies: primary mouse anti-HA (Sigma, H9658; diluted 1:5000), primary rabbit anti-ATOM40 (Tb927.9.9660) ([58], dilution 1:10 000) secondary IRDye 680LT goat anti-mouse, and secondary IRDye 800CW goat anti-rabbit (both from LI-COR Biosciences, dilution 1:20 000).

RNA extraction and northern blotting

RNA was purified following an acid guanidinium thiocyanate–phenol–chloroform extraction protocol [59]. Five micrograms of total RNA extract per sample were separated on an agarose gel and transferred to a nylon transfer membrane. The open reading frame of the reporter gene was detected with a P^{32}-radioactively labeled probe synthesized from a gel purified PCR amplicon of the eGFP open reading frame using the Prime-a-Gene labeling system (Promega).

Results

Kinetoplastids lack genome-wide Kozak consensus sequences

Kozak consensus sequences describe a bias in the nucleotide frequency of eukaryotic mRNAs around the start codon. This bias can be quantified as the "information" content, which is based on the Shannon entropy, of a sequence pattern which can be visualized in nucleotide sequence logos [50, 51].

Kozak consensus sequences have been thoroughly investigated in most of the well-established opisthokont model organisms. These include human (*Homo sapiens*), fruit fly (*Drosophila melanogaster*), and budding yeast (*Saccharomyces cerevisiae*). All three species feature genome-wide Kozak consensus sequences: an increase in the information content around the start codon for the sum of mRNAs encoded by the genome (Fig. 1A). Most prominently, we find an enrichment of A or G at position −3, furthermore positions −5 to −1 upstream of the start codon, as well as in human position +4 immediately following the start codon, show the largest sequence bias. To compare the information content contained in the Kozak consensus sequences between species, we quantified the information content for entire sequences logos, from the −10 nucleotide upstream of the start codon to the +6 nucleotide within the ORF (the AUG itself was excluded from the analysis). The results showed that the genome-wide sequence logos of the Kozak consensus sequences range between 1.38 and 1.60 bits (Fig. 1B).

There is not much known about Kozak consensus sequences in kinetoplastid species. To study this group, we constructed genome-wide Kozak consensus sequence logos for twelve kinetoplastid species representing ten genera and two families. The quantification of the information content of these sequence logos revealed an unexpectedly low sequence consensus in ten of these species when compared to opisthokonts (Fig. 1B). Only the sequence logos of *Paratrypanosoma confusum* (1.31 bits) and *Crithidia fasciculata* (1.30 bits), have a total information content that is comparable to the species shown in Fig. 1A. In contrast, the sequence logos of some kinetoplastids, such as *T. brucei* or *Blechomonas ayalai*, have information contents of 0.33 and 0.32 bits which is almost five-fold less than what is seen in opisthokonts.

While *P. confusum* and *C. fasciculata* display weakly increased information contents in the nucleotide positions of the

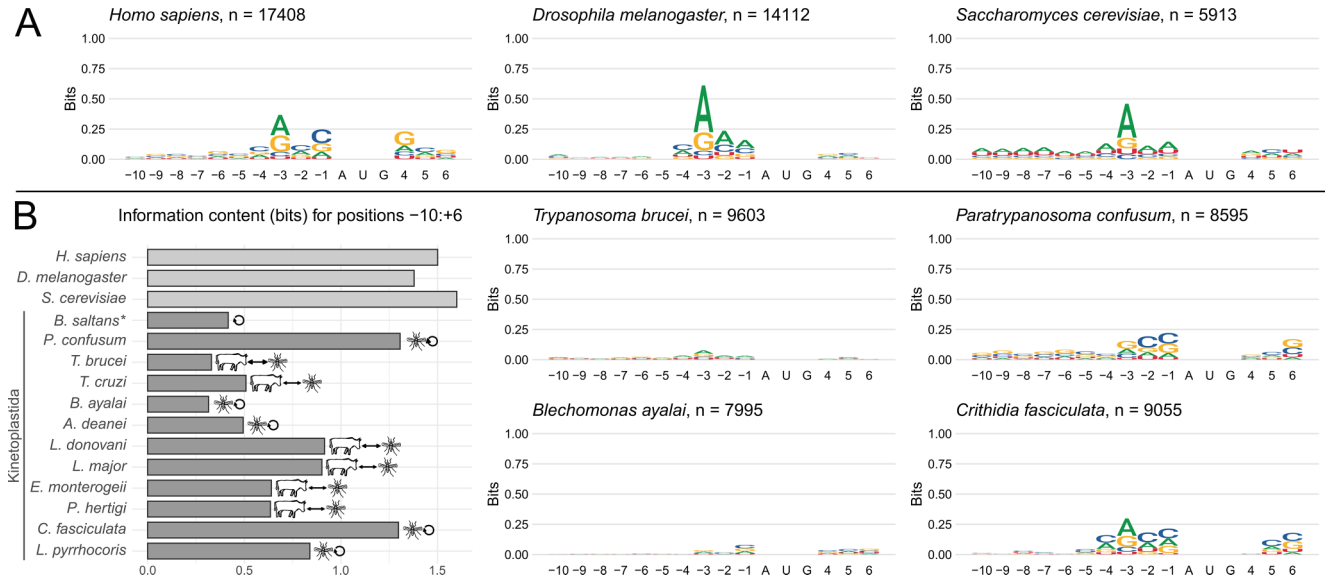


Figure 1. Kinetoplastids lack genome-wide Kozak consensus sequences. (**A**) Nucleotide sequence logos of the sequence context around the start codon in genes of *H. sapiens*, *D. melanogaster*, and *S. cerevisiae*. The range of 10 nucleotides upstream to 6 nucleotides downstream of the AUG start codon is shown. *Y*-axes are truncated at 1.0 bits (maximum is 2.0). *n* indicates the number of included sequences. (**B**) (Left) Total information contents (sum of bits values of a sequence logo from positions −10:+6) of sequences logos in (A) compared to sequence logos constructed identically for twelve species of Kinetoplastida. Symbols depict if species are free-living (closed circle), monoxenous parasites (arthropod symbol), or dixenous parasites (vertebrate and arthropod symbols). The included genome of *B. saltans* (marked by an asterisk) is a draft genome. (Right) Nucleotide sequence logos as in (A) of two kinetoplastid species with low (*T. brucei* and *B. ayalai*) and moderate information content (*P. confusum* and *C. fasciculata*).

canonical Kozak consensus sequence, such a trend is not seen in *T. brucei* or *B. ayalai* (Fig. 1B). Interestingly, the varying information content within Kinetoplastida does neither reflect their inner phylogeny nor does it correlate with a mono- or dixenous parasitic lifestyle of the species. In summary, these analyses suggest that most kinetoplastids lack genome-wide Kozak consensus sequences.

