Nucleic Acids Research, 2025, **53**, gkaf528
https://doi.org/10.1093/nar/gkaf528
Synthetic Biology and Bioengineering

High-resolution profiling reveals coupled transcriptional and translational regulation of transgenes

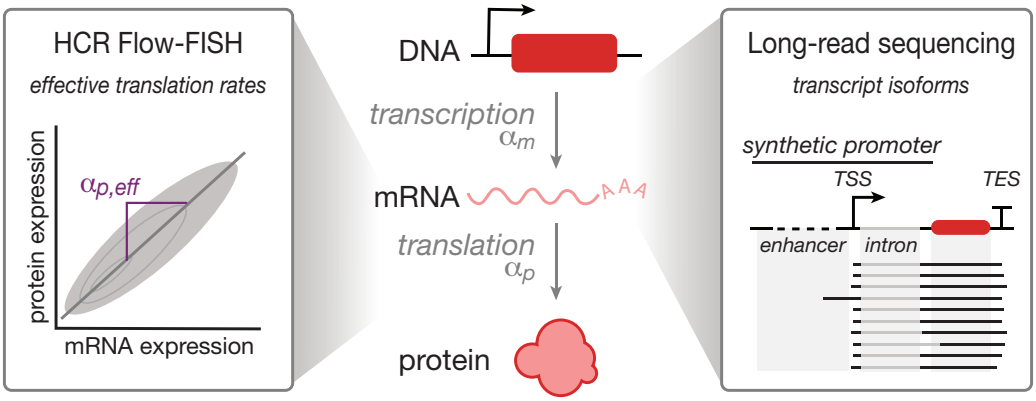
Emma L. Peterman [1], Deon S. Ploessl [1], Kasey S. Love [2], Valeria Sanabria [3], Rachel F. Daniels [3], Christopher P. Johnstone [1], Diya R. Godavarti [4], Sneha R. Kabaria [1], Conrad G. Oakes [5], Athma A. Pai [3], Kate E. Galloway [1,*]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[3]RNA Therapeutics Institute, University of Massachusetts Chan Medical School, Worcester, MA 01605, United States
[4]School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA 30332, United States
[5]Department of Bioengineering, California Institute of Technology, Pasadena, CA 91125, United States
[*]To whom correspondence should be addressed. Email: katiegal@mit.edu

Abstract

Concentrations of RNAs and proteins provide important determinants of cell fate. Robust gene circuit design requires an understanding of how the combined actions of individual genetic components influence both messenger RNA (mRNA) and protein levels. Here, we simultaneously measure mRNA and protein levels in single cells using hybridization chain reaction Flow-FISH (HCR Flow-FISH) for a set of commonly used synthetic promoters. We find that promoters generate differences in both the mRNA abundance and the effective translation rate of these transcripts. Stronger promoters not only transcribe more RNA but also show higher effective translation rates. While the strength of the promoter is largely preserved upon genome integration with identical elements, the choice of polyadenylation signal and coding sequence can generate large differences in the profiles of the mRNAs and proteins. We used long-read direct RNA sequencing to define the transcription start and splice sites of common synthetic promoters and independently vary the defined promoter and 5′ UTR sequences in HCR Flow-FISH. Together, our high-resolution profiling of transgenic mRNAs and proteins offers insight into the impact of common synthetic genetic components on transcriptional and translational mechanisms. By developing a novel framework for quantifying expression profiles of transgenes, we have established a system for building more robust transgenic systems.

Graphical abstract



Introduction

Intracellular levels of key proteins and RNAs govern gene regulatory programs and cell states. Similarly, levels of RNA and protein components can set the activity of gene circuits and influence the robustness and performance of the circuits. Eukaryotic gene expression requires multiple co- and post-transcriptional processing steps of messenger RNA (mRNA) transcripts, including splicing, 3′ end cleavage and polyadenylation, and nuclear export [1, 2]. The degree of post-transcriptional or post-translational processing influences the

Received: December 3, 2024. **Revised:** April 24, 2025. **Editorial Decision:** May 7, 2025. **Accepted:** May 30, 2025

correlation between mRNA and protein levels [3]. This multi-scale process generates endogenous mRNA and protein levels that are much less correlated in eukaryotes than in prokaryotes [4]. While the number of studies profiling the expression of endogenous RNA transcripts is rapidly growing [5, 6], there is still limited understanding of the abundance and composition of mRNA expressed from synthetic transgenic systems. Developing a predictive understanding of the levels of mRNA isoforms and their rates of processing may improve the design of gene circuits in diverse cell types—including primary cells and induced pluripotent stem cells (iPSCs) [7–11].

Forward design of gene circuits requires composable and well-characterized genetic elements. Given the distribution of gene expression profiles from transgenes, models that accurately predict the performance of dynamic circuits require parameters that capture the ensemble features such as the mean and variance of mRNA and protein molecules. Accurate estimation of the average level of protein expression can inform selection of genetic elements for predictive design [12]. Tracking distributions of both transgenic mRNAs and proteins over time can augment the design of systems that amplify or attenuate noise and reveal the underlying network structures of biological systems [13–15]. However, synthetic parts are often characterized by the mean level of expression for a single mRNA or protein species. The complex nature of gene regulation in mammalian cells calls for high-resolution, systematic characterization of genetic parts across transcriptional and translational processes. Defining the combined effects of genetic elements, such as promoters, polyadenylation signals (PASs), and untranslated regions (UTRs), on gene expression profiles and transcript isoforms will offer insight into sources of variability. For instance, are mRNA isoforms uniform within a single construct? Or do transgenes generate a variety of isoforms that may have unique processing rates? Understanding both mRNA levels and compositions will support improved design of transgenic systems in mammalian cells.

As multi-modal circuits rely on levels of both RNAs and proteins, predictable circuit design requires high-resolution characterization of both molecules and their distributions across populations. Previous characterizations of genetic parts relied on easily assayable metrics of expression, such as protein fluorescence or enzymatic activity, which cannot capture species involved in post-transcriptional regulation such as microRNAs, alternative splicing isoforms, and ribozymes [2–4, 12, 16–18] (Fig. 1A). Forward design of RNA-based control systems requires quantification of RNA levels in single cells [19]. Bulk methods such as reverse transcription quantitative polymerase chain reaction (RT-qPCR) obscure variance across single cells [20]. While transcriptional imaging systems and single-molecule fluorescence *in situ* hybridization enable high-resolution quantification of transcript profiles in single cells, these methods suffer from being very low throughput [21–24]. Alternatively, flow cytometry-based RNA fluorescence *in situ* hybridization (Flow-FISH) offers a method to measure levels of specific RNAs in single cells. Specifically, hybridization chain reaction Flow-FISH (HCR Flow-FISH) reduces background fluorescence, enabling high-throughput, sensitive RNA readouts that can be coupled with simultaneous protein quantification in a single cell [25, 26]. Thus, HCR Flow-FISH allows us to measure single-cell mRNA distributions while integrating existing protein expression analysis pipelines for a more comprehensive characterization of existing and novel genetic elements. HCR Flow-FISH paired with methods to analyze full-length isoforms would enable an understanding of how genetic elements influence variability in transcriptional and translational processes.

In this work, we use HCR Flow-FISH to simultaneously quantify levels of transgenic mRNAs and proteins in single cells. With these data, we can quantify the impact of individual genetic elements on different gene regulatory steps. Specifically, we characterize a panel of commonly used constitutive promoter sequences in HEK293T cells and benchmark expression levels against three inducible promoter systems (Tet-On, COMET [27], and synZiFTR [28]). We find that promoter sequences impact both the abundance of mRNA transcripts and the effective translation rate of these transcripts. Moreover, the combination of promoter, coding, and 3′ UTR sequences alter the effective translation rate, suggesting a role for UTRs in transgene regulation. To examine RNA isoforms—including their UTRs—at high resolution, we use long-read sequencing to profile full-length transcripts from transgenes. We find that mature transgenic transcripts are highly uniform, rarely impacted by local sequence context, and exert minimal burden on endogenous gene expression. Together, our work to establish high-resolution profiling of expression distributions and isoforms of transgenic mRNAs offers a novel framework for systematically comparing native and synthetic gene regulation and building more robust transgenic systems.

