Modular 5-UTR hexamers for context-independent tuning of protein expression in eukaryotes

Petersen, Søren D.; Zhang, Jie; Lee, Jae S.; Jakoinas, Tadas; Grav, Lise M.; Kildegaard, Helene F.; Keasling, Jay D.; Jensen, Michael K.

Published in:
Nucleic acids research

Link to article, DOI:
10.1093/nar/gky734

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Published online 14 August 2018

Nucleic Acids Research, 2018, Vol. 46, No. 21 ***e127***
doi: 10.1093/nar/gky734

Modular 5′-UTR hexamers for context-independent tuning of protein expression in eukaryotes

Søren D. Petersen[1], Jie Zhang[1], Jae S. Lee[1], Tadas Jakočiūnas[1], Lise M. Grav[1], Helene F. Kildegaard[1], Jay D. Keasling[1,2,3,4,5,6] and Michael K. Jensen[1,*]

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, [2]Joint BioEnergy Institute, Emeryville, CA 94608, USA, [3]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [4]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, CA 94720, USA, [5]Department of Bioengineering, University of California, Berkeley, CA 94720, USA and [6]Center for Synthetic Biochemistry, Institute for Synthetic Biology, Shenzhen Institutes of Advanced Technologies, Shenzhen 518055, China

Received March 09, 2018; Revised July 24, 2018; Editorial Decision July 30, 2018; Accepted August 01, 2018

ABSTRACT

Functional characterization of regulatory DNA elements in broad genetic contexts is a prerequisite for forward engineering of biological systems. Translation initiation site (TIS) sequences are attractive to use for regulating gene activity and metabolic pathway fluxes because the genetic changes are minimal. However, limited knowledge is available on tuning gene outputs by varying TISs in different genetic and environmental contexts. Here, we created TIS hexamer libraries in baker's yeast *Saccharomyces cerevisiae* directly 5′ end of a reporter gene in various promoter contexts and measured gene activity distributions for each library. Next, selected TIS sequences, resulted in almost 10-fold changes in reporter outputs, were experimentally characterized in various environmental and genetic contexts in both yeast and mammalian cells. From our analyses, we observed strong linear correlations (R^2 = 0.75–0.98) between all pairwise combinations of TIS order and gene activity. Finally, our analysis enabled the identification of a TIS with almost 50% stronger output than a commonly used TIS for protein expression in mammalian cells, and selected TISs were also used to tune gene activities in yeast at a metabolic branch point in order to prototype fitness and carotenoid production landscapes. Taken together, the characterized TISs support reliable context-independent forward engineering of translation initiation in eukaryotes.

INTRODUCTION

Control of protein expression is critical for cellular development, differentiation and adequate response to intra- and extracellular conditions (1). From simple bacteria to multicellular eukaryotes, control of protein expression involves the sequence composition of the 5′-untranslated regions (5′-UTRs) of existing messenger RNAs (mRNAs). Specifically, translation initiation, where the AUG start codon is identified by ribosomes and decoded by methionyl-(transfer RNAs) tRNAs (met-tRNAs), is recognized as one of the most crucial steps in translation (1–4). As a consequence, a large number of studies have been performed to deduce the relationship between 5′-UTR sequences and protein expression (5–11).

In bacteria, simple base-pairing between the 6-nt Shine–Dalgarno (SD) sequence located immediately 5′ end of the start codon and the anti-SD sequence in the peptidyl decoding site of the 16S ribosomal subunit controls translation initiation from 50 to 70% transcripts by modulating the ribosomal accessibility to the SD sequence (12,13). Moreover, deep sequence–function characterization of SD libraries has enabled the development of predictive algorithms for tuning of protein expression over several orders of magnitude by simple modulation of the SD sequence (4–6).

In eukaryotes, translation is initiated at the 5′ end of mRNA by the recruitment of the 40S ribosomal subunit, auxiliary initiation polypeptides and the met-tRNA, collectively the 43S pre-initiation complex (PIC) (14). Different from translation initiation in bacteria, once recruited, PIC scans along a much larger sequence space of eukaryotic 5′-UTRs, often several hundred nucleotides in length, until encountering an AUG codon (14–16). During scanning, a number of 5′-UTR sequence features are known to affect translation, including mRNA secondary structures,

*To whom correspondence should be addressed. Tel: +45 6128 4850; Fax: +45 4525 8001; Email: mije@biosustain.dtu.dk
Present address: Jae S. Lee, Department of Molecular Science and Technology, Ajou University, Suwon 16499, Republic of Korea.

decoy AUG codons, PIC stalling at upstream open reading frames (uORFs) and the sequence context surrounding the cognate AUG for translation, commonly referred to as the Kozak sequence (17–29).

Similar to the algorithms established in bacteria for tuning protein expression, major efforts have been performed to mine the causal sequence elements of native eukaryotic 5′-UTRs, in order to attempt to model and forward engineer 5′-UTRs with predictive protein expression outputs (7,9). Initially, Kozak *et al*. reported GCCRCCAUGGG (R = A/G, start codon underlined) to effectively control ribosomal recognition of AUG and thereby initiation of translation (30–32). In particular the positioning of a purine at position -3 and a guanine at position +4 from the AUG codon has later been adopted for efficient translation initiation (33–35). Expanding on this, in mammalian cell lines, Noderer *et al*. have systematically probed the efficiency of start codon recognition for all possible translation initiation sites (TISs) flanking the AUG start codon at positions -6 to -1, and +4 and +5, totaling ∼65 000 TIS sequences, concluding that the motif RYMRMVAUGGC (Y = U or C, M = A or C, R = A or G and V = A, C, or G, start codon underscored) enhanced start codon recognition and GFP translation efficiency (8). Likewise, in yeast, recent studies have attempted to accurately estimate TIS efficiencies on reporter protein expression by randomizing 5′-UTR elements up to 50 nt upstream AUG (uAUG) and training computational models on smaller subfractions (4×10^{-26} – 0.2%) of these libraries (7,9). Here, both studies pointed out that in addition to uORFs and mRNA secondary structure, positions -3 to -1 from the AUG start codon are the most important parameters for tuning protein expression, ultimately enabling the construction of an algorithm explaining up to 70% of the observed variation in protein levels (7). More recently, the computational model by Dvir *et al*. has been further validated with new experimental data, again investigating the interactions with polymorphism in nucleotides at positions -10 to -1 relative to the start codon, but this time also placing the TIS in genomic contexts of two different reporter proteins and two different promoters (10). Here, similar correlations between experimental data and model predictions were observed ($R^2 = 0.36$–0.73), ultimately suggesting that the ∼30% variation observed for which the current models cannot account for arises from experimental noise and yet-uncharacterized biological factors (7). Likewise, though Noderer *et al*. showed strong linear relationships between GFP expression in different mammalian cell lines and cultivation media, comparisons of TIS efficiencies of GFP expression compared to other reporter genes suggested some context-dependence of TISs and open reading frame (ORF; $R^2 = 0.39$–0.76), in line with the model being trained on ORF-specific library sequences including the +4 and +5 positions (8). Taken together, the above studies indicate that systematic characterization of the impact of short TISs on protein expression in broad contexts still remains to be elucidated before TISs can be used as a tool for predictable tuning of protein expression.

In this study, we sought to establish a robust, simple and experimentally validated workflow to assess the sequence–function relationship of TISs in diverse genomic and environmental contexts (Figure 1). To do so, we created three TIS libraries spanning more than 4500 designs for nucleotide positions -6 to -1 directly upstream of an ORF of GFP controlled by three different promoters in yeast. Based on fluorescence-activated cell sorting (FACS) and single clone validations, a diverse sample of TIS hexamers with a robust output range of ∼10-fold was selected for further characterization in the context of different ORFs, promoters, host chassis, growth medium and cell densities (Figure 1). In general, the linear relationship between relative fluorescence output from selected TIS sequences obtained in different contexts was high, with correlation coefficients ranging from 0.77 to 0.98. Moreover, testing TISs derived from yeast in mammalian cell lines, we specifically uncovered a TIS sequence stronger than the Kozak element commonly used to drive protein production in mammalian cell factories. In addition, we used selected TIS hexamers to investigate the carotenoid production landscape in yeast by tuning dual protein activities at an essential metabolic branch point, thereby prototyping the fitness and carotenoid production landscape in a simple and cost-effective manner. Our detailed, experimental analyses allow us to put forward a list of short sequence-validated TISs to be used as a method for predictable tuning of protein expression in diverse genomic and environmental contexts.