Kozak sequence-binding factors are conserved between the Opisthokonta and Kinetoplastida

A widespread lack of Kozak consensus sequences in kinetoplastid mRNAs could be the result of kinetoplastid-specific mutations in factors that recognize these mRNA features in other organisms. In opisthokonts, nucleotides of the Kozak sequence bind to eIF1, eIF1A, eIF2α and the two ribosomal proteins rpS26e and rpuS19 (previously rpS15). A previous study reported that these proteins including their key residues for mRNA contact are well conserved across a broad range of eukaryotes [12].

Multiple sequence alignments of these five proteins in twelve representative kinetoplastid and three opisthokont species revealed the striking conservation of nearly all residues implicated in Kozak sequence recognition (Fig. 2). In detail, the two arginines of eIF2α that interact with the −3 and −2 positions and the rpuS19 domain contacting the +4 nucleotide are identical (with the possible exception of *B. saltans* but for which only a draft genome is available). *A. deanei* appears to have the most diverged factors, this species has a tryptophane to tyrosine substitution at the eIF1A site contacting the +4 position, yet the lysine recognizing the +5 position is conserved. A few other lineage-specific substitutions were also observed. Overall, the residues directly contacting the Kozak sequence in species of the Opisthokonta are well conserved in all analyzed kinetoplastid species, indicating that their ribosomes likely retain the capability to engage with canonical Kozak sequences.

T. brucei mRNAs encoding highly abundant proteins have a Kozak consensus sequence

T. brucei is the best studied kinetoplastid species which is why we made it the focus of our study. The lack of genome-wide Kozak consensus sequences in *T. brucei* (Fig. 1B) does not exclude that a smaller set of mRNAs may show a Kozak or Kozak-like sequence consensus. The canonical Kozak consensus sequence of Opisthokonta promotes efficient translation initiation. Thus, our working hypothesis was that if such a consensus sequence exists in kinetoplastids, it would most likely be found in mRNAs coding for the most highly expressed proteins.

To test this hypothesis, we reanalyzed a previously published dataset reporting absolute protein abundances for both the PCF and BSF of *T. brucei* [55]. We grouped proteins according to their expression pattern: proteins expressed at comparable levels in both PCF and BSF (Group 1), PCF-specific proteins (Group 2), BSF-specific proteins (Group 3), and a forth group including proteins with low abundance in PCF or BSF and proteins with possibly error prone copy number estimation (Group 4) (Fig. 3A). Group-wide nucleotide sequence logos of the corresponding mRNAs revealed that neither of the groups feature a defined Kozak consensus sequence (Fig. 3A).

Next, we ranked proteins within the respective groups according to their abundance, rank 1 being the most abundant. For groups 1 and 4, protein abundances were ordered by the rank sums of the PCF and BSF measurements and new ranks were assigned. To evaluate the correlation between protein abundance and Kozak consensus sequence on the mRNA, nucleotide sequence logos for subgroups with mRNAs coding

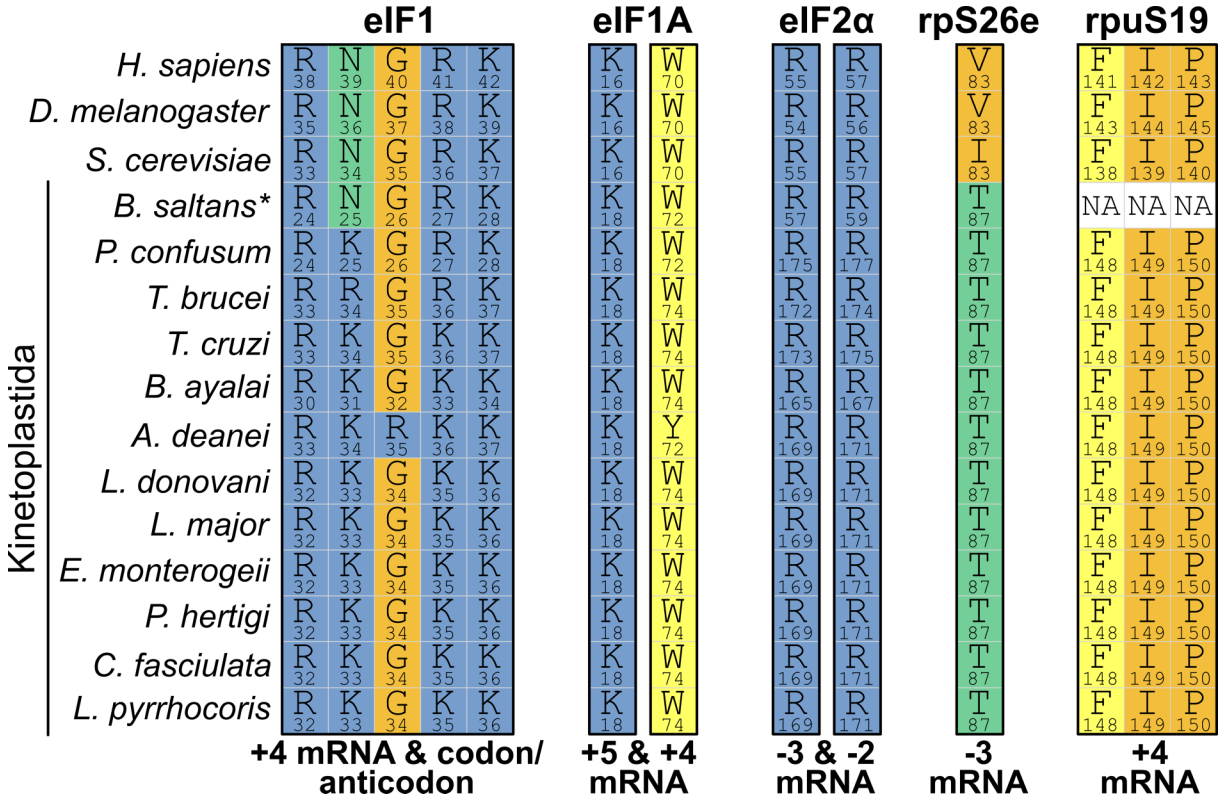


Figure 2. Kozak sequence-binding factors are conserved between the Opisthokonta and Kinetoplastida. Summary of multiple sequence alignments constructed for the eukaryotic translation initiation factor 1 (eIF1), factor 1A (eIF1A), factor 2α (eIF2α), and the cytosolic ribosomal proteins S26e (rpS26e) and uS19 (rpuS19, formerly rpS15) for the indicated species. Only key amino acids known to interact with the Kozak sequence are shown as indicated. The included genome of *B. saltans* (marked by an asterisk) is a draft genome. Amino acids in each ortholog are shown in single letter notation and respective amino acid positions are displayed. Colors indicate side chain characteristics [blue: positively charged, green: uncharged hydrophilic, orange: hydrophobic, yellow: aromatic, white: not applicable (NA)/deletion].

for 100 ordered proteins at a time were constructed. This was done using sliding windows for which the total information contents of the region of the Kozak consensus sequence were calculated (Fig. 3B).