Materials and methods

Cloning

Expression plasmids were generated using a multi-level Golden Gate cloning scheme. First, individual genetic part fragments were amplified or digested from commercial DNA sources. Each fragment was inserted into a corresponding part positioning vector (pPV) backbone via Gibson Assembly using Hifi DNA Assembly Master Mix (NEB, M5520). On each pPV, the region of interest is flanked by BsaI restriction sites detailed in Supplementary Table S3. A full list of pPVs used in this work and their DNA sources are reported in Supplementary Table S4. Full transcriptional units were assembled in a BsaI (NEB, R3733L) Golden Gate reaction to yield the kanamycin-resistant plasmids (pShips) used in transfection experiments. In a second round of PaqCI (NEB, R0745L) Golden Gate reactions, full transcriptional units were inserted into backbones for genomic integration (pHarbors) Bxb1 recombinase, PiggyBac transposase, and lentivirus.

To clone the homology-directed repair (HDR) donor template for targeting *Rogi2* with a Bxb1 landing pad (LP), 5′ and 3′ *Rogi2* homology arms were PCR amplified from HEK293T genomic DNA. To facilitate PCR genotyping of CRISPR-edited clones, the lengths of the homology arms were selected based on the co-design of randomly generated 5′ and 3′ barcode sequences and genotyping primers with primer pairs that (i) flank the HDR junction of the donor DNA and the *Rogi2* locus and (ii) were not predicted by PrimerBLAST to produce off-target amplicons of similar size. The resulting barcodes were encoded on the primers used to amplify the *Rogi2* 5′ and 3′ homology arms. These were assembled with pHarbor backbone fragments PCR amplified with primers that encoded the

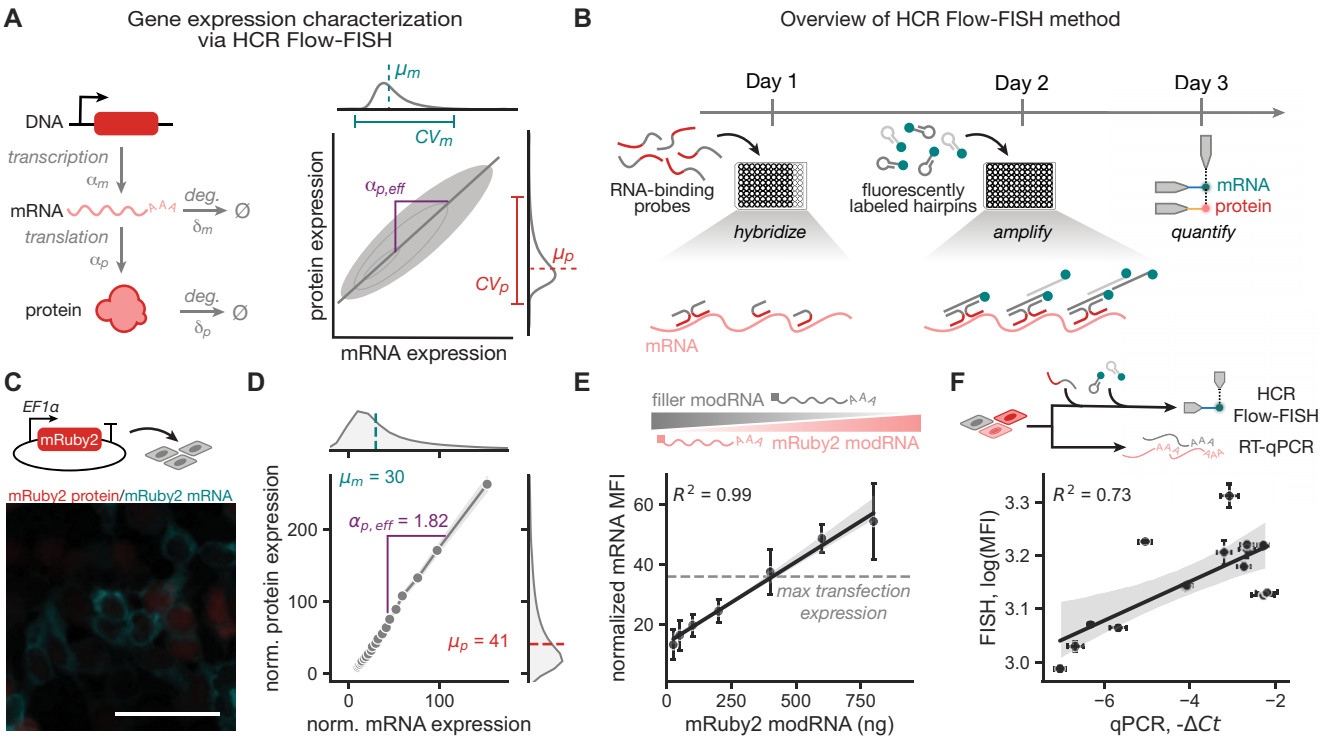


Figure 1. HCR Flow-FISH enables high-throughput quantification of mRNA and protein levels in single cells via flow cytometry. (**A**) In a simple model of gene expression, mRNA and protein levels (μ_m and μ_p) are governed by four main parameters: transcription rate (α_m), mRNA degradation rate (δ_m), translation rate (α_p), and protein degradation rate (δ_p). HCR Flow-FISH data allows for calculation of an effective translation rate, $\alpha_{p,\,eff} = \Delta\mu_p/\Delta\mu_m \propto \alpha_p/\delta_p$, and coefficient of variation (CV). (**B**) RNA-binding probes specific to the mRNA species of interest are first added to fixed and permeabilized cells. Fluorescently labeled hairpins complementary to the probes are then added to amplify the FISH signal. (**C**) Sample HCR FISH imaging for HEK293T cells transfected with an EF1α-mRuby2-bGH plasmid and labeled with Alexa Fluor™ 647 HCR amplifiers. Scale bar represents 50 μm. (**D**) Sample HCR Flow-FISH for HEK293T cells transfected with an EF1α-mRuby2-bGH plasmid and labeled with Alexa Fluor™ 514 HCR amplifiers. Data for one representative biological replicate are binned by transfection marker level into 20 equal-quantile groups. Points represent geometric mean of protein and mRNA expression (mean fluorescence intensity, MFI) for cells in each bin, and shaded regions represent the 95% confidence interval. Effective translation rate is calculated as the slope of a line fitted to the binned data ($R^2 = 0.999$). Normalized expression is calculated as the fold change of fluorescence intensity relative to a nontransfected sample. (**E**) Measurement of HCR Flow-FISH signal for HEK293T cells transfected with varying dosages of mRuby2 modRNA at 4 h post-transfection. Error bars represent the standard deviation across three biological replicates, and the shaded region represents the bootstrapped 95% confidence interval of the linear regression ($P = 7 \times 10^{-6}$). Gray dashed line indicates the mean mRNA MFI for the highest expressing construct (CAG-mRuby2-bGH) in HEK293T transfection. Normalized fluorescence is calculated as the fold change of fluorescence intensity relative to a nontransfected sample. (**F**) Measurement of mRuby2 mRNA level via HCR Flow-FISH (y-axis) and RT-qPCR (x-axis) for PiggyBac-integrated HEK293T cells with varying levels of mRuby2 expression. Error bars represent the standard deviation between technical replicates, and the shaded region represents the bootstrapped 95% confidence interval of the linear regression ($P = 2 \times 10^{-3}$). Each point represents an individual biological replicate. All HCR Flow-FISH data are in arbitrary units from a flow cytometer.

Rogi2 protospacer and protospacer adjacent motif (PAM) sequences. Fragments were designed such that each homology arm is flanked by the *Rogi2* protospacer and PAM sequence needed for in trans paired nicking (ITPN) editing [29]. The PCR fragments were assembled into a pHarbor using an Esp3I (NEB, R0734L) Golden Gate reaction.

The LP architecture inserted at *Rogi2* was based on the STRAIGHT-IN platform [30, 31] with modifications. The LP consists of two transcriptional units Supplementary Fig. S14. The 5′ unit consists of an EF1α-BsdR-bGH cassette, which confers blasticidin resistance and is used for selecting cells that underwent HDR. The 3′ unit consists of a Bxb1 attB site and a PuroR-bGH cassette lacking a promoter and start codon, which is used for enriching cells which underwent Bxb1-mediated integration of attP donor plasmids at the LP. Each transcriptional unit was assembled in a BsaI Golden Gate reaction to generate pShips, which were subsequently assembled into the *Rogi2*-targeting pHarbor in a PaqCI Golden Gate reaction, yielding the final ITPN donor plasmid used to create the HEK293T *Rogi2* LP cell line.