MATERIALS AND METHODS

Strains, cell lines and growth media

Baker's yeast *Saccharomyces cerevisiae* strains were derived from CEN.PK2-1C (EUROSCARF, Germany). Yeast strains were cultured in yeast synthetic drop-out media (Sigma-Aldrich) at 30°C. CHO-S cells (ThermoFisher) and derivative cells were maintained in CD CHO medium (Gibco Cat. #10743-029) supplemented with 8 mM L-glutamine (Lonza Cat. #BE17-605F) and 2 ml/L anti-clumping agent (Gibco Cat.#0010057AE) in 125 ml Erlenmeyer shake flasks (Corning Inc., Acton, MA), incubated at 37°C, 5% CO_2 at 120 rpm and passaged every 2–3 days. *Escherichia coli* DH5α were cultured in Luria-Bertani (LB) medium containing 100 mg/l ampicillin (Sigma-Aldrich) at 37°C.

Plasmid and strain construction

Yeast integrative plasmids were created by USER cloning (36) and propagated in *E. coli* DH5a. Yeast transformations were performed by LiAc/SS carrier DNA/PEG method (37). Plasmids and polymerase chain reaction (PCR) products were purified using kits from Macherey-Nagel. Biobricks for USER assembly were amplified using Phusion U Hot Start PCR Master Mix (ThermoFisher), parts for transformation by Phusion High-Fidelity PCR Master Mix with HF Buffer (ThermoFisher), whereas colony PCRs were performed using 2xOneTaq Quick-Load Master Mix with Standard Buffer (New England Biolabs). Oligos, duplex oligos and gBlocks were purchased from Integrated DNA Technologies (IDT). Sequencing was performed by Eurofins. All primers, plasmids, yeast strains and CHO cell lines are listed in Supplementary Tables S1, S2, S3 and S4, respectively.

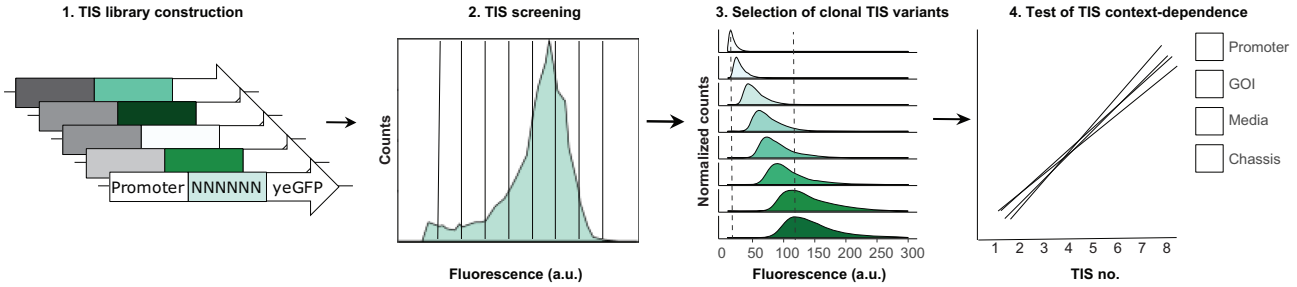

Figure 1. Workflow for TIS library construction and characterization.

Constructions of TIS libraries

Three TIS libraries were constructed using the EasyClone method (38) with slight modifications. Briefly, 0.1 pmol of promoter REV1 (389 bp upstream position +1 of YOR346W; Supplementary Table S5), RPL18B (700 bp; upstream position +1 of YNL301C) and TEF1 (420 bp; upstream position +1 of YPR080W) were cloned together with yeast-enhanced GFP (yeGFP; Supplementary Table S6) (39) with six randomized nucleotides upstream the start codon, into 0.03 pmol linearized EasyClone vector pCfB261 amplified from a vector excluding a *ccdB* cassette (suicide gene (40); pCFB8168) to counterselect for non-linear vector template. Correctly assembled EasyClone plasmids (0.5 pmol) containing the yeGFP cassette under either REV1, RPL18B or TEF1 promoter was linearized by NotI (Fermentas) and transformed into TC-3 cells for recombination at the EasyClone site XII-5 facilitated by Cas9 (41). *Escherichia coli* and *Saccharomyces cerevisiae* library colonies were scraped and pooled from five individual transformation plates. In *E. coli*, TIS library sizes in the context of REV1, RPL18B and TEF1 promoters were estimated by colony count to ∼1180, 4600 and 4520, respectively. In *S. cerevisiae*, REV1, RPL18B and TEF1 library sizes were estimated to ∼845, 3300 and 2750, respectively.

Construction of promoter and reporter strains

Thirty-two yeast strains were constructed similarly to the TEF1 promoter TIS library, by combinatorial assembly of promoter, TIS and reporter sequences using the CasEMBLR method with assemblies integrated into EasyClone site XII-5 (38,42). For these strains, the TIS was defined as one of eight sequences TGATAT, CGACTT, ACGTTC, GGGGGT, TAGGTT, AGGACA, TGTGAA or TCGGTC. Eight of the 32 strains were constructed by transformation of the alcohol dehydrogenase II (ADH2) promoter fragment (700 bp upstream position +1 of YMR303C), and one of eight TIS-yeGFP into strain TC-3 (43) for recombination at the XII-5 site. Homology between fragments was 30 bp, with up and down fragments of at least 450 bp for homologous recombination into EasyClone site XII-5. The remaining 24 strains were constructed similarly by transformations of the previously used TEF1 promoter fragment, with each of 24 fragments made by combination of the eight TIS sequences with the three reporter genes encoding ymUkG1 (44), yeGFP and mKate2 (45).

Construction of carotenoid strains

A background carotenoid expression strain was created by first, amplifying genes *crtI* and *crtYB* from plasmid YIplac211-YB/I/E* (46). Second, genes were USER cloned together with bidirectional promoter (pTDH3_pTEF1) into linearized vector pCfB390 (38) to create a plasmid pTAJAK-11. Third, linearized pTAJAK-11 was integrated into CEN.PK2-1C strain, XI-3 site as described in Jensen *et al.* (38), and the strain was named TC-9. Fourth, TC-9 was transformed with the Cas9 expression plasmid pCfB176 to create strain TC-10. Further, carotenoid strains were constructed by transforming (i) linear 90-bp DNA donor fragments spanning *Erg9* promoter and coding sequence introducing either TIS no. 1 (TGATAT), no. 5 (TAGGTT) or no. 8 (TCGGTC) directly upstream the start codon, (ii) a pCfB261 upstream part with phosphoglycerate kinase gene (PGK1) promoter (YCR012W; 984 bp) and (iii) either TIS no. 1 (TGATAT), no. 5 (TAGGTT) or no. 8 (TCGGTC) directly upstream the *crtE* start codon with a pCfB261 downstream homology part into EasyClone site XII-5 of strain TC-10 by the CasEMBLR method (Erg9 gRNA sequence: CACATATCACACACACACAA; XII-5 gRNA sequence: TTGTCACAGTGTCACATCAG) (42).

Construction of the CHO reporter cell pools

CHO reporter cell pools were derived from a master cell line harboring a recombinase-mediated cassette exchange (RMCE) landing pad. The master cell line was made by CRISPR-mediated homology directed targeted integration of CHO-S cells as previously described (47), with minor changes in the homology-directed repair (HDR) donor plasmid. The mCherry coding sequence in the HDR-donor plasmid (pCFB8173) has been flanked by a *loxP* sequence at the 5′ end and a *lox2272* sequence at the 3′ end (pEF1α-*loxP*-mCherry-*lox2272*-BGHpA), and the 5′ and 3′ homology arms target a non-coding region. Promoterless and polyAless RMCE vectors were constructed by assembly of PCR fragments containing TIS sequences no. 1 (TGATAT), no. 5 (TAGGTT), no. 8 (TCGGTC) or the mammalian consensus TIS (GCCACC) in combination with mammalian-enhanced GFP (meGFP) (48) or ZsGreen1 (Clontech #632428) that were flanked by *loxP* and *lox2272* sequences. The CHO master cell line at a concentration of 1×10^6 cells/ml was transfected with TIS-GFP or TIS-ZsGreen1 RMCE reporter plasmids and Cre-

recombinase vector in 3:1 ratio (w:w) in six-well plates using FreeStyle™ MAX transfection reagent to exchange mCherry coding sequence with TIS-GFP or TIS-ZsGreen1 cassettes. For Cre recombinase expression, PSF-CMV-CRE recombinase vector (OGS591, Sigma-Aldrich) was used. Transfected cell pools were passaged two times after transfection. After 7 days, cell pools were analyzed by flow cytometry. Flow cytometry revealed that 1–3% of the cells in all cell pools were changed from mCherry to GFP positive.