This analysis still did not detect any Kozak consensus sequences for mRNAs encoding proteins of groups 2, 3, or 4. However, for group 1 proteins this was very different. The top ranked ∼700 proteins are encoded from mRNAs that feature strong Kozak consensus sequences with a total information content similar to or even higher than genome-wide Kozak consensus sequences observed in the three opisthokont species (Fig. 3B). Additionally, within these ∼700 proteins, the highest total information content was found in the Kozak consensus sequences of mRNAs encoding for the 100 top ranked proteins. The Kozak consensus sequence of this subgroup shows a strong bias toward a −3A and +5C as well as low/moderate biases in the −9, −8, −7, −4, −2, and +4 nucleotide positions (Fig. 3B).

As the total information content does not necessarily imply a specific sequence consensus, we expanded this analysis and calculated nucleotide frequencies of each of the positions −6 to +6 (excluding the start codon) separately. This analysis confirmed that within mRNAs coding for group 1 proteins there is a strong correlation of an A at position −3 and of a C at position +5 with abundance of the corresponding protein (Supplementary Fig. S1). To further investigate the putative causal relationship between −3A and +5C with translation initiation efficiency, we repeated the same analysis using translation efficiency data retrieved from a published ribosome profiling study [45]. To that end we ordered the mRNAs coding for group 1 proteins by the rank sum of their PCF and BSF translation efficiencies (low ranks corresponding to high translation efficiency) and analyzed the nucleotide frequencies of positions −6 to +6. Intriguingly, the results mirrored the findings obtained with protein abundance ranks (Supplementary Fig. S1) confirming the correlation between −3A and +5C with high translation efficiency and protein abundance, respectively (Supplementary Fig. S2).

Kozak consensus sequences regulate translation efficiency in *T. brucei*

There is a strong correlation between high protein abundance and trypanosomal mRNAs that closely match the Kozak consensus sequence gaa**A**ag<u>AUG</u>g**C**c (start codon underlined, highly conserved nucleotides in bold, less conserved nucleotides in lower case) (Fig. 3B right). However, the analysis suffers from the following limitations: mRNAs encoding different ORFs and steady state protein levels, that do not necessarily reflect translation initiation rates, are being compared.

We wanted to find out whether the observed correlation reflects a causal relationship. To that end we tested mRNAs variants containing different Kozak sequences (positions −6 to +6) in the context of the same ORF. The reporter protein used was the C-terminally triple HA-tagged enhanced green fluorescent protein (eGFP-3xHA). The resulting constructs were stably integrated into an rDNA locus and the eGFP-3xHA mRNA was expressed under tetracycline control.

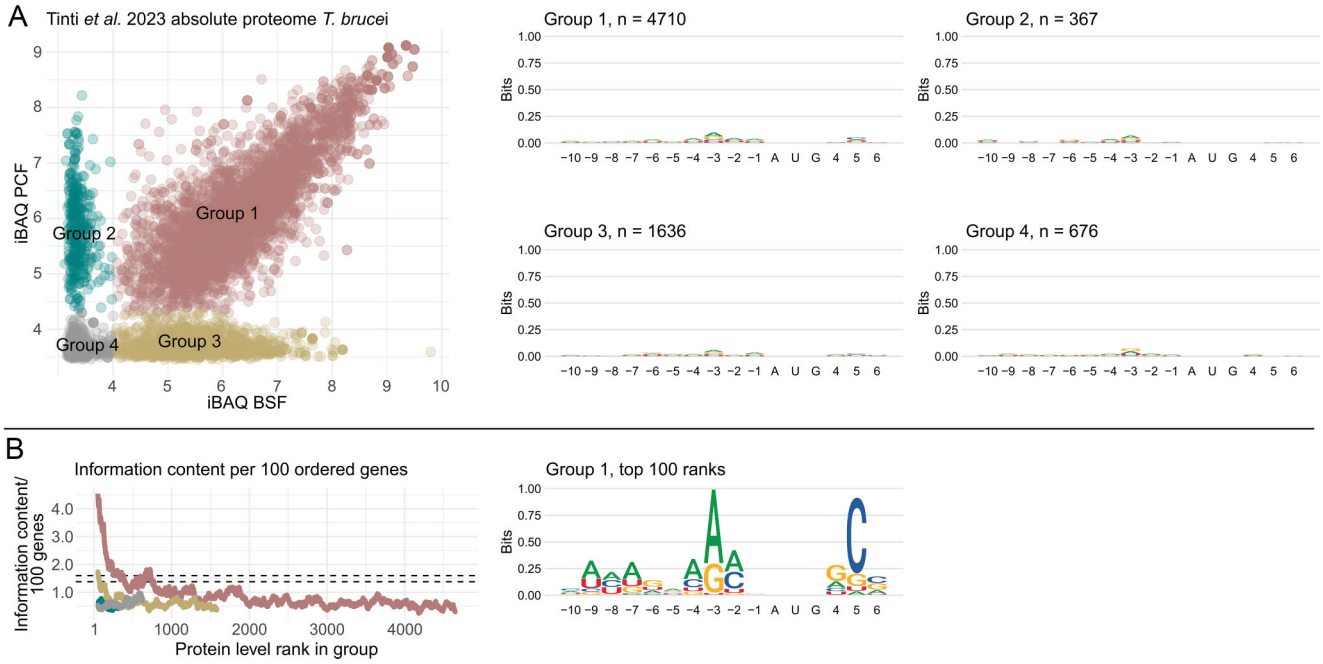


Figure 3. *T. brucei* mRNAs encoding highly abundant proteins have a Kozak consensus sequence. (**A**) (Left) Protein abundance shown as iBAQ in PCF and BSF *T. brucei* (data taken from Tinti *et al.*, 2023). Proteins were grouped into four groups as indicated by colors (see Materials and methods). (Middle and right) Group-wide nucleotide sequence logos of the sequence context around the start codon; *n* indicates the number of sequences per group. (**B**) (Left) Total information content in positions −10:+6 of a sliding window (width = 100 ranks) through genes ordered by their protein product abundance in the respective groups from (A), colors as in (A). Low ranks correspond to high protein abundance. Horizontal black dashed lines indicate the range of genome-wide information contents observed for *H. sapiens*, *D. melanogaster*, and *S. cerevisiae* (Fig. 1). (Right) Nucleotide sequence logo of the sequence context around the start codon for the mRNAs corresponding to the top 100 protein level ranks in group 1.

Newly synthesized eGFP-3xHA was quantified by measuring eGFP fluorescence during 12 h after tetracycline addition. To compare different cell lines the largely linear slope reflecting the increase of eGFP fluorescence between 4 and 8 h after expression induction was quantified (Fig. 4A and B). This slope should be a much better proxy for the translation efficiency than simply scoring the steady state levels of the protein.