Cell culture

HEK293T

HEK293T cells (ATCC, CRL-3216) were cultured using Dulbecco's Modified Eagle's Medium (DMEM, Genesee Scientific, 25-501) plus 10% fetal bovine serum (FBS, Genesee Scientific, 25-514H) and incubated at 37°C with 5% CO2. Cells were passaged at 80%–90% confluence, in which spent media was aspirated, and cells were washed with PBS (Sigma–Aldrich, P4417-100TAB) and then subsequently dissociated with 0.25% Trypsin-EDTA (Genesee Scientific, 25-510) diluted in phosphate-buffered saline (PBS). After 4 min, cells were spun down at 400 rcf for 5 min, resuspended in media, then transferred to a new flask. Media was replaced with fresh DMEM + 10% FBS every 2–3 days.

CHO-K1

CHO-K1 cells (ATCC, CCL-61) were cultured using DMEM/F12 (Corning, 10-090-CV) plus 10% FBS and

incubated at 37°C with 5% CO2. CHO-K1 cells were passaged identically to HEK293T cells.

iPS11

iPS11 cells (Alstem, iPS11) were cultured using mTeSR™ Plus (STEMCELL Technologies, 100-1130) on Geltrex™-coated (Thermo Fisher Scientific, A1413302) plates and incubated at 37°C with 5% CO2. Cells were passaged in clumps using ReLeSR™ (STEMCELL Technologies, 100-0484) according to the manufacturer's instructions.

Transfection

HEK293T

In preparation for experiments, HEK293T cells were counted using a hemocytometer, seeded with 0.1% gelatin coating (Sigma–Aldrich, G1890-100G) at a density of 150 000 cells per 12-well, and transfected 24 h later. For imaging experiments, cells were seeded in Geltrex-coated glass-bottom 96-well plates. Transfection was performed using linear polyethylenimine (PEI, Fisher Scientific, 4389603). Transfection mixes were prepared using a ratio of 4 μg PEI to 1 μg DNA. First, a master mix of PEI and KnockOut™ DMEM (Thermo Fisher Scientific, 10-829-018) was prepared and incubated for a minimum of 10 min. This master mix was then added to DNA mixes containing 450 ng of output plasmid and 450 ng of transfection marker plasmid per 12-well. These condition mixes were further incubated for 10–15 min and then added on top of the growth media in the plate.

For experiments with constitutive promoters, media was replaced with fresh DMEM + 10% FBS after 24 h. At 2 days post transfection, HEK293T cells were dissociated by adding 0.25% Trypsin-EDTA (diluted in PBS) for 4 min, followed by quenching with an equal volume of DMEM + 10% FBS. After centrifuging at 500 rcf for 5 min, cells were resuspended in PBS and transferred to a v-bottom plate for flow cytometry or subsequent HCR Flow-FISH.

For experiments with inducible promoters, 24 h after transfection media was replaced with fresh DMEM + 10% FBS containing the corresponding small molecule inducer or solvent control. Inducer stocks were prepared as follows: doxycycline (dox; Sigma–Aldrich, D3447) in water at 1 mg/m, grazoprevir (GZV; MedChem Express, HY-15298) in dimethyl sulfoxide (DMSO) at 1 mM, and rapamycin (Rap; Millipore Sigma, 553210) in DMSO at 200 μM. Grazoprevir and rapamycin stocks were stored at −80°C until use, then kept at 4°C for up to 2 weeks. Doxycycline stocks were stored at −20°C until use, then kept at 4°C. Small molecule stocks were diluted in DMEM + 10% FBS to the following concentrations for experiments: 1 μg/ml doxycycline, 1 μM grazoprevir, and 0.1 μM rapamycin. Two days after small molecule addition (3 days post-transfection), HEK293T cells were dissociated by adding 0.25% Trypsin-EDTA (diluted in PBS) for 4 min, followed by quenching with an equal volume of DMEM + 10% FBS. After centrifuging at 500 rcf for 5 min, cells were resuspended in PBS and transferred to a v-bottom plate for flow cytometry or subsequent HCR Flow-FISH.

CHO-K1

Similar to HEK293T cells, CHO-K1 cells were seeded 24 h prior to transfection with 0.1% gelatin coating at a density of 150 000 cells per 12-well. Using PEI, 450 ng of output plasmid and 450 ng of transfection marker plasmid were delivered to each 12-well. Media was replaced with fresh DMEM/F12 + 10% FBS after 24 h. At 2 days post transfection, CHO-K1 cells were dissociated, resuspended in PBS, and transferred to a v-bottom plate for subsequent HCR Flow-FISH.

iPS11

For transfection experiments, iPS11 cells were dissociated using Gentle Cell Dissociation Reagent (STEMCELL Technologies, 100-1077) according to manufacturer's instructions and counted using a hemocytometer. Cells were plated 3 days prior to transfection in mTeSR™ Plus with 10 μM ROCK inhibitor (Millipore Sigma, Y0503-5MG) and 100 U/ml penicillin-streptomycin (Gibco, 15140122) at ∼10% confluency per 12-well. After 24 h, ROCK inhibitor was removed. On the day of transfection, transfection mixes were prepared with FUGENE® HD (FuGENE, HD-1000) using a ratio of 3 μl reagent to 1 μg DNA, and the media was changed to Opti-MEM™ (Thermo Fisher Scientific, 31985062). Four hundred nanograms of output plasmid and 400 ng of transfection marker plasmid were delivered to each 12-well. Fresh mTeSR™ Plus with penicillin-streptomycin was added 4 h after transfection, and the media was changed 24 h after transfection. At 2 days post transfection, cells were dissociated using Gentle Cell Dissociation Reagent, resuspended in PBS, and transferred to a v-bottom plate for subsequent HCR Flow-FISH.

modRNA synthesis and titration curve experiments

The modRNA used in this study was synthesized from the plasmid templates indicated in Supplementary Table S6. The linear template for *in vitro* transcription (IVT) was generated via PCR using Q5 DNA Polymerase (New England Biolabs, M0491) with the primer sequences reported in Supplementary Table S6. The PCR product was isolated on a 1% agarose gel, excised, and purified using the Monarch PCR and DNA Cleanup Kit (New England Biolabs, T1030). Two hundred nanograms of purified product served as template in a 20 μl IVT reaction using the HiScribe T7 High Yield RNA Synthesis Kit (New England Biolabs, E2040), fully substituting UTP with N1-methylpseudouridine-5′-triphosphate (TriLink Biotechnologies, N-1081) and co-transcriptionally capping with CleanCap Reagent AG (TriLink Biotechnologies, N-7114). IVT reactions were incubated at 37°C for 4 h, at which point reactions were diluted to 50 μl, treated with 2 μl DNase I (New England Biolabs, M0303), and incubated at 37°C for 30 min to degrade the IVT PCR template DNA. Synthesized modRNA was column purified and eluted with 60 μl water using the 50 μg Monarch RNA Cleanup Kit (New England Biolabs, T2040). A small sample was nanodropped and ran on a native denaturing gel to determine modRNA concentration and verify full-length product. The modRNA was dispensed in single-use aliquots and stored at −80°C.

For modRNA titration experiments, HEK293T cells were seeded on 12-well plates with 150 000 cells per well 3 days before modRNA transfection. Two days before modRNA transfection, plasmid control conditions were transfected as described above. One day before modRNA transfection, media was replaced with fresh DMEM + 10% FBS. The following day, each modRNA mixture was transfected in triplicate using 1.6 μl Lipofectamine MessengerMAX (Thermo Fisher Scientific, LMRNA) according to manufacturer's instructions. Each well was transfected with varying amounts of mRuby2 or

tagBFP modRNA. To normalize transfection efficiency across conditions, each condition was adjusted to a total modRNA amount of 800 ng with a similar length modRNA that has no predicted affinity with the FISH probes. For experiments in Supplementary Fig. S13B, 400 ng of mRuby2-encoding modRNA was used along with 400 ng of a marker β-globin-tagBFP modRNA. Four hours (Fig. 1E) or 12 h (Fig. 6 and Supplementary Fig. S9 and S13B) after modRNA transfection, cells were dissociated for HCR Flow-FISH by adding 0.25% Trypsin-EDTA (diluted in PBS) for 4 min, followed by quenching with an equal volume of DMEM + 10% FBS.

Site-specific integration

Generation of the HEK293T *Rogi2* Bxb1 LP cell line

We generated a Bxb1 attP LP cell line for facile site-specific integration of genetic cargoes in HEK293T cells. We chose the *Rogi2* locus for integration as it is far from other genes and regulatory elements [32]. To perform the genomic integration, we leveraged ITPN [29]. This CRISPR-based strategy flanks the *Rogi2* homology arms on the donor DNA plasmid with the cognate *Rogi2* protospacer sequence targeted on the genome, 5′-CATCAGACTTGATAGCACTGAGG-3′ (PAM underlined). Subsequent *in situ* nicking of the *Rogi2* locus and the donor DNA plasmid by a high-fidelity Cas9 nickase variant facilitates precise installation of donor DNA at the target locus while minimizing random integration events associated with double-strand breaks generated by wild-type (WT) Cas9.