Next-generation sequencing of TIS libraries

Genomic DNA was extracted from over night cultures using PureLink Genomic DNA Purification Kit (Invitrogen). Genomic DNA extracts were used as template in PCR amplifying ∼300 bp overlapping the TIS sequence within the first 50 bp. Purified PCR products were indexed with Nextera XT indexing. The indexed amplicons were quantified using Qubit 2.0 Fluorometer (Life Technologies), pooled in equimolar quantities and sequenced on Illumina MiSeq using 75-bp reads. TIS sequences were extracted from sequencing reads using the cutadapt command line tool (49) treating the TIS sequence flanking regions as anchored adapters. Reads shorter than 70 bp was removed and up to 5 bp mismatches in the flanking regions were allowed in total. From 4.1 million usable reads from the three sequenced libraries, we identified 4037, 4093 and 4037 (or 98.6–99.9%) of the 4096 possible hexameric TIS sequences for each library. Of these 1721, 2174 and 844 TISs exceeded our cutoff of 100 reads for reliable quantification.

Flow cytometry and TIS library sorting by FACS

Yeast cells were grown in 96-well microtiter plates ON to saturation, diluted to OD_{600} 0.025 (measured by reading the absorbance at 600 nm on Microplate Reader, BioTek) and incubated for 4–6 h (until OD_{600} reached 0.1–0.2) before being measured by flow cytometry using a MACSquant VYB (Miltenyi) or BD Fortessa (BD Biosciences) flow cytometer. CHO reporter cell pools in exponential phase were analyzed by flow cytometry using a BD FACSJazz cell sorter (BD Biosciences). Fluorescence of yeGFP, ymUkG1, meGFP and ZsGreen1 was measured after excitation by 488 nm laser and detected through 525/50 nm bandpass filters. Fluorescence of mKate2 and mCherry was measured after excitation by 561 nm laser and detected through a 615/20 nm bandpass filter. Cells were gated based on FSC-A and FSC-H (singlets) as well as FSC-A and SSC-A profiles for robust measurements. All fluorescence data presented are median values for at least 10 000 or 5000 cells from yeast and CHO cells, respectively. Flow cytometry data were analyzed using FlowLogic version 700.2A (Inivai Technologies). Fluorescence measurements from each fluorescent protein were mean normalized (each measurement divided by the mean and multiplied by 100).

The TIS library in the context of the TEF1 promoter was divided into 10 equal gates based on yeGFP signal, and 48 cells were sorted out from each gate using BD FACS ARIA II (BD Biosciences). In total 480 cells were spotted onto agar plates, grown in liquid cultures and validated by flow cytometry (BD Fortessa, BD Biosciences). Eight colonies spanning the range of fluorescence were selected and sequenced to reveal the corresponding TISs.

Characterization of carotenoid strains

Pre-cultures were inoculated from glycerol stock and incubated for 48 h before 2 μl culture was spotted onto SD agar. Pictures of colonies on agar plates were taken after incubation for 48 h at 30°C and 72 h at 5°C. Maximum specific growth rates (μ_{max}) were calculated using the Easylinear function from the growth rates R package (50) and setting the number of consecutive data points to 10 ($h = 10$). Growth were measured in 200 μl cultivations (Growth Profiler 960, EnzyScreen).

β-Carotene extraction and quantification by HPLC

Measurements were performed using a method described by (51) with a few modifications. β-Carotene was extracted from 2 ml culture broth. The pelleted cells were lysed with 250 μl glass beads and in 500 μl ethyl acetate supplemented with 0.01% 3,5-di-tert-4-butylhydroxy toluene (BHT). Finally, 300 μl ethyl acetate was evaporated from cell extracts and the pellet was redissolved in 1.5 ml ethanol with 0.01% BHT for high pressure liquid chromatography (HPLC) measurements.

Data analysis

Data analysis were mainly done using the *R* statistical environment (version 3.4.1). Additional analysis using RNAfold from the Vienna RNA package (version 2.4.6), the yUTR-calculator by Decoene *et al.* (10) and cutadapt version 1.13 was performed in a Python 3.5 environment. Systematic names of yeast genes using one of TISs 1 to 8 were found using the find Motif search tool in CLC Main Workbench (version 7.7.2) on the CEN.PK113-7d genome (Genbank ID: AEHG01000000).

RESULTS

Construction and characterization of TIS libraries

Our first aim was to characterize in high-throughput how short TIS sequences affect protein expression in eukaryotes. In yeast and mammalian cells, earlier studies have characterized TIS libraries ranging from -50 to -1, -10 to -1, -6 to +4 and -6 to +5 positions relative to the AUG start codon (7–9,33,52). From those studies, it has been inferred that (i) positions -3 to -1 from the AUG start codon, (ii) a purine (R) in position -3 or lack of G in position +4, (iii) mRNA secondary structure and (iv) out-of-frame uAUGs are the most important parameters for tuning protein expression. In order to experimentally investigate in greater detail the potential for identifying a list of hexameric TISs spanning the -6 to -1 position that can be used to predictably tune protein expression in a context-independent manner (i.e. promoter or gene of interest (GOI) proximal sequences), we initially constructed three TIS libraries, each containing a yeGFP expression cassette controlled by either a weak (REV1), medium (RPL18B) or strong (TEF1) constitutive

promoter, which span approximately three orders of magnitude in expression level (53) and have <54% similarity in the -50 to -7 position (Supplementary Figure S1). Between promoter and gene, we cloned randomized hexameric TISs and genomically integrated the reporter expression cassettes (Figure 2A). For each of the three libraries, we also constructed a corresponding control strain by substituting the randomized nucleotides with -6 to -1 positioned nucleotides, AAAACA, from the strong PGK1 promoter (54) (Figure 2A). From the 4096 possible hexamer variants, DNA sequencing revealed library coverages of 42% (1721), 53% (2174) and 21% (844) from the REV1, RPL18B and TEF1 TIS libraries, respectively (Supplementary Figure S2). Both the per base position and overall frequency of nucleotides A, C, G and T were ∼20%, 10%, 45% and 25%, respectively (Supplementary Figure S3). Flow cytometry analysis of the three TIS libraries revealed variances in yeGFP fluorescence of ∼2, >3 and >4 orders of magnitude for the REV1, RPL18B and TEF1 promoters, respectively (Figure 2B–D). Also, none of the three libraries included TIS variants that exceeded the fluorescence measurements observed in cells expressing the PGK1 TIS AAAACA (Figure 2B–D, red).

In order to identify, and further characterize, TISs showing a high degree of protein expression tunability, we sorted the TIS library in the context of the TEF1 promoter, which showed the highest detectable variance of the three libraries. Briefly, we selected 480 single cells based on gating (Supplementary Figure S4), thereby covering a large fraction of the yeGFP expression range (Figures 1 and 3A). Following single clonal validation of fluorescence, we selected eight strains uniformly covering the maximum 10-fold range in yeGFP expression observed in our TIS library and determined the TISs by sequencing (Figure 3A). Importantly, as the chromophore formation is an O_2-dependent autocatalytic process (55), it is critical to consider cell density of the small cultivation volumes (150 μl) used when comparing fluorescence intensities of cell populations expressing individual TIS variants. Accordingly, as we observed a rapid decrease in maximum per cell fluorescence with increasing cell densities, we analyzed yeGFP expression at OD_{600} = 0.1–0.2 as also reported from studies in bacteria (Supplementary Figure S5) (56).

Modular TISs show context-independent tuning of protein expression in yeast

In eukaryotes, earlier studies of TIS libraries, spanning larger sequence spaces (e.g positions -50 to -1 or -6 to +5), have focused on building computational models to enable forward engineering of protein expression levels (7–10). Though these efforts have enabled high-throughput enumeration of sequence parameters of importance for predictive tuning of protein expression, the predicted protein expression levels from placing TISs in new genomic contexts (i.e. promoter or gene of interest) have so far not been able to explain >30% of the variation observed between experimentally deduced behavior compared to the genomic context in which the TIS algorithms were originally designed (7,10).

To investigate if shorter TIS variants selected from our FACS analysis (Figure 3A) would have context-dependent effects in protein expression, we placed the TIS sequences in the context of a different promoter, different genes of interest (GOI) and in another cultivation medium, and used standard flow cytometry to analyze protein expression from a total of 32 genomic designs (Figure 3B and Supplementary Figure S9). Specifically, in addition to the strong constitutive TEF1 promoter, we also tested the eight selected TISs in the context of the glucose-repressed ADH2 promoter, and for GOI we included two other fluorescent reporters: ymUKG1 and mKate2 (44). The selection of GOIs and promoter contexts was based upon maximal sequence diversity and carbon source dependent expression, respectively (57) (Supplementary Figure S1). Next, from the 32 designs we experimentally validated, fluorescence measurements revealed from 2- to 10-fold variation, with glucose and TEF1 contexts displaying the largest fold changes between the weakest (TIS 1, TGATAT) and the strongest (TIS 8, TCGGTC) TISs (Figure 3C–E). The five strongest TISs all had a purine in position -3, which is in line with earlier reports (7,8,10,52). Furthermore, the TIS sequence dictates the fluorescence in a similar manner across all tested genomic and environmental conditions, as evidenced by the linear correlations between mean fluorescence values between individual promoter or reporter contexts (Pearson's, R^2 = 0.75–0.98) (Figure 4). This range overlaps with previous studies based on larger number of TISs (8).