The *T. brucei* genome contains more than a dozen rDNA repeat regions on different chromosomes [32]. In which of these rDNA loci and where within a given rDNA repeat a construct gets integrated can influence the expression levels of the mRNAs it encodes. To control for this, we initially analyzed five independent clonal cell lines for each construct (Supplementary Fig. S2A and B). Out of these cell lines the median performing clone was selected for quantification (Fig. 4A and B) and RNA was extracted to control for similar eGFP-3xHA mRNA expression by northern blot analyses (Supplementary Fig. S2).

First, we tested a Kozak sequence exactly matching the consensus gaa**A**agAUG**G**c and two derivatives thereof which are highly similar. How close a given sequence matches the Kozak consensus sequences is expressed by the Kozak similarity scores: 1.0 means a perfect match, 0 means no similarity (see Materials and methods). The three tested sequences had a Kozak similarity score of >0.97 (Fig. 4A) and were compared to a sequence designed to give the lowest possible Kozak similarity score and two derivatives thereof, all with Kozak similarity scores of <0.15 (Fig. 4B). The results showed that Kozak sequences with a Kozak similarity score of >0.97 were an average two-fold more efficiently translated than the three sequences with a Kozak similarity score of <0.15.

Next, we investigated whether the two most conserved position of the Kozak consensus sequence alone, −3A and +5C, can promote efficient translation. To that end we tested three new constructs whose Kozak sequences were based on the construct tested in Fig. 4A right (GAAA**A**AAUGG**C**C) with a Kozak similarity score of >0.99. In the first construct the −3A of the Kozak sequence was replaced by −3U (GAAUAAAUGG**C**C), in the second the +5C was replaced by +5A (GAA**A**AAAUGG**A**C), and in the third both positions were replaced (GAAUAAAUGG**A**C). The results showed that replacing the conserved +5C by an A significantly reduces translation efficiency whereas the same is not seen when −3A is replaced by a U (Fig. 4C and Supplementary Fig. S2C).

In summary, these results strongly suggest that the observed correlation between the high abundance of a protein and a high Kozak similarity score of the corresponding mRNA reflects a causal relationship. Moreover, the translation efficiency of a given trypanosomal protein can be amplified by a high Kozak similarity score in its mRNA, and having a C at position +5 within the Kozak sequence seems to be an important determinant for this.

A strong Kozak sequence allows translation initiation at a CUG codon

Non-AUG start codons can be used in a variety of organisms and their occurrence in cytosolic mRNAs of eukaryotes has been linked to strong Kozak sequences [12, 60–62]. While non-AUG start codons have so far not been found in cytosolic mRNAs of *T. brucei*, we wondered whether a strong Kozak sequence might allow translation initiation from the CUG which codes for leucine.