To implement ITPN, we adopted a staggered delivery approach similar to CRISPR for long-fragment integration via pseudovirus [33], where the donor DNA is delivered first, followed by delivery of CRISPR components 24 h later (Supplementary Fig. S14A). One day before transfection, 150 000 HEK293Ts were plated on a single 12-well. Cells were then transfected with 1000 ng of donor DNA plasmid using PEI. The next day, cells were transfected with 300 ng of nCas9(1.1) modRNA (synthesized in-house with the HiScribe T7 High Yield RNA Synthesis Kit, New England Biolabs, E2040) and 100 ng of *Rogi2* single guide RNA (sgRNA, synthesized with the EnGen sgRNA synthesis kit, New England Biolabs, E3322) using 0.8 µl of Lipofectamine MessengerMax.

Two days after RNA delivery, cells were passaged onto a single six-well. The following day, cells were treated with 10 µg/ml blasticidin (Tocris, 5502) for 4 days to enrich cells that genomically integrated the donor DNA. Single cells from the polyclonal population were sorted into individual wells of a 96-well plate using the Sony MA-900 flow sorter. Two weeks post-sort, confluent monoclonal lines were passaged to 24-well plates, with half of the cells harvested for genotyping PCR. Harvested cells were pelleted and resuspended in 50 µl Cell Lysis Buffer (10×) (Cell Signaling Technology, 9803S) and 0.5 µl of Proteinase K (New England Biolabs, P8107S). Cells were lysed by incubating the suspension for 45 min at 85°C.

As amplifying the *Rogi2* locus proved challenging, we developed an efficient PCR screening method to detect for the desired insertion at *Rogi2*. This consisted of designing the donor DNA with short barcode sequences located between the attP LP and the *Rogi2* homology arms (Supplementary Fig. S14B). These barcode sequences were co-designed with the genotyping primers using PrimerDesign (NCBI) to minimize potential off-target amplicons. To enhance PCR sensitivity, we used nested PCR [34]. For the first, outer PCR, 1 µl of the cell lysate was used as template in a 20 µl PCR using Apex Taq RED Master Mix, 2× (Genesee Scientific, 42-138) with a 30 s extension time and a "touchdown" annealing temperature [35]. This consisted of setting the annealing temperature of the first PCR cycle at 72°C, with each subsequent PCR cycle decreasing the annealing temperature by 1°C until reaching a final annealing temperature of 57°C, followed by an additional 12 cycles at this annealing temperature. One microliter of the first PCR was used as template for a second, inner PCR, following the same PCR conditions. PCR products obtained from the second PCR reaction were resolved on a 2% agarose gel (Supplementary Fig. S14C).

After identifying monoclones that passed the genotyping PCR screen, each candidate was phenotypically screened for the ability to effectively integrate and express genetic cargoes encoded on Bxb1 attB donor plasmids via Bxb1-mediated recombination. Clone #14 emerged as the best candidate from this screen, hereafter referred to as *Rogi2* LP, and was subsequently used in this study.

Bxb1-mediated integration of cargoes in the HEK293T *Rogi2* LP cell line

The procedure outlined here functions analogous to the STRAIGHT-IN iPSC LP platform (Supplementary Fig. S14D) [30, 31]. The *Rogi2* LP line contains a puromycin resistance gene missing a promoter and start codon. Upon Bxb1-mediated recombination between the attB site on the donor plasmid (Supplementary Fig. S14E) and attP site in the LP, an EF1α promoter and start codon is placed in-frame of the resistance gene, conferring recombined cells resistance to puromycin (Supplementary Fig. S14F). A total of 18 different donor plasmids were cloned encoding mRuby2 driven by six different promoters and three different PAS sequences. To integrate the donor plasmids, *Rogi2* LP was plated on 24-well plate at a seeding density of 75 000 cells per well 1 day before transfection. Cells were then co-transfected with 300 ng of donor attB plasmid and 200 ng of CAG-Bxb1 (gift from the Wong Lab at Boston University) using PEI. Once confluent (2–3 days post-transfection), cells were passaged onto a six-well, with puromycin (1 µg/ml, Invivogen, ant-pr-1) administered the following day. Once confluent (5–6 days post puromycin selection), cells were passaged at a split ratio of 1:10 to dilute out residual donor plasmid, at which point cells were ready for use in downstream analyses.

PiggyBac integration

Six different genetic cargoes were randomly integrated into HEK293T cells, encoding expression of mRuby2-2A-PuroR-bGH driven by CAG, EF1α, CMV, UbC, EFS, or hPGK. For PiggyBac transposase-mediated integration, 100,000 HEK293T cells per well were seeded in a 24-well plate coated with 0.1% gelatin. Each well was transfected as described above with 225 ng of donor plasmid and 225 ng of PiggyBac transposase plasmid (gift from the Elowitz Lab). At 1 day post-transfection, media was replaced with fresh DMEM + 10% FBS. At 2 days post-transfection, cells in each 24-well were passaged to a six-well. One day after passaging, media was replaced with fresh DMEM + 10% FBS with 1 µg/ml puromycin for selection of integrated cells. Selection media

was replaced daily for 5 total days of selection. After selection, cells were returned to DMEM + 10% FBS for outgrowth on six-well plates.

After the outgrown cells reached ∼80% confluence, cells were trypsinized (0.25% Trypsin-EDTA diluted in PBS at a 3 PBS : 2 trypsin ratio) and resuspended in DMEM + 10% FBS supplemented with 100 U/ml penicillin-streptomycin.

Separately, "half and half" conditioned media was made by removing media from a confluent flask of HEK293T cells, filtering it through a 0.22 μm filter, and mixing it 1:1 with fresh media. The resulting conditioned media was supplemented with 100 U/ml penicillin-streptomycin.

Cells were sorted on a Sony MA-900 flow sorter using the gates shown in Supplementary Fig. S15. Briefly, live single cells were identified using forward scatter and side scatter gates, and the mRuby2-positive cells were gated using a roughly rectangular mRuby2-FSC gate. Cells were recovered onto gelatin-coated, pre-warmed plates containing the conditioned media. After a media change and outgrowth in media not containing penicillin-streptomycin, cells were confirmed myco-negative (Lonza MycoAlert).

Lentiviral integration

Lentivirus production

Lenti-X HEK293T cells (Takara Bio, 632180) grown in DMEM + 10% FBS were seeded at 10^6 cells per well of a six-well plate. The following day (day 1), 1 μg of the third-generation lentiviral expression plasmid, 1 μg of the packaging plasmid (psPAX2, Addgene #12260), and 2 μg of the envelope plasmid (pMD2.G/VSVG, Addgene #12259) per well were co-transfected using PEI as described above. After 6 h, the media was aspirated and replaced with 1.25 ml of DMEM + 10% FBS + 25 mM HEPES (Sigma–Aldrich, H3375). On the following day (day 2), the media was collected, stored at 4°C, and replaced with HEPES-buffered DMEM + 10% FBS. On day 3, the media was again collected. The collected media was filtered through a 0.45 μm PES filter.

To the filtered virus-containing media, Lenti-X Concentrator (Takara Bio, 631232) was added in a 3 parts media : 1 part concentrator volume ratio, mixed gently, and left overnight at 4°C. On day 4, the media was centrifuged at 1500 rcf at 4°C for 45 min. The supernatant was aspirated, and the resulting pellet was resuspended to a total volume of 200 μl in HEPES-buffered DMEM + 10% FBS. Virus was used immediately or stored at −80°C.

Lentivirus titration

Regularly passaged HEK293T cells were seeded at a concentration of 15 000 cells per well of a 96-well plate in DMEM + 10% FBS on the day of the transduction. Cells were combined with 5 μg/ml polybrene (hexadimethrine bromide, Sigma–Aldrich, H9268-5G) and a two-fold serial dilution of the produced virus (highest concentration: 5.0 μl concentrated virus per well). The resulting cell, polybrene, and virus mixture was plated onto 96-well plates coated with 0.1% gelatin.

Three days later, the resulting cells were dissociated using 0.25% Trypsin-EDTA diluted in PBS at a 3 PBS : 2 trypsin ratio, and data was collected using an Attune NxT flow cytometer. Single, live cells were selected using forward scatter and side scatter gates. Transduced cells were identified using an mRuby2 gate that excluded untransduced cells.