Benchmarking measured TIS efficiencies

When comparing the measured relative expression values of all 32 combinatorial designs with existing computational algorithms for predicting translation initiation (8,10), we observe substantial correlation coefficients (R^2 = 0.44–0.86) between measured and predicted values (Supplementary Figures S6 and S7), with the model inferred by Noderer *et al.* generally displaying stronger correlations compared to the model generated by Decoene *et al.* However, as reported in these previous modeling studies, translation efficiency for some TISs (in our case TIS no. 4) would require additional experimental validation (incl. context) to further refine predictive tuning of protein expression (8).

Complementary to these observations, we performed a genome-wide search for the occurence of the eight different TISs and compared the result with the translation efficiency of the native gene products as reported by Lathvee *et al*. (58). Briefly, this analysis identified TISs 2, 3, 5, 6 and 7 in the -6 to -1 position in a total of 11 genes of which four genes with TISs 5, 6 and 7 had translation efficiencies reported (58). From this small number of hits, the translation efficiencies of the two genes with TIS 6 were higher than the efficiency reported for the gene with TIS 5, whereas the gene with TIS 7 had the lowest translation efficiency reported among the four genes (Supplementary Table S6).

Moreover, one critical parameter known to influence translation initiation efficiency is the folding propensity of 5′-UTRs (17,20,30). To further benchmark translation initiation efficiencies of TIS 1–8, we calculated the minimum free energy (kcal/mol) for bases at positions -15 to +50 by RNAfold as a function of normalized mean fluorescence

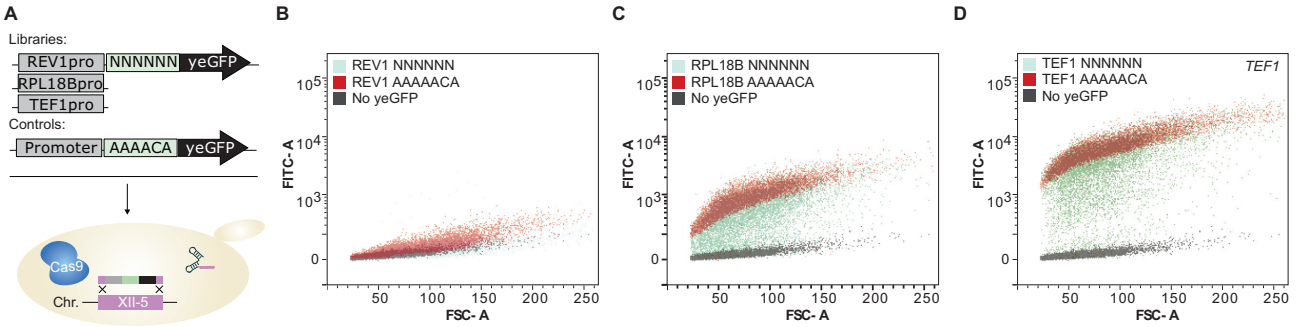

Figure 2. Distribution of reporter gene activities of three TIS libraries. (**A**) Schematic outline of the TIS library designs in the context of three different promoters (REV1, REPL18B and TEF1) controlling the expression of yeGFP. Negative control strain was without yeGFP expression. Fluorescence outputs for all three libraries were compared to yeGFP expression under the control of the TIS AAAACA from the strong PGK1 promoter. Libraries and control designs were integrated into yeast chromosome XII, EasyClone site 5 by CRISPR-mediated double-strand breaking and homologous recombination. (**B**) Fluorescence (FITC-A) as a function of forward scatter (FSC-A) of the TIS library in the context of REV1 promoter. (**C**) Fluorescence (FITC-A) as a function of forward scatter (FSC-A) of the TIS library in the context of RPL18b promoter. (**D**) Fluorescence (FITC-A) as a function of forward scatter (FSC-A) of the TIS library in the context of TEF1 promoter. Scatter plots in (B–D) are displayed together with control populations having TIS AAAACA (red) and wild-type cells without yeGFP expressed (gray).

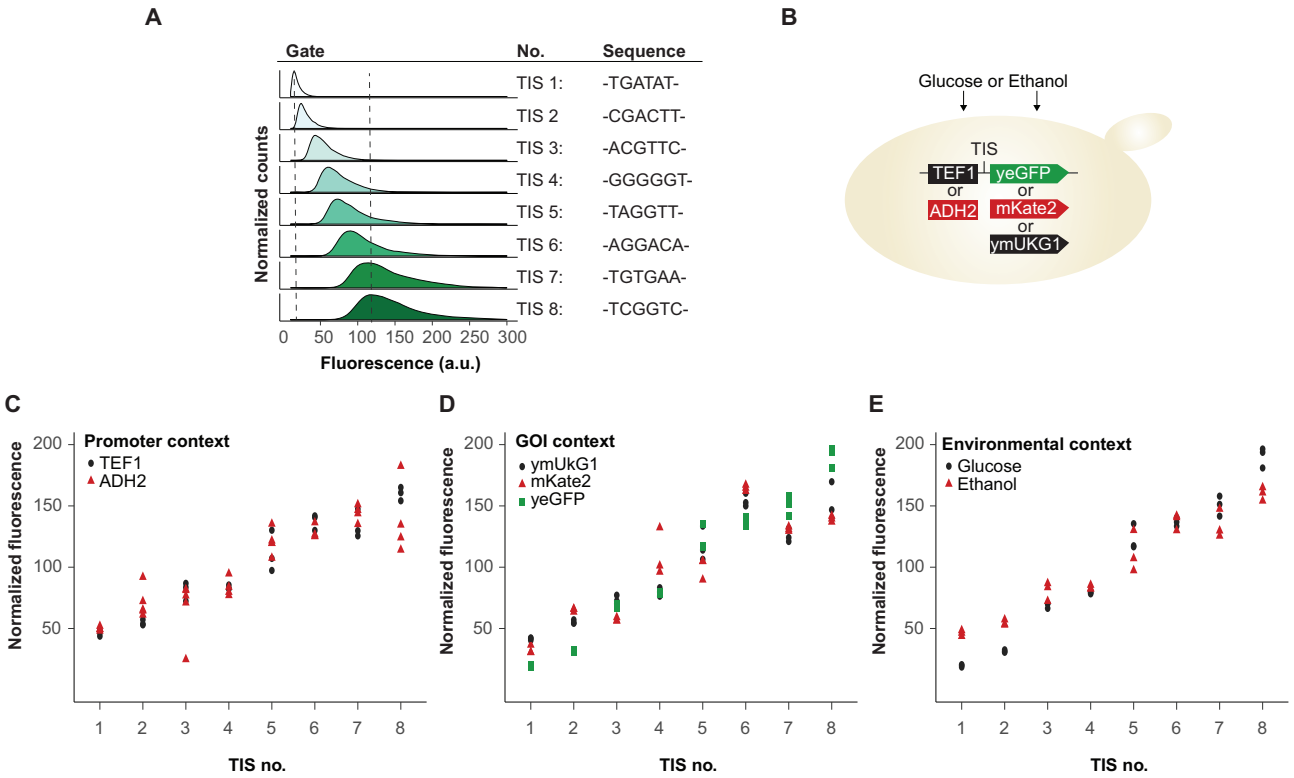

Figure 3. Investigation of interaction effects between TIS sequences and promoter, gene of interest (GOI) or growth culture condition. (**A**) Fluorescence histograms of gated populations of single clonal variants from the TIS library in the context of the TEF1 promoter. To the right, selected TISs from each of the populations are represented. (**B**) Schematic outline of the experimental design used to investigate interaction effects. (**C**) Normalized median fluorescence measured for the eight different selected TIS sequences in the context of a constitutive promoter (TEF1) and a glyconeogenic promoter (ADH2). (**D**) Normalized median fluorescence measured for the eight different selected TIS sequences in the context of three different GOI; ymUkG1, mKate2 and yeGFP. (**E**) Normalized median fluorescence measured for the eight different selected TIS sequences in the context of two different carbon sources present in the growth medium; glucose or ethanol. In plots (D–E), the TIS sequence order is selected from the TIS library controlling yeGFP expression under the control of the TEF1 promoter. In plot (C–E), median fluorescence values are shown for at least three biological replicates each based on at least 5000 single cell measurements.