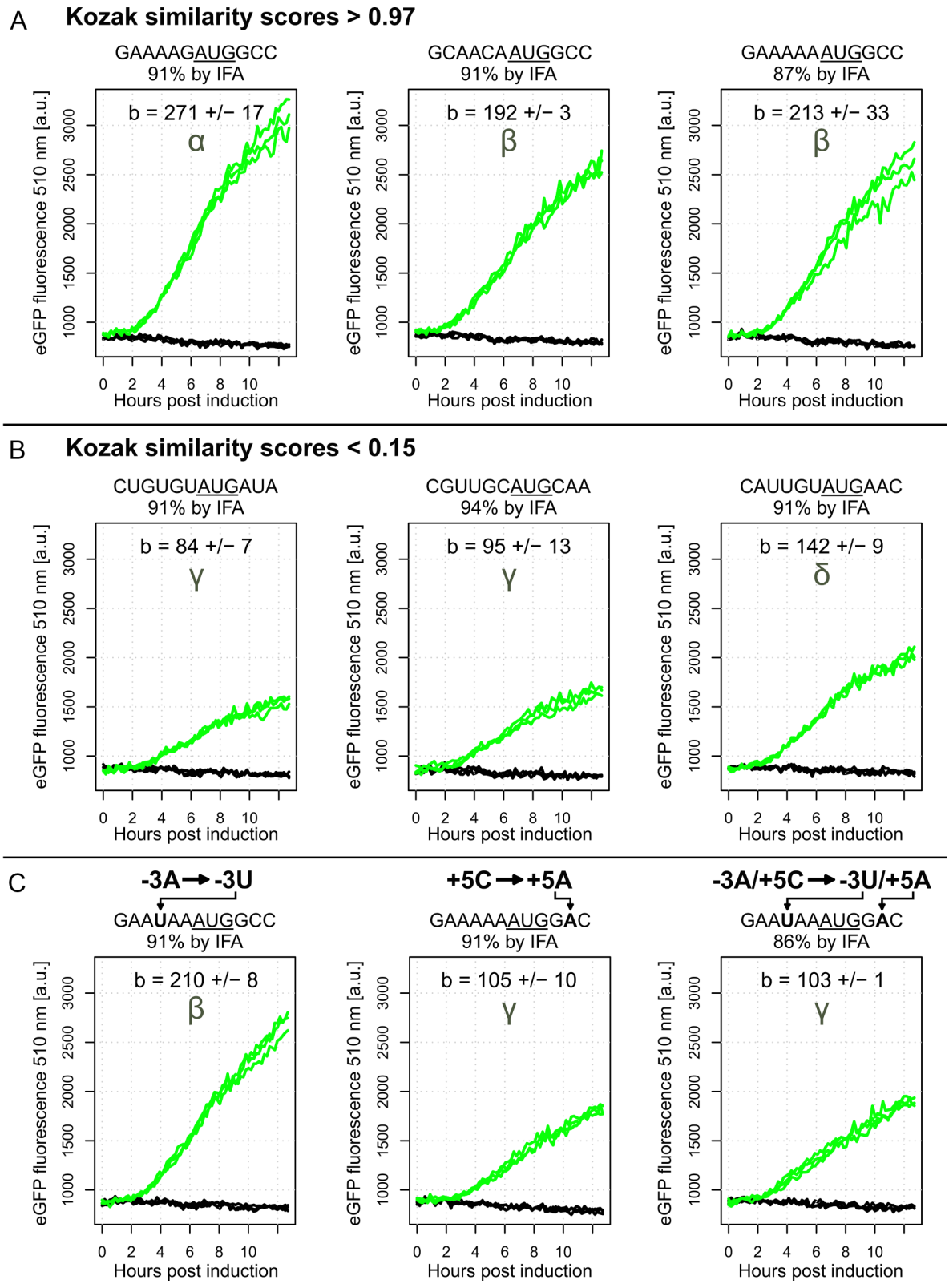


Figure 4. Kozak consensus sequences regulate translation efficiency in *T. brucei*. (**A**) (Left) eGFP fluorescence versus time post induction with tetracycline in PCF *T. brucei* expressing eGFP-3xHA from a genetic context with a high Kozak similarity score as indicated (start AUG underlined). Kozak similarity scores were calculated against the Kozak motif in Fig. 3B (right) (see Materials and methods). $n = 3$ technical replicates are shown (black: uninduced control cultures, green: induced cultures). Uniform eGFP-3xHA expression in each cell line was measured by IFA and is indicated. The fluorescence increase rate between 4 and 8 h post induction (*b* value) was quantified as a measure of translation efficiency and is indicated as a mean ± standard deviation (arbitrary units/hour). (Middle and right) as in (left) but with two varying Kozak consensus sequences as indicated. (**B**) As in (A) but eGFP-3xHA expressed from genetic contexts with low Kozak similarity scores. (**C**) As in (A and B) but eGFP-3xHA expressed from derivatives of the Kozak sequence in (A-right) with indicated point mutations. Different Greek letters in graphs indicate statistically significant differences between the performances of the Kozak sequences ($P < 0.05$, ANOVA with *post hoc* two-sided Student's *t*-tests, Benjamini–Hochberg adjustment).

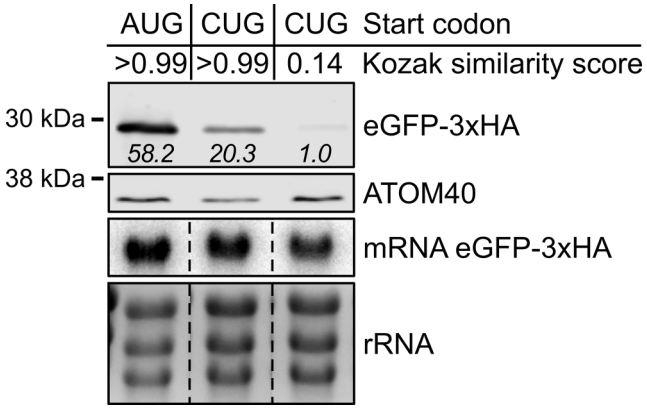


Figure 5. A strong Kozak sequence allows translation initiation at a CUG codon. Cell lines expressing eGFP-3xHA with an AUG or CUG start codon in a sequence context with a high or low Kozak similarity score were analyzed by immunoblotting (6 h post induction with tetracycline). eGFP-3xHA was probed by HA-tag specific antibodies, italicized numbers show the means of relative densitometric measurements ($n = 3$). The signal of the atypical translocase of the outer membrane 40 (ATOM40) serves as a loading control. mRNA expression was monitored by northern blotting using a probe targeting the ORF of eGFP. The signals of cytosolic rRNAs stained with ethidium bromide serve as loading controls.

To test this, we compared expression of two eGFP-3xHA encoding mRNAs, one with a Kozak similarity score >0.99 (GAAAAAC UGGCC) and another one with the Kozak similarity score of 0.14 (CAUUGUC UGAAC). eGFP-3xHA expressed from a construct with the canonical start codon AUG and a Kozak similarity score of >0.99 was used as a control. The immunoblots in Fig. 5 show that the eGFP-3xHA encoded on the mRNA with the canonical AUG start codon was efficiently expressed, as expected. Moreover, eGFP-3xHA encoded on the mRNA with the alternative CUG start codon was also expressed, albeit at ∼3x lower levels, and only if it contained a strong Kozak sequence, while the mRNA with the alternative CUG start codon in the context of the weak Kozak sequence produced nearly no detectable eGFP-3xHA.

In summary, these results (Fig. 5) corroborate our previous findings (Fig. 4) that Kozak sequences with a high similarity score amplify translation efficiency of a given trypanosomal ORF. Moreover, they also show that translation initiation from the CUG codon in *T. brucei* is possible, provided that it is expressed in the context of a strong Kozak sequence.

Discussion

Unlike other eukaryotes, trypanosomes and their relatives (Kinetoplastida) express all their protein-coding genes by polycistronic transcription from a few hundred promoters. As a consequence, mRNA levels from individual genes cannot be regulated by transcription. Trypanosomes therefore mainly rely on post-transcriptional mechanisms, including controlling translation and mRNA decay, to regulate gene expression both at steady-state conditions and for adaptation to different environments [33].

Here we propose a previously unrecognized way of how trypanosomes can regulate the abundance of proteins at steady state. They use the Kozak consensus sequence to increase the abundance of proteins that need to be expressed at high levels. This scenario is supported by the following lines of evidence:

(i) The Kozak consensus sequence could readily be detected in alignments of sequences around the AUG start codon in mRNAs of essentially any given eukaryote (Fig. 1A) [11, 12]. However, the same analysis failed to find such a sequence consensus in mRNAs of ten out of twelve representative species of Kinetoplastida including *T. brucei* (Fig. 1B).

(ii) However, a deeper analysis showed that a Kozak consensus sequence was not absent from all trypanosomal mRNAs, but could be found in mRNAs coding for proteins that are highly abundant in the PCF and BSF lifecycle stages. The presence of a Kozak sequence in a subpopulation of kinetoplastid mRNAs is consistent with the observed strong conservation of the key amino acid positions in proteins that interact with it (Fig. 2).

(iii) No Kozak consensus sequences could be detected in mRNAs encoding proteins that are stage specifically expressed in either PCF or BSF cells irrespectively of their expression levels (Fig. 3, Groups 2 and 3). This would be expected because the AUG start codon sequence context of any mRNA is identical throughout the lifecycle and because the conserved Kozak consensus sequence interacting proteins are expressed at comparable levels in PCF and BSF forms [55].

(iv) Using transgenic *T. brucei* cell lines expressing the tetracycline-inducible eGFP-3xHA reporter with different Kozak sequences showed that the low and high translation efficiencies of eGFP-3xHA were directly linked to this motif: high Kozak similarity scores led to high translation efficiency, low scores to low efficiency. This suggests that the observed proteome-wide correlation of proteins that are highly abundant in both lifecycle stages with high Kozak similarity scores is due to a causal relationship (Fig. 3).

(v) When a Kozak sequence with a very high Kozak similarity score is used eGFP-3xHA translation can be initiated at the non-conventional start codon CUG (Fig. 5). Similar results have also been obtained in other organisms [12, 60]. However, up to date no physiological examples of alternative start codon usage in cytosolic translation of trypanosomal mRNAs have been described [39, 45, 63]. These results, together with the examples above, illustrate that the trypanosomal Kozak sequence directly influences protein levels in a largely predictable way.

Based on the sequence logo in Fig. 3B the Kozak consensus sequence of *T. brucei* can be described as gaa**A**an<u>AUG</u>g**C**c. The most prominent features of this sequence are a conserved A at position −3 and a conserved C at position +5.

A −3 purine (A/G) is found in the Kozak consensus sequences of essentially all eukaryotes [11, 12].