HEK293T transduction

Regularly passaged HEK293T cells were seeded on the day of viral transduction in suspension at 150 000 cells per 12-well. Each 12-well was transduced with concentrated lentivirus produced from six-well plate and titered to have a multiplicity of infection (MOI) of 2. Fresh DMEM + 10% FBS was included to reach a final volume of 2 ml per 12-well, and 5 μg/ml polybrene was added to increase transduction efficiency. Three days later, the resulting cells were dissociated and labeled for HCR Flow-FISH.

HCR RNA-FISH

In all HCR RNA-FISH experiments here, we use Molecular Instruments probe sets for mRuby and tagBFP compatible with B2 amplifiers conjugated to Alexa Fluor™ 514, Alexa Fluor™ 647 (imaging only), or Alexa Fluor™ 488 (Supplementary Figs S4 and S7 only). The FISH protocol as well as hybridization and wash buffer compositions were based on those reported by Choi *et al.* and modified to improve cell recovery for flow cytometry [25]. Amplification buffer composition was based on that reported by Jia *et al* [36]. Compositions of the hybridization buffer, wash buffer, 5X SSCT buffer, and amplification buffer are reported in Supplementary Table S7.

FISH imaging

Cells grown on Geltrex-coated glass-bottom plates were fixed by incubating with 4% paraformaldehyde (PFA) solution (EMD Millipore, 818715) for 1 h at 4°C. After washing the cells three times with cold PBS, the cells were permeabilized using 0.5% Tween-20 (Sigma–Aldrich, P2287) overnight at 4°C. Next, cells were washed twice with 2× saline-sodium citrate (SSC) and then incubated with hybridization buffer for 30 min at 37°C. Probe set stock solution was diluted to 16 nM in hybridization buffer. Cells were incubated in this probe solution overnight at 37°C.

Following hybridization, cells were incubated with wash buffer for 5 min at 37°C, and this was repeated for a total of four washes. Then, cells were incubated with 5× SSCT for 5 min at room temperature, and this was repeated for a total of two washes. After these washes, cells were incubated in amplification buffer for 30 min at room temperature. Amplifier solution was prepared by combining separately snap-cooled hairpins h1 and h2 at a concentration of 60 nM in amplification buffer. Cells were incubated with this amplifier solution for 45 min at room temperature.

Following amplification, cells were incubated with 5× SSCT for 5 min at room temperature, and this was repeated for a total of five washes. Finally, wells were filled with PBS for imaging using a Nikon Ti2-E fluorescence microscope.

HCR Flow-FISH

After suspension in PBS, cells were transferred to 96-well v-bottom plate for HCR Flow-FISH. After each resuspension, spins were performed at 500 rcf for 5 min with default settings, unless otherwise noted. Cells were first fixed through incubation in 4% PFA for 15 min at room temperature. After spinning, cells were then permeabilized using 0.5% Tween-20 for 15 min at room temperature. Next, cells were spun and resuspended in hybridization buffer for 30 min at 37°C. During this incubation, probe set stock solution was diluted in hybridization buffer to a concentration of 14 nM for transfected cells or 28 nM for integrated cell lines. Cells were spun and

resuspended in this probe solution for incubation overnight at 37°C. Due to the viscosity of the hybridization buffer, these spins were performed with reduced deceleration speed to minimize cell loss.

Following hybridization, cells were spun and resuspended in wash buffer for 15 min at 37°C. Then, cells were spun and resuspended in 5× SSCT for 5 min at room temperature. After these washes, cells were spun and resuspended in amplification buffer for 30 min at room temperature. Amplifier solution was prepared by combining separately snap-cooled hairpins h1 and h2 at a concentration of 130 nM in amplification buffer. Cells were spun and then incubated with this amplifier solution overnight at room temperature.

Following amplification, cells were spun and resuspended in 5× SSCT for one 30 min incubation and one 5 min incubation at room temperature. Finally, cells were spun and resuspended in PBS for flow cytometry.

Flow cytometry

All flow cytometry data were collected using an Attune NxT flow cytometer with channel mappings and voltages reported in Supplementary Table S8. Data from HCR Flow-FISH experiments were compensated using the matrix reported in Supplementary Table S9. To account for differences in background across experiments, normalized fluorescence is used where indicated and is calculated as the fold change of fluorescence intensity relative to a WT or non-transfected sample labeled at the same time. Single cells were selected using forward scatter and side scatter gates. Transfected cells were gated based on expression of a co-transfected marker. For lentiviral transduction, the top 85% of cells were gated based on mRuby2 expression, assuming a poisson distribution of integration events corresponding to an MOI of 2.

RT-qPCR

For concurrent RT-qPCR and HCR Flow-FISH analysis, PiggyBac-integrated cell lines were plated in biological triplicate at a density of 300 000 cells per well in six-well plates. After 3 days of growth, cells were dissociated by adding 0.25% Trypsin-EDTA (diluted in PBS) for 4 min, followed by quenching with an equal volume of DMEM + 10% FBS. Each sample of dissociated cells was split in half, with equal amounts going to HCR Flow-FISH processing or RNA isolation.

RNA was isolated using the Monarch Total RNA Miniprep Kit (New England Biolabs, T2010) with an additional Dnase I (New England Biolabs, M0570) treatment step. RNA samples were eluted into 50 µl of nuclease-free water. Complementary DNA (cDNA) was synthesized from 6 µl of eluted RNA using the ProtoScript First Strand cDNA Synthesis Kit (New England Biolabs, E6300) with oligo-dT primers. cDNA samples were stored at −20°C until qPCR.

qPCR was performed at the MIT BioMicro Center on a Roche LightCycler 480 with four technical replicates per condition. Reaction mixes were assembled using 2.5 µl KAPA SYBR FAST qPCR Master Mix (2×) Universal (Kapa Biosystems, KK4600), 0.5 µl 2 µM forward and reverse primers, 0.5 µl cDNA product, and 1.5 µl nuclease-free water. The primer sequences used for each gene are reported in Supplementary Table S10. Using the "High Sensitivity" analysis mode, C_t values were called. Pooling over technical replicates by taking the mean, ΔC_t values were calculated relative to the GADPH levels for each sample and used to calculate relative expression levels.

Total transcription quantification

Total RNA yield

PiggyBac-integrated cell line and WT HEK293T cells were cultured for 3 days and counted via hemocytometer to isolate samples of 500K cells each. RNA was isolated using the Monarch Total RNA Miniprep Kit with an additional Dnase I treatment step. RNA samples were eluted into 50 µL of nuclease-free water, and concentrations were measured via Nanodrop.

5-ethynyluridine labeling

Cells were counted using a hemocytometer, seeded with 0.1% gelatin coating (Sigma–Aldrich, G1890-100G) at a density of 150 000 cells per 12-well, and grown in culture for 3 days. Thirty minutes prior to dissociation, cells were fed media containing 1 mM 5-ethynyluridine (EU) (Sigma–Aldrich, 909475). Cells were first fixed through incubation in 4% PFA for 15 min at room temperature. After spinning, cells were then permeabilized using 0.5% Tween-20 for 15 min at room temperature. Next, cells were incubated for 30 min on a rotator at room temperature in a label mixture containing 0.05 mM copper sulfate (Fisher Scientific, AC197730010), 0.25 mM THPTA (Vector Laboratories, CCT-1010-100), 8 µM Pacific Blue azide (Click Chemistry Tools, 1413-1), and 20 mg/ml ascorbic acid (Sigma–Aldrich, A4544) in PBS. Cells were washed twice with 0.1% Tween-20 in PBS and analyzed via flow cytometry.

Long-read sequencing

RNA from polyclonal PiggyBac-integrated cell lines and WT HEK293T cell lines was isolated using the Monarch Total RNA Miniprep Kit with an additional Dnase I treatment step. RNA samples were eluted into 50 µl of nuclease-free water.

Direct RNA sequencing was performed on an Oxford Nanopore GridION device using the Direct Sequencing Kit (SQK-RNA004, date accessed 15 May 2024), MinION RNA flow cell (FLO-MIN00RA), and data pre-processing was performed with MinKNOW (v24.06.10). Libraries were constructed individually with the following modifications to optimize fragment yield and quantity: (i) ∼1.2 µg of total RNA was used in 8 µl total volumes; and (ii) all binding and elution steps were doubled, with a minimum bead binding time of 5 min. Basecalling was performed on-device using the "super-accurate basecalling" model in Dorado version 7.4.12. The resulting .fastq files were aligned using minimap2 version 2.26 (flags: -ax splice -uf -k14) to custom human reference genomes combining GRCh38 v108 with the plasmid sequence for each construct.