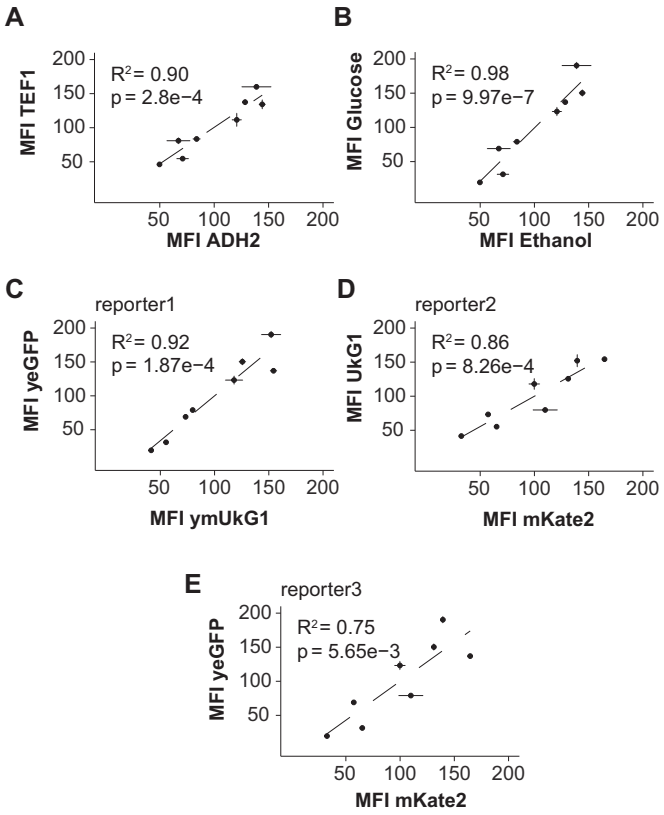

Figure 4. Linear correlations between normalized fluorescence measurements for eight TIS sequences in diverse genomic and environmental contexts. (**A**) Correlation between yeGFP fluorescence measurements for the two tested promoters (TEF1 and ADH2). (**B**) Correlation between yeGFP fluorescence measurements when using the two different carbon sources, glucose or ethanol. (**C**–**E**) Correlation between fluorescence measurements between each pair of the three different fluorescent reporters tested, ymUkG1, mKate2 and yeGFP. The fluorescence values are means of normalized median fluorescence values from at least three biological replicates each based on at least 5000 single cell measurements.

values for all 32 strain designs in this study. From this analysis, we observed no significant positive correlation ($R^2 = 3.0 \times 10^{-3}$, $P = 0.77$) (Supplementary Figure S8), which could indicate that the effect of varying the relative small hexameric TISs reported in this study, only have a modest effect on the minimum free energy observed for the sequence space analyzed (-15 to +50) (21).

Taken together, benchmarking the eight TISs with existing translation initiation prediction tools (8,10,21) and experimentally measured translation efficiencies (58) reveals that the model output overall correlated with our measured TIS efficiencies. However, when investigating the correlation between genome-wide occurrences of the identified TISs and their translation efficiency, the numbers are too low to infer statistical significance. Finally, we observe no significant correlation between the TIS strength of the 32 different designs studied and the folding propensity of their -15 to +50 regions (21), indicating that the hexameric TISs only modestly affect folding propensity of the 5′-UTRs.

TISs show context-independent tuning of protein expression in mammalian cells

In mammalian cells, the TIS sequence RYMRMVAUGGC (Y = U or C, M = A or C, R = A or G and V = A, C or G, start codon underscored) has been reported as a high-efficiency TIS, with positions -4, -3, -2, +4 and +5 as the most critical for efficient translation initiation (8,33). This consensus dictates the use of the CGx anticodon of ala-tRNA following incorporation of the AUG start codon for efficient translation initiation. Our best sequence TCGGTC (motif YYRRYVAUG–) is not fully in accordance with the earlier reported high-efficiency TIS motif RYMRMVAUGGC, as it deviates at position no. -6, -4 and -2. Moreover, neither does our TIS toolkit take into consideration the use of specific codons following the AUG start codon. Still, to further investigate if yeast-derived TIS variants spanning only positions -6 to -1 could also tune context-independent protein expression in mammalian cells, we decided to engineer CHO cells, the biotechnology workhorse for recombinant therapeutic protein production (59). Here, we constructed six CHO cell pools containing three different TISs derived from our FACS-based selection (Figures 1 and 5A) in combination with either meGFP or ZsGreen1, selected for their low sequence similarity, and optimized fluorescence intensity for CHO cells (Supplementary Figure S1) (48). Additionally, we created a cell pool containing the mammalian consensus TIS GCCACC (32) in combination with meGFP (Figure 5A and B; Supplementary Figure S10). First, testing meGFP expression in both yeast and CHO cells showed that the TIS strength was maintained between the two chassis and revealed fluorescence measurements with almost 10-fold variation between weakest and strongest TISs (Figure 5B). Importantly, considering the inherent efficiency of CHO cells for protein production, we observed that meGFP expression in combination with TIS TCGGTC was almost 50% stronger than the mammalian consensus TIS GCCACC (TIS no. 11) (Figure 5B). Moreover, just as was observed in yeast, TISs dictated the fluorescence of each of the reporter genes tested (Figure 5C). Finally, we observed strong linear correlations between fluorescence outputs from yeast versus CHO cells ($R^2 = 0.98$) as well as for mean meGFP versus ZsGreen fluorescence values across biological duplicates ($R^2 = 0.91$) (Figure 5D–E).

Tuning metabolic fluxes using modular TISs

The short length and context-independence of TISs make them particularly useful for engineering regulatory branch points of cellular metabolism, similar to earlier reports from MAGE-derived replacement of SD sequences in bacteria (60). To demonstrate an application of simple tuning of a metabolic branch point by the use of hexameric TIS variants, we aimed to tune metabolic fluxes through the native mevalonate biosynthesis pathway, and the *Xanthophyllomyces dendrorhous* 4-step β-carotene pathway (Figure 6A) (46). Specifically, this included targeting the genes encoding squalene synthase (*ERG9*) and the heterologous geranylgeranyl diphosphate (GGPP) synthase (*crtE*) at the farnesyl diphosphate (FPP) branch point (Figure 6A), both have

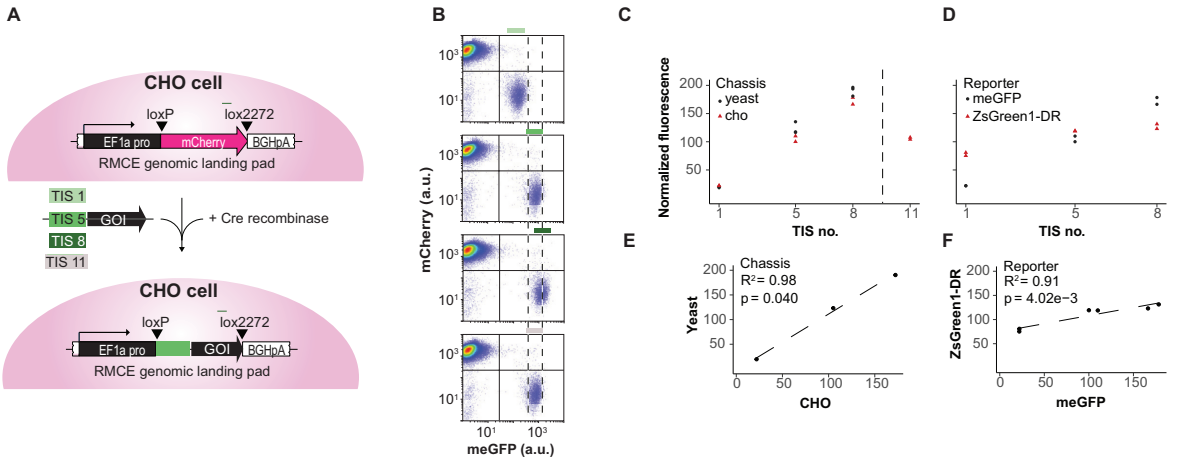

Figure 5. Comparison of interaction effects between TIS sequences and eukaryote chassis. (**A**) Schematic outline of the RMCE methodology used for introducing GOI with varying TISs into the genomic landing pad. In this study we compared hexameric TIS no. 1 (TGATAT), 5 (TAGGTT) and 8 (TCGGTC), with the mammalian consensus TIS no. 11 (GCCACC). (**B**) Scatter plots of CHO singlets with genomically integrated reporter meGFP in the context of TIS no. 1, 5, 8 or 11. Dashed lines indicate the distribution of the gated population from meGFP expressed in the context of TIS no. 11. (**C**) Comparison of yeast and CHO cells expressing meGFP under the control of varying TISs. (**D**) Comparison of two reporters (meGFP and ZsGreen1) expressed under the control of varying TISs. (**E**) Correlation between fluorescence measurements for three TIS sequences in yeast and CHO. (**F**) Correlation between fluorescence measurements for two reporters (meGFP and ZsGreen1) in CHO cells. In (E), the data shown are means of the normalized median fluorescence for each of the three TIS sequence with paired measurements from (C). In (C–F), the fluorescence values are means of normalized median fluorescence values from at least two biological replicates each based on at least 5000 single cell measurements.