It was therefore surprising that replacement of −3A by −3U did not affect translation rates of eGFP-3xHA in *T. brucei* (Fig. 4C and Supplementary Fig. S2C). A possible explanation could be that the eGFP-3xHA mRNA in transgenic *T. brucei* is transcribed from the procyclin promotor by RNA polymerase I which is more efficient than RNA polymerase II that transcribes the bulk of all mRNAs [64, 65]. Thus, a putative regulatory role of -3A on protein abundance might have been masked by the much higher transcription rate of RNA polymerase I.

A conserved +5C is not part of the originally reported canonical Kozak consensus sequence [9, 10] and an early report showed no effect of the +5 nucleotide on AUG recognition [66]. In contrast, another study found that the +5 and +6 positions are determinants in the context of non-AUG start codons [62]. However, both of these studies were largely

based on *in vitro* translation systems which might have affected translation regulation. Other analyses have shown that a low enrichment of +5C is found in many eukaryotes including humans and suggested that +5C promotes translation initiation [20–23]. Moreover eIF1A directly interacts with the +5 position [24, 28], finding a bias for a specific nucleotide at this site would therefore not be surprising. In line with the high correlation of +5C with protein abundance in trypanosomes (Supplementary Figs S1 and S2), we found that its replacement by an A resulted in an approximately two-fold reduced translation rate of eGFP-3xHA (Fig. 4C and Supplementary Fig. S2C). This suggests that in the *T. brucei* Kozak consensus sequence the +5C is an important determinant for highly efficient translation initiation. In fact the position +5 alone can in principle explain the results shown in Fig. 4. All three Kozak sequences with a Kozak similarity score of >0.97 had the +5C and showed high translation rates of eGFP-3xHA, whereas the three sequences with a Kozak similarity score of <0.15 had a U or an A at position +5 which resulted in an approximately two-fold lower translation rate. Note that the large effect of +5C on translation rate did not depend on the strength of the promoter. It was seen in the proteome-wide analysis where all mRNAs are transcribed by RNA polymerase II (Fig. 3B, and Supplementary Figs S1 and S2) but also with the reporter protein eGFP-3xHA which was transcribed by RNA polymerase I (Fig. 4). Intriguingly, even though a +5C is much less prevalent in the human consensus sequence than in *T. brucei* it appears also to correlate with a higher translation efficiency in humans [21].

There is a caveat: since the +5C, a primary determinant of high translation efficiency in *T. brucei*, lies within the ORF, its replacement will inevitably alter the identity of the second amino acid. Thus, the observed effects on protein abundance following +5C substitutions could, in principle, result from such an amino acid change. However, we consider this unlikely, as neither the fluorescence nor the stability of our reporter protein eGFP-3xHA have been shown to be affected by changes in amino acids near the N-terminus [67–69].

Using the eGFP-3xHA reporter gene allowed for well-controlled unbiased investigation of the effect of Kozak sequences on translation efficiency. We cannot exclude though that *in vivo*, where Kozak sequences represent just one of many regulatory layers, this effect might be less evident or even completely negligible for some genes.

T. brucei is an excellent experimental model system that offers a large toolbox of molecular genetic techniques. Various amounts of transgenic mRNAs can easily be expressed from either procyclin or ribosomal RNA (rRNA) RNA polymerase I promoters, or alternatively from RNA polymerase II promoters. However, the levels of the produced proteins often do not parallel the levels of the produced mRNAs. Our study now provides a novel tool to modulate the expression levels of a given protein. Expressing it in the context of a strong Kozak sequence, such as gaa**A**anAUG**C**c, will results in high levels of the protein, whereas expressing the same protein with a weak Kozak sequence, such as nnn(C/G/U)nnAUGn(A/G/U)n, will yield much lower amounts.

It remains unclear whether the Kozak consensus sequence was already present in the last eukaryotic common ancestor (LECA), implying a monophyletic origin, or whether it arose independently in different lineages through convergent evolution. The Kozak consensus sequence in *T. brucei* and its relatives is evident only in mRNAs encoding for highly abundant proteins, suggesting it has a more prominent regulatory role than in most other eukaryotes. Our findings indicated that this use of Kozak consensus sequences likely evolved because of the virtually exclusive polycistronic transcription of protein-coding genes in kinetoplastid species, which shifts gene expression control to post-transcriptional processes, including translation initiation.

Acknowledgements

We thank Elke Horn for excellent technical assistance.

Author contributions: Philip Stettler (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Project administration [equal], Resources [equal], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [equal], Writing—review & editing [equal]), Marina Cristodero (Data curation [supporting], Writing—review & editing [supporting]), Norbert Polacek (Resources [supporting], Writing—review & editing [supporting]), and André Schneider (Conceptualization [supporting], Funding acquisition [lead], Project administration [equal], Resources [equal], Supervision [lead], Writing—original draft [equal], Writing—review & editing [equal]).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

The authors declare no conflict of interest.

Funding

This work was supported by a grant of the NCCR RNA & Disease, a National Centre of Competence in Research (grant number 205601) to A.S. funded by the Swiss National Science Foundation (https://www.snf.ch/en). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Funding to pay the Open Access publication charges for this article was provided by Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, National Centre of Competence in Research, NCCR RNA & DiseaseGrant 205601.

Data availability

All data generated or analyzed during this study were either previously published or are included in the article and its Supplementary Information files.

References

1. Hashem Y, Frank J. The jigsaw puzzle of mRNA translation initiation in eukaryotes: a decade of structures unraveling the mechanics of the process. *Annu Rev Biophys* 2018;47:125–51. https://doi.org/10.1146/annurev-biophys-070816-034034
2. Merrick WC, Pavitt GD. Protein synthesis initiation in eukaryotic cells. *Cold Spring Harb Perspect Biol* 2018;10:a033092. https://doi.org/10.1101/cshperspect.a033092
3. Sokabe M, Fraser CS. Toward a kinetic understanding of eukaryotic translation. *Cold Spring Harb Perspect Biol* 2019;11:a032706. https://doi.org/10.1101/cshperspect.a032706