Unique reads mapping to the construct sequences were isolated using bedtools. Major isoform start sites were manually identified from the reads looking at a density distribution of read starts. For any analyses quantifying the start or end positions of the reads, reads were filtered to remove any reads whose 5′ end was >25 nt away from major isoform start sites to avoid any artifacts introduced by 5′ truncation prevalent in long-read RNA sequencing data. Major transcription start sites (TSSs) and intron locations for each promoter sequence are reported in Supplementary Table S1.

Differential expression of endogenous genes was analyzed by comparing data from each cell line to data from WT HEK293T cells. For each gene, the fold change in the read counts per million relative to the WT baseline was calculated, and any fold change with an absolute value >1.5 was considered different from the baseline. Differentially expressed genes common to all six cell lines are listed in Supplementary Table S2. Gene Ontology analysis (GO Ontology database [37], released 6 February 2025) was carried out on this set of genes using a PANTHER overexpression test (released 7 August 2024, Binomial test with Bonferroni correction).

Statistical analysis

Unless noted, all data points used in statistical analysis represent the mean of three biological replicates. Statistical significance was determined using t-tests. Comparisons with directional hypotheses based on previously observed differences in protein expression (e.g. promoter strength, effective translation rate) were performed using one-tailed tests. All other comparisons were performed using two-tailed tests.

Results

HCR Flow-FISH enables simultaneous, high-throughput quantification of mRNA and protein levels in single cells

To simultaneously assess mRNA and protein distributions, we used a Flow-FISH method, which allows for the concurrent measurement of mRNA and protein levels in single cells [38]. Hybridization chain reaction RNA-FISH (HCR Flow-FISH) amplifies signal, improving the signal-to-noise ratio and enabling better mRNA detection at low expression levels [25, 26]. HCR Flow-FISH uses a two-stage amplification approach (Fig. 1B). First, RNA-binding probes complementary to the mRNA transcript of interest are hybridized overnight in fixed and permeabilized cells. The following day, probe-specific, fluorescently labeled hairpins are added to amplify the FISH signal. HCR leads to higher fluorescence levels while minimizing background, making it well-suited for flow cytometry quantification.

We optimized and validated an HCR Flow-FISH protocol based on methods reported in Choi *et al.* [26]. We transfected HEK293T cells with a plasmid expressing the fluorescent protein mRuby2 with the EF1α promoter and bGH PAS (EF1α-mRuby2-bGH). We quantified expression of the mRuby2 mRNA using sequence-specific FISH probes and compatible HCR amplifiers. Given the limited size or absence of introns in the transgenic transcripts, we do not distinguish between nascent and mature transcripts in this work. Through fluorescence imaging (Fig. 1C, Alexa Fluor™ 647) and flow cytometry (Fig. 1D, Alexa Fluor™ 514), we measured both mRNA and protein expression in the same single cells. Binning cells based on expression of a co-transfected fluorescent marker, we observe a linear dependence between mRNA signal and protein signal. We quantified a dimensionless effective translation rate of the transcripts, $\alpha_{p,\,eff}$ by calculating the slope between mRNA signal and protein signal via least squares regression (Fig. 1A). This parameter is proportional to the number of proteins translated from a single mRNA transcript and combines the contributions of mRNA transport, translation initiation, and protein stability. We selected the plasmid dosage to maximize expression while minimizing the cell death possibly caused by the transfection reagent (Supplementary Fig. S1).

To verify that HCR Flow-FISH detects quantitative changes in mRNA level, we transfected varying dosages of mRuby2-encoding modRNA into HEK293T cells. To minimize any differences in transfection efficiency, we added a non-fluorescent, "filler" modRNA to maintain a consistent total modRNA amount across all conditions. As expected, the mean HCR Flow-FISH signal increases linearly with modRNA dosage, indicating that this method can detect the anticipated mean differences in modRNA levels between populations of cells (Fig. 1E, and Supplementary Fig. S9A and B). The protein expression at 12 h post-modRNA transfection is also linearly related to the modRNA levels, indicating that effective translation rates are not dependent on modRNA dosage, as expected (Supplementary Fig. S9C and D). Importantly, when transfecting a plasmid with a strong constitutive promoter, CAG (dashed line), the measured HCR Flow-FISH signal sits within the linear detection regime and does not reach a nonlinear saturating regime. Compared to genomic integration, transfection typically leads to higher transgene DNA copy numbers, so this condition represents the maximum relevant amount of mRNA for detection. Therefore, because HCR Flow-FISH signal correlates linearly with modRNA dosage across the range relevant for transgene expression, we can quantitatively compare mRNA levels between different compositions of genetic elements. Additionally, we find that the mean signal from HCR Flow-FISH correlates positively with the signal from RT-qPCR across five cell lines with varying mRuby2 expression (Fig. 1F). Thus, we conclude that HCR Flow-FISH and RT-qPCR provide similar relative estimates for mean mRNA levels, while HCR Flow-FISH offers the additional benefit of quantifying the distribution of mRNA levels. Together, HCR Flow-FISH enables single-cell quantification of mRNA levels for simultaneous characterization of mRNA and protein expression distributions.

Promoter sequences affect RNA transcript abundance and effective translation rate

With the quantitative HCR Flow-FISH protocol validated, we characterized a commonly used set of genetic elements including promoters. In the field of synthetic biology, genetic components that drive transcription are determined heuristically from native genomes and often include the combination of a minimal promoter, a TSS, upstream regulatory elements, and in some cases an associated 5′ UTR sequence (Fig. 2A). Since the precise boundaries of each of these elements within a region are not always clear, we use a functional definition of the promoter referring to this entire set of sequences until they can be precisely defined. Choice of promoter can set the level of protein expression. However, as promoters differ in recruitment of transcriptional machinery, 5′ UTR sequences, and splicing within 5′ UTRs, protein data alone cannot define how promoters influence transcription rates. We chose to evaluate a set of six constitutive promoters with varying expression levels, composition, and origins: CAG, a hybrid of the cytomegalovirus enhancer and chicken beta-actin promoter; EF1α, the human elongation factor 1-alpha promoter; CMV, a strong promoter derived from cytomegalovirus; UbC, the human polyubiquitin C promoter; hPGK, the human phosphoglycerate kinase promoter; and EFS, a derivative of the EF1α promoter lacking the intron (Supplementary Table S1).