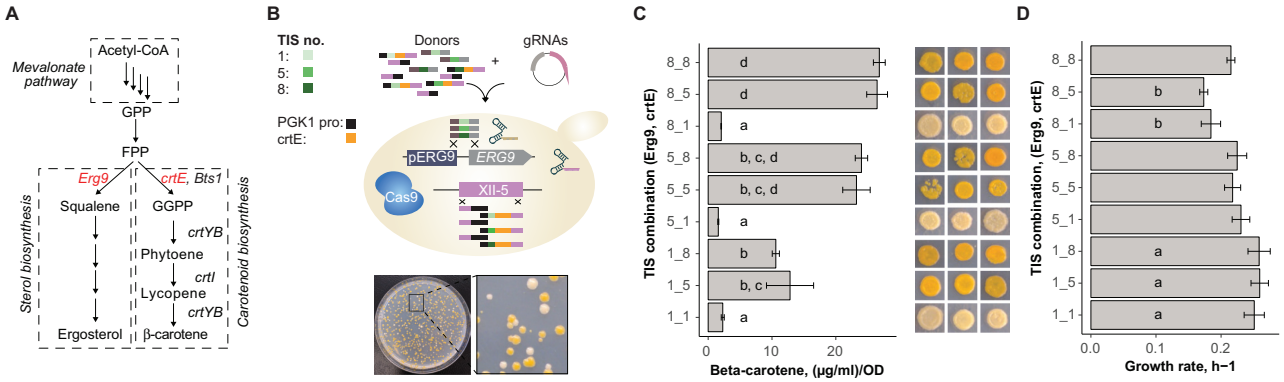

Figure 6. Effect of balancing Erg9 and crtE protein expression on carotenoid production. (**A**) Map of the carotenoid pathway and its connection to the native metabolism in yeast. The dashed lined boxes indicate sectioned native and heterologous metabolic pathways, with the branch point genes, *Erg9* and *crtE*, colored in red. (**B**) Schematic outline of the TISs tested and the genome engineering approach (top), and the resulting phenotypic landscape observed from multiplex TIS targeting of crtE (on Chr. XII-5) and *Erg9* loci on the yeast genome (bottom). (**C** and **D**) β-Carotene quantification and phenotype of carotenoid strains with variable duplex TISs controlling the expression of Erg9 and CrtE. Mean β-carotene content and maximum specific growth rate are shown (with standard errors, $n = 6$). Mean values with different lettering are significantly different according to pairwise Tukey HSD test ($P < 0.05$). Phenotypes are depicting three biological replicates.

earlier been reported to impact isoprenoid production in yeast (43,61,62).

Here, starting from a baseline strain with Cas9 and the 4-step β-carotene pathway genes integrated, we performed a one-pot transformation of a double guide RNA (gRNA) construct and repair templates introducing TISs (no. 1, 5 and 8) controlling the expression of ERG9 at the native site, as well as the *crtE* under the control of the PGK1 promoter genomically integrated at EasyClone site XII-5 (Figure 6B).

Following library transformation, colonies stably displayed diverse carotenoid-associated orange coloring and colony sizes (Figure 6B). Having observed the wide phenotypic distribution offered from combinatorial perturbation of TISs, we next re-constructed nine defined designs by duplex integration of TISs no. 1, 5 and 8 to control Erg9 and CrtE protein expression in all combinations in our background strain (Figure 6B) (46), and then quantified β-carotene levels as well as measured growth rates for all designs. From this analysis we observed up to 16-fold differences in β-carotene levels (Figure 6C), as well as up to 50% differences in growth rate (Figure 6D, and Supplementary Figure S11, $R^2 = 0.79$–0.99 with a mean $R^2 = 0.97$). Though no linear effect between fitness and production was observed, it is evident that stronger TISs (TISs 5 or 8) are needed to drive the expression of crtE in order to direct flux toward carotenoid production (Figure 6C). Interestingly, the strains with TIS 1 controlling ERG9 showed the highest growth rates. This finding is surprising in light of Erg9p

being an essential enzymatic step for conversion of FPP to squalene. However, acknowledging the intricate transcriptional and product-inhibited regulation of Erg9, low translation initiation efficiency of Erg9 could relieve ergosterol feedback inhibition and lead to upregulation of transcription (63,64). Alternatively, we could imagine the accumulation of toxic intermediates causing a reduction in growth rates for some of the strains with TISs 5 and 8 engineered to control Erg9 translation initiation (65).

Taken together, this example corroborates the simple design, rapid construction and testing of intricately regulated production and fitness landscapes offered from library transformations of hexameric TIS variants.

DISCUSSION

In this study, we have characterized and classified hexameric TISs according to their impact on protein expression in yeast and mammalian cells. Starting from three TIS libraries collectively covering 4739 TIS variants in yeast, we identified TISs that can tune protein expression up to 10-fold irrespective of the diverse genomic (38–54% similarity of the -54 to +13 positions, Supplementary Figure S1) and environmental (cell density or growth medium) conditions. Importantly, in terms of applicability, we showed that TIS TCGGTC was stronger than the mammalian consensus TIS GCCACC frequently used for protein production in CHO cells, and that a multiplex transformation of TIS variants targeting an essential metabolic branch point could be used to probe the production and fitness landscape of yeast cell factory designs. Though the combined use of large *de novo* synthesized TIS libraries and FACS screens to deduce sequence to function relationships has recently been reported in both bacteria and eukaryotes (6–9,66), the sequence space (positions -6 to -1) covered in this study is to our knowledge the smallest space systematically studied in broad genomic and environmental contexts, yet the dynamic range covered is similar to variants selected from larger TIS sequence spaces (7). Moreover, as the TISs characterized in this study only cover positions upstream the AUG start codon, protein expression of any ORF should technically be possible by a simple hexameric 5′-end primer extension using said ORF as a template. As such, both scalability and cost-effectiveness in both design and construction of engineered cells are ensured.

In the further positioning of our findings in relation to earlier studies, we find the five strongest TISs identified in our study have a purine at position -3 (Figure 3), consistent with earlier studies (33,34). Also, the degree of tunability observed in this study is similar to the ∼7-fold changes in protein expression observed from studies characterizing larger 5′-UTR sequence space (e.g positions -50 to -1 or -6 to +5) (7,8), underscoring the potential to use hexameric TISs for efficient protein expression tuning. Interestingly, among the five different fluorescent reporter genes tested in this study, the ones displaying the largest tunability in the context of varying TISs are the yeGFP and mammalian GFP (Figures 3 and 5; Supplementary Figure S1). The ORFs of these two genes are the only ones not having a guanine at position +4, otherwise reported to be important for efficient translation initiation (8,33–34), suggesting that the TISs identified in this study could be recalcitrant to ORF sequence diversity at this exact position.

More generally speaking, one immediate observation from studies of TISs in eukaryotes is that even though small TIS sequence spaces can robustly tune protein expression in a predictable manner, the degree of tuning is several orders of magnitude lower than the tuning offered by TIS variants in bacteria (6,67). This is largely due to more intricate regulatory mechanisms associated with translation initiation in eukaryotes compared to bacteria, including ribosome scanning mode-of-action, longer 5′-UTRs, 5′-end capping of mRNA, assembly of eukaryotic initiation factors, internal ribosome entry sites, uAUGs and uORFs observed in eukaryotes (10). Yet, engineering excessively short 5′-UTR (≤20 nt) may not provide TISs with higher tunability, as genome-wide mapping of yeast 5′-UTRs with such short 5′-UTRs has been observed to be detrimental to translation initiation control and exhibit below-average translational efficiency (68), and hence is not considered a viable route to dereplicate the impact combinations of native 5′-UTR elements would have on translation initiation.

Furthermore, looking ahead, it is important to consider system-level limitations of protein expression (69), and continue to improve current and new models for predicting TIS strengths in broad genomic contexts and, cellular and environmental conditions (5–10). Also, from the range of fluorescent outputs observed in our TIS libraries in three promoter contexts (Figure 2), it is evident that the native transcriptional regulation, conferred by promoter usage, controls the absolute quantitative impact TISs will have on protein expression, as was recently reported from genome-wide studies in yeast (58). As such, we envision that in order to engineer synthetic translation initiation elements with higher dynamic output ranges based on the existing features of the translation machinery, a more detailed understanding of both *cis* and *trans* initiation mechanisms is expected to enhance our ability to predictably control larger spans of protein expression levels.