4. Pelletier J, Sonenberg N. The organizing principles of eukaryotic ribosome recruitment. *Annu Rev Biochem* 2019;88:307–35. https://doi.org/10.1146/annurev-biochem-013118-111042
5. Hinnebusch AG. Structural insights into the mechanism of scanning and start codon recognition in eukaryotic translation initiation. *Trends Biochem Sci* 2017;42:589–611. https://doi.org/10.1016/j.tibs.2017.03.004
6. Lind C, Aqvist J. Principles of start codon recognition in eukaryotic translation initiation. *Nucleic Acids Res* 2016;44:8425–32. https://doi.org/10.1093/nar/gkw534
7. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* 1986;44:283–92. https://doi.org/10.1016/0092-8674(86)90762-2
8. Kozak M. Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. *Nucl Acids Res* 1981;9:5233–52. https://doi.org/10.1093/nar/9.20.5233
9. Kozak M. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucl Acids Res* 1984;12:857–72. https://doi.org/10.1093/nar/12.2.857
10. Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl Acids Res* 1987;15:8125–48. https://doi.org/10.1093/nar/15.20.8125
11. Osnaya VG, Gomez-Romero L, Moreno-Hagelsieb G *et al.* AUGcontext DB: a comprehensive catalog of the mRNA AUG initiator codon context across eukaryotes. *RNA Biol* 2025;22:1–5. https://doi.org/10.1080/15476286.2025.2465196
12. Hernandez G, Osnaya VG, Perez-Martinez X. Conservation and variability of the AUG initiation codon context in eukaryotes. *Trends Biochem Sci* 2019;44:1009–21. https://doi.org/10.1016/j.tibs.2019.07.001
13. Zhong V, Archibald BN, Brophy JAN.. Transcriptional and post-transcriptional controls for tuning gene expression in plants. *Curr Opin Plant Biol* 2023;71:102315. https://doi.org/10.1016/j.pbi.2022.102315
14. Wallace EWJ, Maufrais C, Sales-Lee J *et al.* Quantitative global studies reveal differential translational control by start codon context across the fungal kingdom. *Nucleic Acids Res* 2020;48:2312–31. https://doi.org/10.1093/nar/gkaa060
15. Sugio T, Matsuura H, Matsui T *et al.* Effect of the sequence context of the AUG initiation codon on the rate of translation in dicotyledonous and monocotyledonous plant cells. *J Biosci Bioeng* 2010;109:170–3. https://doi.org/10.1016/j.jbiosc.2009.07.009
16. Lutcke HA, Chow KC, Mickel FS *et al.* Selection of AUG initiation codons differs in plants and animals. *EMBO J* 1987;6:43–8. https://doi.org/10.1002/j.1460-2075.1987.tb04716.x
17. Li J, Liang Q, Song W *et al.* Nucleotides upstream of the Kozak sequence strongly influence gene expression in the yeast S. cerevisiae. *J Biol Eng* 2017;11:25. https://doi.org/10.1186/s13036-017-0068-1
18. Hernandez G, Garcia A, Weingarten-Gabbay S *et al.* Functional analysis of the AUG initiator codon context reveals novel conserved sequences that disfavor mRNA translation in eukaryotes. *Nucleic Acids Res* 2024;52:1064–79. https://doi.org/10.1093/nar/gkad1152
19. Xie J, Zhuang Z, Gou S *et al.* Precise genome editing of the Kozak sequence enables bidirectional and quantitative modulation of protein translation to anticipated levels without affecting transcription. *Nucleic Acids Res* 2023;51:10075–93. https://doi.org/10.1093/nar/gkad687
20. Grzegorski SJ, Chiari EF, Robbins A *et al.* Natural variability of Kozak sequences correlates with function in a zebrafish model. *PLoS One* 2014;9:e108475. https://doi.org/10.1371/journal.pone.0108475
21. Ambrosini C, Destefanis E, Kheir E *et al.* Translational enhancement by base editing of the Kozak sequence rescues haploinsufficiency. *Nucleic Acids Res* 2022;50:10756–71. https://doi.org/10.1093/nar/gkac799
22. Niimura Y, Terabe M, Gojobori T *et al.* Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucleic Acids Res* 2003;31:5195–201. https://doi.org/10.1093/nar/gkg701
23. Nakagawa S, Niimura Y, Gojobori T *et al.* Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res* 2008;36:861–71. https://doi.org/10.1093/nar/gkm1102
24. Petrychenko V, Yi SH, Liedtke D *et al.* Structural basis for translational control by the human 48S initiation complex. *Nat Struct Mol Biol* 2025;32:62–72. https://doi.org/10.1038/s41594-024-01378-4
25. Llacer JL, Hussain T, Marler L *et al.* Conformational differences between open and closed states of the eukaryotic translation initiation complex. *Mol Cell* 2015;59:399–412. https://doi.org/10.1016/j.molcel.2015.06.033
26. Lomakin IB, Steitz TA. The initiation of mammalian protein synthesis and mRNA scanning mechanism. *Nature* 2013;500:307–11. https://doi.org/10.1038/nature12355
27. Rabl J, Leibundgut M, Ataide SF *et al.* Crystal structure of the eukaryotic 40S ribosomal subunit in complex with initiation factor 1. *Science* 2011;331:730–6. https://doi.org/10.1126/science.1198308
28. Hussain T, Llacer JL, Fernandez IS *et al.* Structural changes enable start codon recognition by the eukaryotic translation initiation complex. *Cell* 2014;159:597–607. https://doi.org/10.1016/j.cell.2014.10.001
29. Basu I, Gorai B, Chandran T *et al.* Selection of start codon during mRNA scanning in eukaryotic translation initiation. *Commun Biol* 2022;5:587. https://doi.org/10.1038/s42003-022-03534-2
30. Grosely R, Alvarado C, Ivanov IP *et al.* eIF1 and eIF5 dynamically control translation start site fidelity. *Nat Struct Mol Biol* 2025; 32:2308–18. https://doi.org/10.1038/s41594-025-01629-y
31. Burki F, Roger AJ, Brown MW *et al.* The New Tree of Eukaryotes. *Trends Ecol Evol* 2020;35:43–55. https://doi.org/10.1016/j.tree.2019.08.008
32. Berriman M, Ghedin E, Hertz-Fowler C *et al.* The genome of the African trypanosome *Trypanosoma brucei*. *Science* 2005;309:416–22. https://doi.org/10.1126/science.1112642
33. Clayton C. Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol* 2019;9:190072. https://doi.org/10.1098/rsob.190072
34. Kostygov AY, Karnkowska A, Votypka J *et al.* Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biol* 2021;11:200407. https://doi.org/10.1098/rsob.200407
35. Hollingshead CM, Bermudez R. Treasure Island (FL): StatPearls, 2025.
36. Trenaman A, Tinti M, Wall RJ *et al.* Post-transcriptional reprogramming by thousands of mRNA untranslated regions in trypanosomes. *Nat Commun* 2024;15:8113. https://doi.org/10.1038/s41467-024-52432-0
37. Tinti M, Horn D. Decoding post-transcriptional gene expression controls in trypanosomatids using machine learning. *Wellcome Open Res* 2025;10:173. https://doi.org/10.12688/wellcomeopenres.23817.2
38. Jeacock L, Faria J, Horn D. Codon usage bias controls mRNA and protein abundance in trypanosomatids. *eLife* 2018;7:e32496. https://doi.org/10.7554/eLife.32496
39. Jensen BC, Ramasamy G, Vasconcelos EJ *et al.* Extensive stage-regulation of translation revealed by ribosome profiling of *Trypanosoma brucei*. *Bmc Genomics [Electronic Resource]* 2014;15:911. https://doi.org/10.1186/1471-2164-15-911
40. de Freitas Nascimento J, Kelly S, Sunter J *et al.* Codon choice directs constitutive mRNA levels in trypanosomes. *eLife* 2018;7:e32467. https://doi.org/10.7554/eLife.32467
41. Fervers P, Fervers F, Makalowski W *et al.* Life cycle adapted upstream open reading frames (uORFs) in *Trypanosoma congolense*: a post-transcriptional approach to accurate gene