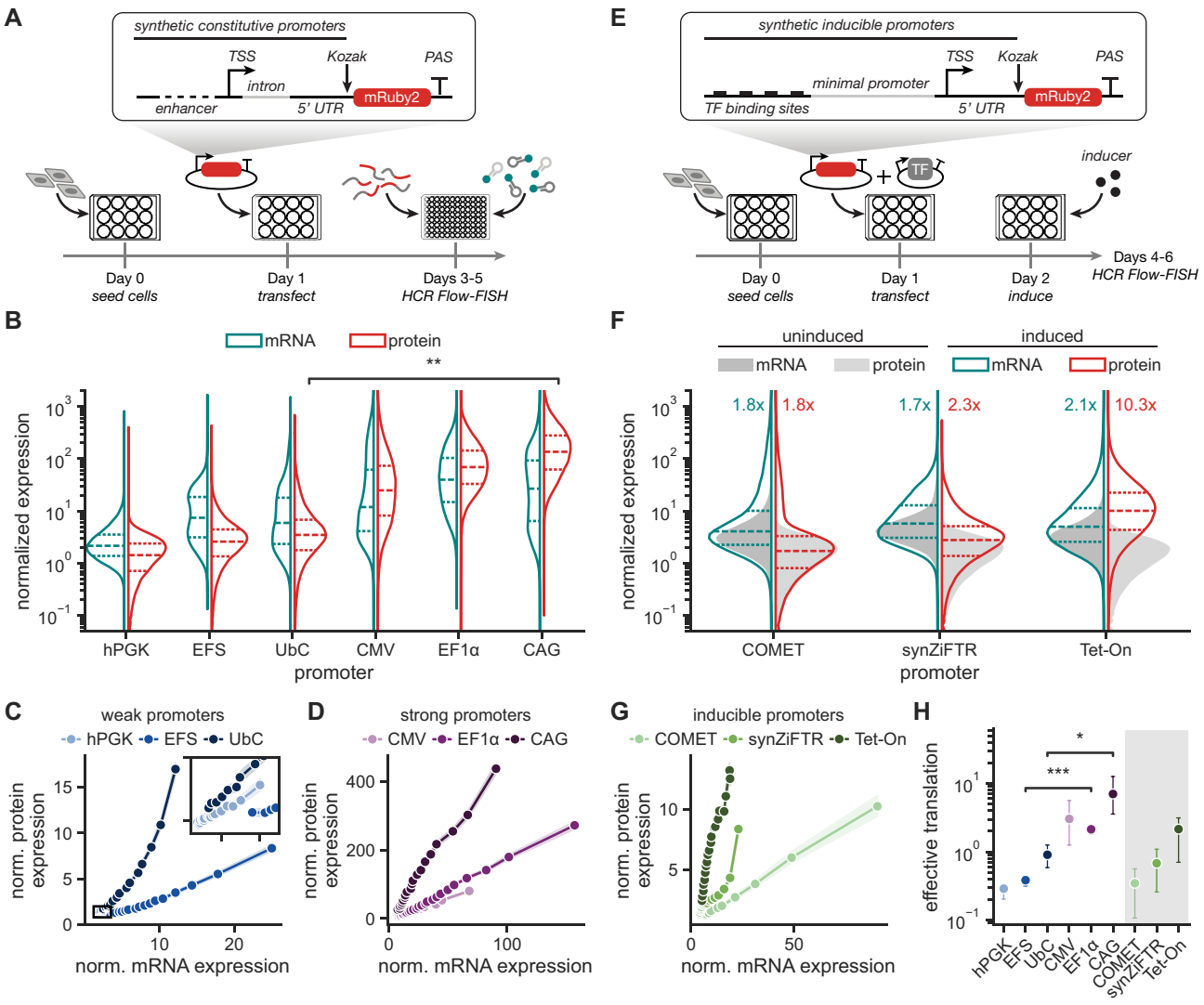


Figure 2. Promoter sequences affect RNA transcript abundance and effective translation rate. (**A**) HEK293T cells were transfected with a plasmid encoding mRuby2 driven by one of six different constitutive promoters (CAG, EF1α, CMV, UbC, EFS, or hPGK) followed by a bGH PAS. (**B**) Normalized expression distributions for mRNA (Alexa Fluor™ 514) and protein (mRuby2) with six constitutive promoters as measured by flow cytometry. .01 ≥ **P > .001, one-sided *t*-test. (**C**) Normalized protein expression as a function of normalized mRNA expression with three weak constitutive promoters. Inset axes show the indicated low expression domain. Data for one representative biological replicate are binned by marker level into 20 equal-quantile groups. Points represent geometric means of mRNA and protein levels for each bin. Shaded regions represent the 95% confidence interval. (**D**) Normalized protein expression as a function of normalized mRNA expression for three strong constitutive promoters for one representative biological replicate. (**E**) HEK293T cells were transfected with plasmids encoding mRuby2 driven by one of three different inducible promoters (Tet-On, COMET, synZiFTR) and the corresponding transcription factor. (**F**) Normalized expression distributions for mRNA (Alexa Fluor™ 514) and protein (mRuby2) with three inducible promoters following small molecule induction as measured by flow cytometry. Fold-change in expression upon induction is annotated for each condition. (**G**) Normalized protein expression as a function of normalized mRNA expression with three inducible promoters for one representative biological replicate. (**H**) Effective translation rate as calculated by the slope of a line fitted to the binned data. Inducible promoters are denoted by the gray box. Points represent the mean of three biological replicates ± the 95% confidence interval. .05 ≥ *P > .01, one-sided *t*-test. Normalized expression is calculated as the fold change of fluorescence intensity relative to a nontransfected sample. All data are in arbitrary units from a flow cytometer. TSS: transcription start site; and PAS: polyadenylation signal.

To assess the relative expression levels of this promoter set, we used each promoter to drive expression of mRuby2 with a bGH PAS (Fig. 2). For each condition, we included a co-transfected marker plasmid encoding a separate fluorescent protein, whose expression we used as a proxy for copy number [39]. We observed consistent marker expression and transfection efficiency regardless of the identity of the co-transfected promoter (Supplementary Fig. S3). Constitutive promoter conditions were analyzed via HCR Flow-FISH at two days post-transfection (Fig. 2A). As expected from previous characterizations [16], our data ranks promoters by strength—as measured by mean protein expression—from highest to lowest: CAG, EF1α, CMV, UbC, EFS, and hPGK (Supplementary Fig. S2).

In transfection, mRNA expression exhibits more variance across the population than protein expression (Fig. 2B and Supplementary Fig. S2). Bimodality observed in levels of mRNA and not in protein may reflect the higher stability of the proteins compared to mRNA. Despite the increased variance, the relative ordering of promoter strength, as deter-

mined by the mean mRNA expression level, generally matches that of the mean protein expression level (Supplementary Fig. S2). However, CAG achieves the highest protein expression but has only the second-highest level of mRNA, indicating post-transcriptional processing and translation rates influence protein levels. Binning the cells by expression of the transfection marker, we find that higher mRNA levels are strongly correlated with higher protein levels (Fig. 2C and D). However, the relative slopes of these curves differ, indicating differences in effective translation rate across promoters. In addition to having higher mRNA levels, the strong promoters (CAG, EF1α, and CMV) exhibit more efficient translation than the weak promoters (UbC, EFS, and hPGK) (Fig. 2H). These differences are likely sequence-specific and could be attributed to factors such as RNA nuclear export, localization, and secondary structure.

To assess generality of trends, we characterized profiles of expression across different cell types and integration methods. First, we quantified profiles of expression for transfection in Chinese hamster ovary (CHO-K1) and iPS11 cells. Both cell types maintain the same relative promoter strengths, with the exception of CMV, which did not express above background in iPS11 cells (Supplementary Fig. S4). Similar to HEK293T cells, we observed higher effective translation rates for stronger promoters in CHO-K1 cells and iPS11 cells; however, the relative differences vary across cell types (Supplementary Fig. S4). Immunogenicity associated with viral sequences in CAG and CMV may suppress gene expression in iPSCs [40]. We next explored how integration into the genome affects profiles of expression. Understanding how transfection profiles translate to integration can accelerate the design-build-test-learn loop, which remains slower due to the time scales of generating cell lines. We quantified expression from each promoter in HEK293T cells after random integration via PiggyBac transposase and lentivirus as well as site-specific integration at a LP at *Rogi2* [32] (see "Materials and methods" section). We find that across integration methods, relative promoter strengths are comparable to those in transfection (Supplementary Fig. S5A–C). However, we observe a smaller range in mRNA expression between weak and strong promoters for PiggyBac-mediated and lentiviral integration (Supplementary Fig. S5D). We hypothesize that this is due to differences in copy numbers between the constructs. In the case of PiggyBac-mediated integration, selection pressure from the co-expressed puromycin-resistance gene may bias the selection of cells integrated with weaker promoters. To survive selection, weaker promoters may need to be integrated at higher copy numbers than are needed for stronger promoters, which would diminish the differences between weak and strong promoters. Similarly, lentivirus titering relies on distinguishing infected cells based on fluorescence, and a larger number of integrations for weaker promoters may be required for this fluorescence to be detectable. While, these effects may be mitigated by selecting or titering using a separate adjacent gene, adjacent genes can alter expression through biophysical coupling [41–43]. Overall, we find that relative promoter strengths remain consistent across cellular and integration contexts, while absolute expression levels may vary.

For many applications that require temporal control over expression, inducible promoters offer the advantage of small-molecule regulation of transcription. We characterized three promoters activated by small molecule-inducible transcriptional activators: doxycycline-inducible rtTA (Tet-On, CMV minimal promoter), rapamycin-inducible zinc-finger activator (COMET [27], YB TATA minimal promoter), and grazoprevir-inducible zinc-finger activator (synZiFTR [28], YB TATA minimal promoter). To facilitate comparison, each inducible promoter drives expression of mRuby2 with a bGH PAS. The cognate transcriptional activators for each synthetic promoter are expressed separately via the EFS promoter. We added small-molecule inducers at 1 day post-transfection and measured expression profiles via HCR Flow-FISH 2 days later (Fig. 2E). Under these conditions, all three promoters display comparable levels of leaky expression in the absence of the inducer, similar to levels of expression from the weakest constitutive promoter, hPGK (Fig. 2F). Interestingly, the induced Tet-On promoter has a similar expression level and effective translation rate to the constitutive CMV promoter. Both of these promoters contain the minimal CMV promoter, suggesting that this sequence drives both transcriptional and post-transcriptional kinetics. Upon induction, the Tet-On and synZiFTR promoters show modest increases in mRNA expression; however, the Tet-On promoter leads to a much higher level of protein expression (Fig. 2F). This larger increase in protein level relative to mRNA level for the Tet-On promoter appears as a higher slope (Fig. 2G) and effective translation rate (Fig. 2H) compared to the other inducible promoters. Altogether, our data suggest that these promoter sequences impact protein expression not just through differences in mRNA transcript levels but also through the effective translation rates of those transcripts, potentially via mRNA processing and transport kinetics.