Finally, when quantifying changes in protein expression within an order of magnitude as observed from varying short TISs, mitigating experimental noise and conforming to standardized experimental procedures become essential for deducting sequence–function relationships (7,70). With the ongoing development of advanced genome engineering technologies, especially in relation to <100-bp edits (71,72), and the drop in DNA synthesis costs, we expect that TISs will be particularly useful baits for multiplex targeting, tuning and optimization of protein expression levels in robust genomic contexts, thereby expectedly improving signal-to-noise ratios, and ultimately enabling predictable and rational tuning of genetic circuits and cellular behavior.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank Saranya Nallapareddy for help with CHO transfections, Nachon Charanyanonda Petersen for help with flow cytometry analyses, and Larissa

Tramontin and Kanchana Kildegaard for help with HPLC. Also, a warm thanks to colleagues at the Novo Nordisk Foundation Center for Biosustainability for fruitful discussions and comments.

FUNDING

Novo Nordisk Foundation; European Commission Horizon 2020 programme (PAcMEN, No. 722287). Funding for open access charge: Novo Nordisk Foundation and European Commission HZ2020 Programme.
Conflict of interest statement. J.D.K. has a financial interest in Amyris, Lygos, Demetrix, Constructive Biology, Maple Bio and Napigen.

REFERENCES

1. Sonenberg,N. and Hinnebusch,A.G. (2009) Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, **136**, 731–745.
2. Jackson,R.J., Hellen,C.U.T. and Pestova,T.V. (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.
3. Allen,G.S., Zavialov,A., Gursky,R., Ehrenberg,M. and Frank,J. (2005) The cryo-EM structure of a translation initiation complex from Escherichia coli. *Cell*, **121**, 703–712.
4. Seo,S.W., Yang,J.-S., Kim,I., Yang,J., Min,B.E., Kim,S. and Jung,G.Y. (2013) Predictive design of mRNA translation initiation region to control prokaryotic translation efficiency. *Metab. Eng.*, **15**, 67–74.
5. Salis,H.M. (2011) The ribosome binding site calculator. *Methods Enzymol.*, **498**, 19–42.
6. Bonde,M.T., Pedersen,M., Klausen,M.S., Jensen,S.I., Wulff,T., Harrison,S., Nielsen,A.T., Herrgård,M.J. and Sommer,M.O.A. (2016) Predictable tuning of protein expression in bacteria. *Nat. Methods*, **13**, 233–236.
7. Dvir,S., Velten,L., Sharon,E., Zeevi,D., Carey,L.B., Weinberger,A. and Segal,E. (2013) Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, E2792–E2801.
8. Noderer,W.L., Flockhart,R.J., Bhaduri,A., Diaz de Arce,A.J., Zhang,J., Khavari,P.A. and Wang,C.L. (2014) Quantitative analysis of mammalian translation initiation sites by FACS-seq. *Mol. Syst. Biol.*, **10**, 748.
9. Cuperus,J.T., Groves,B., Kuchina,A., Rosenberg,A.B., Jojic,N., Fields,S. and Seelig,G. (2017) Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res.*, **27**, 2015–2024.
10. Decoene,T., Peters,G., De Maeseneire,S.L. and De Mey,M. (2018) Toward predictable 5′UTRs in saccharomyces cerevisiae: development of a yUTR calculator. *ACS Synth. Biol.*, **7**, 622–634.
11. Ben-Yehezkel,T., Atar,S., Zur,H., Diament,A., Goz,E., Marx,T., Cohen,R., Dana,A., Feldman,A., Shapiro,E. *et al.* (2015) Rationally designed, heterologous S. cerevisiae transcripts expose novel expression determinants. *RNA Biol.*, **12**, 972–984.
12. Shine,J. and Dalgarno,L. (1974) The 3′-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. U.S.A.*, **71**, 1342–1346.
13. Ludwig,P., Huber,M., Lehr,M., Wegener,M., Zerulla,K., Lange,C. and Soppa,J. (2018) Non-canonical Escherichia coli transcripts lacking a Shine-Dalgarno motif have very different translational efficiencies and do not form a coherent group. *Microbiology*, **164**, 646–658.
14. Leppek,K., Das,R. and Barna,M. (2017) Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat. Rev. Mol. Cell Biol.*, **19**, 158–174.
15. Grillo,G., Turi,A., Licciulli,F., Mignone,F., Liuni,S., Banfi,S., Gennarino,V.A., Horner,D.S., Pavesi,G., Picardi,E. *et al.* (2010) UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **38**, D75–D80.
16. Tuller,T., Kupiec,M. and Ruppin,E. (2009) Co-evolutionary networks of genes and cellular processes across fungal species. *Genome Biol.*, **10**, R48.
17. Ringnér,M. and Krogh,M. (2005) Folding free energies of 5′-UTRs impact post-transcriptional regulation on a genomic scale in yeast. *PLoS Comput. Biol.*, **1**, e72.
18. Hinnebusch,A.G., Dever,T.E. and Asano,K. (2007) 9 Mechanism of translation initiation in the yeast saccharomyces cerevisiae. *Cold Spring Harbor Monogr. Arch.*, **48**, 225–268.
19. Zhang,Z. and Dietrich,F.S. (2005) Mapping of transcription start sites in Saccharomyces cerevisiae using 5′ SAGE. *Nucleic Acids Res.*, **33**, 2838–2851.
20. Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.
21. Robbins-Pianka,A., Rice,M.D. and Weir,M.P. (2010) The mRNA landscape at yeast translation initiation sites. *Bioinformatics*, **26**, 2651–2655.
22. Kochetov,A.V. (2005) AUG codons at the beginning of protein coding sequences are frequent in eukaryotic mRNAs with a suboptimal start codon context. *Bioinformatics*, **21**, 837–840.
23. Nakagawa,S., Niimura,Y., Gojobori,T., Tanaka,H. and Miura,K.-I. (2008) Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.*, **36**, 861–871.
24. Tzani,I., Ivanov,I.P., Andreev,D.E., Dmitriev,R.I., Dean,K.A., Baranov,P.V., Atkins,J.F. and Loughran,G. (2016) Systematic analysis of the PTEN 5′ leader identifies a major AUU initiated proteoform. *Open Biol.*, **6**, 150203.
25. Ben-Yehezkel,T., Zur,H., Marx,T., Shapiro,E. and Tuller,T. (2013) Mapping the translation initiation landscape of an S. cerevisiae gene using fluorescent proteins. *Genomics*, **102**, 419–429.
26. Diaz de Arce,A.J., Noderer,W.L. and Wang,C.L. (2018) Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. *Nucleic Acids Res.*, **46**, 985–994.
27. Chew,G.-L., Pauli,A. and Schier,A.F. (2016) Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.*, **7**, 11663.
28. Tholen,M., Hillebrand,L.E., Tholen,S., Sedelmeier,O., Arnold,S.J. and Reinheckel,T. (2014) Out-of-frame start codons prevent translation of truncated nucleo-cytosolic cathepsin L in vivo. *Nat. Commun.*, **5**, 4931.
29. Zur,H. and Tuller,T. (2013) New universal rules of eukaryotic translation initiation fidelity. *PLoS Comput. Biol.*, **9**, e1003136.
30. Kozak,M. (1986) Influences of mRNA secondary structure on initiation by eukaryotic ribosomes. *Proc. Natl. Acad. Sci. U.S.A.*, **83**, 2850–2854.
31. Kozak,M. (1987) At least six nucleotides preceding the AUG initiator codon enhance translation in mammalian cells. *J. Mol. Biol.*, **196**, 947–950.
32. Kozak,M. (1987) An analysis of 5′-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125–8148.
33. Kozak,M. (1995) Adherence to the first-AUG rule when a second AUG codon follows closely upon the first. *Proc. Natl. Acad. Sci. U.S.A.*, **92**, 2662–2666.
34. Harte,R.A., Farrell,C.M., Loveland,J.E., Suner,M.-M., Wilming,L., Aken,B., Barrell,D., Frankish,A., Wallin,C., Searle,S. *et al.* (2012) Tracking and coordinating an international curation effort for the CCDS Project. *Database*, **2012**, bas008.
35. Pesole,G., Gissi,C., Grillo,G., Licciulli,F., Liuni,S. and Saccone,C. (2000) Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene*, **261**, 85–91.
36. Nour-Eldin,H.H., Hansen,B.G., Nørholm,M.H.H., Jensen,J.K. and Halkier,B.A. (2006) Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res.*, **34**, e122.
37. Gietz,R.D. and Schiestl,R.H. (2007) Quick and easy yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat. Protoc.*, **2**, 35–37.
38. Jensen,N.B., Strucko,T., Kildegaard,K.R., David,F., Maury,J.J., Mortensen,U.H., Forster,J., Nielsen,J. and Borodina,I. (2014) EasyClone: Method for iterative chromosomal integration of multiple genes in Saccharomyces cerevisiae. *FEMS Yeast Res.*, **14**, 238–248.
39. Cormack,B.P., Bertram,G., Egerton,M., Gow,N.A., Falkow,S. and Brown,A.J. (1997) Yeast-enhanced green fluorescent protein