regulation. *PLoS One* 2018;13:e0201461. https://doi.org/10.1371/journal.pone.0201461
42. Siegel TN, Gunasekera K, Cross GA *et al.* Gene expression in Trypanosoma brucei: lessons from high-throughput RNA sequencing. *Trends Parasitol* 2011;27:434–41. https://doi.org/10.1016/j.pt.2011.05.006
43. Urbaniak MD, Guther ML, Ferguson MA. Comparative SILAC proteomic analysis of *Trypanosoma brucei* bloodstream and procyclic lifecycle stages. *PLoS One* 2012;7:e36619. https://doi.org/10.1371/journal.pone.0036619
44. Gunasekera K, Wuthrich D, Braga-Lagache S *et al.* Proteome remodelling during development from blood to insect-form *Trypanosoma brucei* quantified by SILAC and mass spectrometry. *Bmc Genomics [Electronic Resource]* 2012;13:556. https://doi.org/10.1186/1471-2164-13-556
45. Vasquez JJ, Hon CC, Vanselow JT *et al.* Comparative ribosome profiling reveals extensive translational complexity in different *Trypanosoma brucei* life cycle stages. *Nucleic Acids Res* 2014;42:3623–37. https://doi.org/10.1093/nar/gkt1386
46. Alvarez-Jarreta J, Amos B, Aurrecoechea C *et al.* VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center in 2023. *Nucleic Acids Res* 2024;52:D808–16. https://doi.org/10.1093/nar/gkad1003
47. Pagès H, P.A. RG, DebRoy S. Biostrings: efficient manipulation of biological strings. 2022. https://bioconductor.org/packages/Biostrings (1 January 2026, date last accessed).
48. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag, 2016. https://doi.org/10.1007/978-0-387-98141-3
49. Wagih O. ggseqlogo: a versatile R package for drawing sequence logos. *Bioinformatics* 2017;33:3645–7. https://doi.org/10.1093/bioinformatics/btx469
50. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 1990;18:6097–100. https://doi.org/10.1093/nar/18.20.6097
51. Schneider TD, Stormo GD, Gold L *et al.* Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986;188:415–31. https://doi.org/10.1016/0022-2836(86)90165-8
52. Gleason AC, Ghadge G, Chen J *et al.* Machine learning predicts translation initiation sites in neurologic diseases with nucleotide repeat expansions. *PLoS One* 2022;17:e0256411. https://doi.org/10.1371/journal.pone.0256411
53. Bodenhofer U, Bonatesta E, Horejs-Kainrath C *et al.* msa: an R package for multiple sequence alignment. *Bioinformatics* 2015;31:3997–9. https://doi.org/10.1093/bioinformatics/btv494
54. Zhou L, Feng T, Xu S *et al.* ggmsa: a visual exploration tool for multiple sequence alignment and associated data. *Brief Bioinform* 2022;23:bbac222. https://doi.org/10.1093/bib/bbac222
55. Tinti M, Ferguson MA.J. Visualisation of proteome-wide ordered protein abundances in *Trypanosoma brucei*. *Wellcome Open Res* 2022;7:34. https://doi.org/10.12688/wellcomeopenres.17607.2
56. Aeschlimann S, Kalichava A, Schimanski B *et al.* Single p197 molecules of the mitochondrial genome segregation system of *Trypanosoma brucei* determine the distance between basal body and outer membrane. *Proc Natl Acad Sci USA* 2022;119:e2204294119. https://doi.org/10.1073/pnas.2204294119
57. Wirtz E, Leal S, Ochatt C *et al.* A tightly regulated inducible expression system for conditional gene knock-outs and dominant-negative genetics in *Trypanosoma brucei*. *Mol Biochem Parasitol* 1999;99:89–101. https://doi.org/10.1016/S0166-6851(99)00002-X
58. Niemann M, Wiese S, Mani J *et al.* Mitochondrial outer membrane proteome of *Trypanosoma brucei* reveals novel factors required to maintain mitochondrial morphology. *Mol Cell Proteomics* 2013;12:515–28. https://doi.org/10.1074/mcp.M112.023093
59. Chomczynski P, Sacchi N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* 1987;162:156–9. https://doi.org/10.1006/abio.1987.9999
60. Diaz de Arce AJ, Noderer WL, Wang CL. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res* 2018;46:985–94. https://doi.org/10.1093/nar/gkx1114
61. Cao X, Slavoff SA. Non-AUG start codons: expanding and regulating the small and alternative ORFeome. *Exp Cell Res* 2020;391:111973. https://doi.org/10.1016/j.yexcr.2020.111973
62. Boeck R, Kolakofsky D. Positions +5 and +6 can be major determinants of the efficiency of non-AUG initiation codons for protein synthesis. *EMBO J* 1994;13:3608–17. https://doi.org/10.1002/j.1460-2075.1994.tb06668.x
63. Parsons M, Ramasamy G, Vasconcelos EJ *et al.* Advancing *Trypanosoma brucei* genome annotation through ribosome profiling and spliced leader mapping. *Mol Biochem Parasitol* 2015;202:1–10. https://doi.org/10.1016/j.molbiopara.2015.09.002
64. Gunzl A, Bruderer T, Laufer G *et al.* RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Euk Cell* 2003;2:542–51. https://doi.org/10.1128/EC.2.3.542-551.2003
65. Budzak J, Rudenko G. Pedal to the metal: nuclear splicing bodies turbo-charge VSG mRNA production in African Trypanosomes. *Front Cell Dev Biol* 2022;10:876701. https://doi.org/10.3389/fcell.2022.876701
66. Kozak M. Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J* 1997;16:2482–92. https://doi.org/10.1093/emboj/16.9.2482
67. Sarkisyan KS, Bolotin DA, Meer MV *et al.* Local fitness landscape of the green fluorescent protein. *Nature* 2016;533:397–401. https://doi.org/10.1038/nature17995
68. Weinstein JY, Marti-Gomez C, Lipsh-Sokolik R *et al.* Designed active-site library reveals thousands of functional GFP variants. *Nat Commun* 2023;14:2890. https://doi.org/10.1038/s41467-023-38099-z
69. Li X, Zhang G, Ngo N *et al.* Deletions of the Aequorea victoria green fluorescent protein define the minimal domain required for fluorescence. *J Biol Chem* 1997;272:28545–9. https://doi.org/10.1074/jbc.272.45.28545

Received: November 4, 2025. **Revised:** March 14, 2026. **Accepted:** March 19, 2026