(yEGFP): a reporter of gene expression in Candida albicans. *Microbiology*, **143**, 303–311.
40. Bernard,P., Gabant,P., Bahassi,E.M. and Couturier,M. (1994) Positive-selection vectors using the F plasmid ccdB killer gene. *Gene*, **148**, 71–74.
41. Jessop-Fabre,M.M., Jakočiūnas,T., Stovicek,V., Dai,Z., Jensen,M.K., Keasling,J.D. and Borodina,I. (2016) EasyClone-MarkerFree: a vector toolkit for marker-less integration of genes into Saccharomyces cerevisiae via CRISPR-Cas9. *Biotechnol. J.*, **11**, 1110–1117.
42. Jakočiūnas,T., Rajkumar,A.S., Zhang,J., Arsovska,D., Rodriguez,A., Jendresen,C.B., Skjødt,M.L., Nielsen,A.T., Borodina,I., Jensen,M.K. *et al.* (2015) CasEMBLR: Cas9-Facilitated multiloci genomic integration of in vivo assembled DNA parts in Saccharomyces cerevisiae. *ACS Synth. Biol.*, **4**, 1226–1234.
43. Jakočiūnas,T., Bonde,I., Herrgård,M., Harrison,S.J., Kristensen,M., Pedersen,L.E., Jensen,M.K. and Keasling,J.D. (2015) Multiplex metabolic pathway engineering using CRISPR/Cas9 in Saccharomyces cerevisiae. *Metab. Eng.*, **28**, 213–222.
44. Kaishima,M., Ishii,J., Matsuno,T., Fukuda,N. and Kondo,A. (2016) Expression of varied GFPs in Saccharomyces cerevisiae: codon optimization yields stronger than expected expression and fluorescence intensity. *Sci. Rep.*, **6**, 35932.
45. Lee,S., Lim,W.A. and Thorn,K.S. (2013) Improved blue, green, and red fluorescent protein tagging vectors for S. cerevisiae. *PLoS One*, **8**, e67902.
46. Verwaal,R., Wang,J., Meijnen,J.P., Visser,H., Sandmann,G., Van Den Berg,J.A. and Van Ooyen,A.J.J. (2007) High-level production of beta-carotene in Saccharomyces cerevisiae by successive transformation with carotenogenic genes from Xanthophyllomyces dendrorhous. *Appl. Environ. Microbiol.*, **73**, 4342–4350.
47. Lee,J.S., Kallehauge,T.B., Pedersen,L.E. and Kildegaard,H.F. (2015) Site-specific integration in CHO cells mediated by CRISPR/Cas9 and homology-directed DNA repair pathway. *Sci. Rep.*, **5**, 8572.
48. Grav,L.M., Lee,J.S., Gerling,S., Kallehauge,T.B., Hansen,A.H., Kol,S., Lee,G.M., Pedersen,L.E. and Kildegaard,H.F. (2015) One-step generation of triple knockout CHO cell lines using CRISPR/Cas9 and fluorescent enrichment. *Biotechnol. J.*, **10**, 1446–1456.
49. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10–12.
50. Petzoldt,T. (2017) growthrates: Estimate Growth Rates from Experimental Data. *R package version 0.7.1*. https://CRAN.R-project.org/package=growthrates.
51. Kildegaard,K.R., Adiego-Pérez,B., Doménech Belda,D., Khangura,J.K., Holkenbrink,C. and Borodina,I. (2017) Engineering of Yarrowia lipolytica for production of astaxanthin. *Synth. Syst. Biotechnol.*, **2**, 287–294.
52. Kozak,M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell*, **44**, 283–292.
53. Lee,M.E., Aswani,A., Han,A.S., Tomlin,C.J. and Dueber,J.E. (2013) Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Res.*, **41**, 10668–10678.
54. Tuite,M.F., Dobson,M.J., Roberts,N.A., King,R.M., Burke,D.C., Kingsman,S.M. and Kingsman,A.J. (1982) Regulated high efficiency expression of human interferon-alpha in Saccharomyces cerevisiae. *EMBO J.*, **1**, 603–608.
55. Zhang,L., Patel,H.N., Lappe,J.W. and Wachter,R.M. (2006) Reaction progress of chromophore biogenesis in green fluorescent protein. *J. Am. Chem. Soc.*, **128**, 4766–4772.
56. Salis,H.M., Mirsky,E.A. and Voigt,C.A. (2009) Automated design of synthetic ribosome binding sites to control protein expression. *Nat. Biotechnol.*, **27**, 946–950.
57. Reider Apel,A., d'Espaux,L., Wehrs,M., Sachs,D., Li,R.A., Tong,G.J., Garber,M., Nnadi,O., Zhuang,W., Hillson,N.J. *et al.* (2017) A Cas9-based toolkit to program gene expression in *Saccharomycescerevisiae*. *Nucleic Acids Res.*, **45**, 496–508.
58. Lahtvee,P.-J., Sánchez,B.J., Smialowska,A., Kasvandik,S., Elsemman,I.E., Gatto,F. and Nielsen,J. (2017) Absolute quantification of protein and mRNA abundances demonstrate variability in Gene-Specific translation efficiency in yeast. *Cell Syst.*, **4**, 495–504.
59. Kim,J.Y., Kim,Y.-G. and Lee,G.M. (2012) CHO cells in biotechnology for production of recombinant proteins: current state and further potential. *Appl. Microbiol. Biotechnol.*, **93**, 917–930.
60. Wang,H.H., Isaacs,F.J., Carr,P.A., Sun,Z.Z., Xu,G., Forest,C.R. and Church,G.M. (2009) Programming cells by multiplex genome engineering and accelerated evolution. *Nature*, **460**, 894–898.
61. Ro,D.-K., Paradise,E.M., Ouellet,M., Fisher,K.J., Newman,K.L., Ndungu,J.M., Ho,K.A., Eachus,R.A., Ham,T.S., Kirby,J. *et al.* (2006) Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature*, **440**, 940–943.
62. Mitchell,L.A., Chuang,J., Agmon,N., Khunsriraksakul,C., Phillips,N.A., Cai,Y., Truong,D.M., Veerakumar,A., Wang,Y., Mayorga,M. *et al.* (2015) Versatile genetic assembly system (VEGAS) to assemble pathways for expression in S. cerevisiae. *Nucleic Acids Res.*, **43**, 6620–6630.
63. Asadollahi,M.A., Maury,J., Møller,K., Nielsen,K.F., Schalk,M., Clark,A. and Nielsen,J. (2008) Production of plant sesquiterpenes in Saccharomyces cerevisiae: effect of ERG9 repression on sesquiterpene biosynthesis. *Biotechnol. Bioeng.*, **99**, 666–677.
64. Smith,S.J., Crowley,J.H. and Parks,L.W. (1996) Transcriptional regulation by ergosterol in the yeast Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **16**, 5427–5432.
65. Martin,V.J.J., Pitera,D.J., Withers,S.T., Newman,J.D. and Keasling,J.D. (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. *Nat. Biotechnol.*, **21**, 796–802.
66. Ben Yehezkel,T., Rival,A., Raz,O., Cohen,R., Marx,Z., Camara,M., Dubern,J.-F., Koch,B., Heeb,S., Krasnogor,N. *et al.* (2016) Synthesis and cell-free cloning of DNA libraries using programmable microfluidics. *Nucleic Acids Res.*, **44**, e35.
67. Espah Borujeni,A., Channarasappa,A.S. and Salis,H.M. (2014) Translation rate is controlled by coupled trade-offs between site accessibility, selective RNA unfolding and sliding at upstream standby sites. *Nucleic Acids Res.*, **42**, 2646–2659.
68. Arribere,J.A. and Gilbert,W.V. (2013) Roles for transcript leaders in translation and mRNA decay revealed by transcript leader sequencing. *Genome Res.*, **23**, 977–987.
69. Huang,M., Bao,J., Hallström,B.M., Petranovic,D. and Nielsen,J. (2017) Efficient protein production by yeast requires global tuning of metabolism. *Nat. Commun.*, **8**, 1131.
70. Canton,B., Labno,A. and Endy,D. (2008) Refinement and standardization of synthetic biological parts and devices. *Nat. Biotechnol.*, **26**, 787–793.
71. Garst,A.D., Bassalo,M.C., Pines,G., Lynch,S.A., Halweg-Edwards,A.L., Liu,R., Liang,L., Wang,Z., Zeitoun,R., Alexander,W.G. *et al.* (2016) Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.*, **35**, 48–55.
72. Barbieri,E.M., Muir,P., Akhuetie-Oni,B.O., Yellman,C.M. and Isaacs,F.J. (2017) Precise editing at DNA replication forks enables multiplex genome engineering in eukaryotes. *Cell*, **171**, 1453–1467.