Choice of PAS impacts effective translation rate of mRNA transcripts

Given that 3′ UTR sequences can significantly impact RNA stability, protein translation, and lentivirus production efficiency, we sought to quantify the effects of PAS choice on expression kinetics [44–47]. We paired each constitutive promoter with three commonly used 3′ UTR sequences—bGH, derived from the bovine growth hormone gene; SV40, derived from the SV40 virus; and WPRE, derived from the woodchuck hepatitis virus (Fig. 3A)—and transfected these transgenes into HEK293T cells. The bGH and SV40 PASs generate similar levels of protein (Fig. 3C) and mRNA (Fig. 3B). However, the WPRE sequence, which is not a mammalian PAS but enables the most efficient lentivirus production (Supplementary Fig. S5E), causes a reduction in protein levels for strong promoters (CAG and EF1α, Fig. 3C) despite having similar mRNA expression (Fig. 3B).

Analyzing the data binned by marker expression, we find that the WPRE sequence results in the lowest slope of protein expression with respect to mRNA expression for all promoters, regardless of promoter strength (Fig. 3D). This represents a lower effective translation rate for WPRE transcripts compared to bGH transcripts (Supplementary Fig. S6A). These results generalize to CHO-K1 cells and iPS11, where the WPRE sequence results in lower protein expression and lower effective translation for strong promoters (Supplementary Fig. S7). For strong promoters such as CAG and EF1α, which represent the maximum transcriptional output in these transfection experiments, the decrease in effective translation rate results in a decrease in protein levels. For weak promoters such as EFS and hPGK, the decrease in effective translation rate may be balanced by an increase in mRNA levels, resulting in

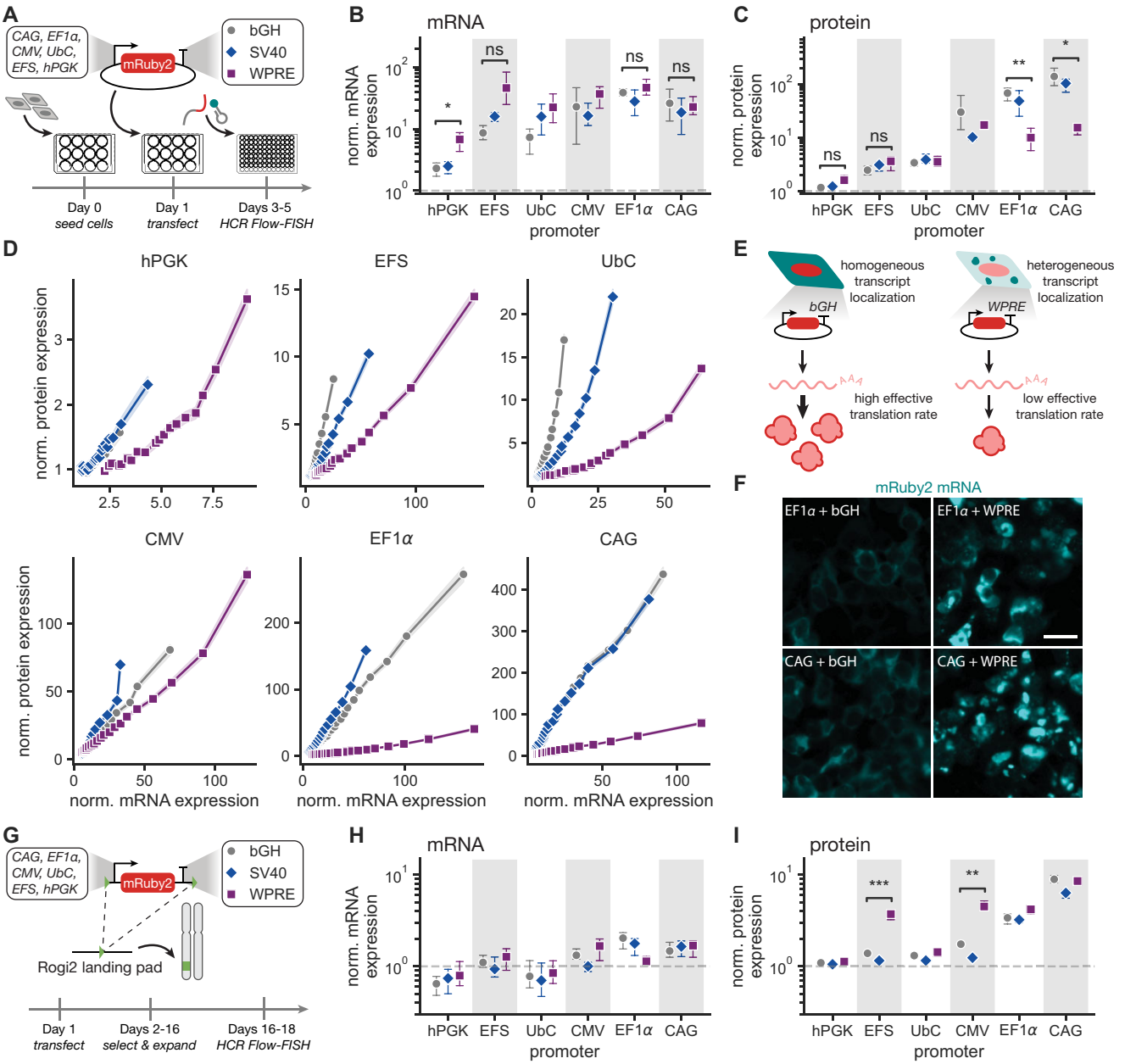


Figure 3. Choice of PAS impacts effective translation rate of mRNA transcripts. (**A**) HEK293T cells were transfected with a plasmid encoding mRuby2 driven by one of six different constitutive promoters along with one of three different 3′ UTR sequences. (**B**, **C**) Normalized geometric mean of mRNA (Fig. 3B, Alexa Fluor™ 514) and protein (Fig. 3C) fluorescence for varying promoter and PAS or 3′ UTR sequence in transfection of HEK293T cells. .05 ≥ *P > .01, .01 ≥ **P > .001, two-sided *t*-test. (**D**) Normalized protein expression as a function of normalized mRNA expression for each promoter and 3′UTR pair. Data for one representative biological replicate are binned by marker level into 20 equal-quantile groups. Points represent geometric means of protein and mRNA expression for cells in each bin, and shaded regions represent 95% confidence intervals. (**E**) Schematic of HCR RNA-FISH imaging of transfected HEK293T cells. Transcripts with a bGH PAS are distributed homogeneously throughout the cytoplasm and are efficiently translated. Transcripts with a WPRE sequence localize in foci and produce less protein. (**F**) HCR FISH imaging for HEK293T cells transfected and labeled with Alexa Fluor™ 647 HCR amplifiers. Scale bar represents 25 μm. All images captured at identical settings. (**G**) The genes shown in Fig. 3A were integrated site-specifically at a *Rogi2* LP in HEK293T cells using Bxb1-mediated recombination. (**H**, **I**) Normalized geometric mean of mRNA (Fig. 3H) and protein (Fig. 3I) fluorescence for varying promoter and PAS integrated at the *Rogi2* locus in HEK293T cells. .01 ≥ **P > .001, .001 ≥ ***P > .0001, two-sided *t*-test. Normalized expression is calculated as the fold change of fluorescence intensity relative to a nontransfected sample. Points represent means of three biological replicates, and error bars represent the 95% confidence interval. All data are in arbitrary units from a flow cytometer.

relatively stable protein levels. With HCR FISH imaging, we observe that transcripts encoding WPRE are heterogeneously distributed across the cytoplasm and form foci, whereas bGH transcripts are distributed uniformly throughout the cytoplasm (Fig. 3E and F). The presence of these foci may indicate aberrant localization of WPRE transcripts, which may limit their accessibility to ribosomes and effective translation rate.

To examine how choice of 3′ UTR sequences affects expression profiles of integrated transgenes, we site-specifically integrated these cassettes into HEK293T cells at the *Rogi2* LP (Fig. 3G). In contrast to the transfection results, we found that, at *Rogi2*, the WPRE sequence significantly increases protein levels for the CMV and EFS promoters despite minimal effects at the mRNA level (Fig. 3H and I). Additionally, we observe no differences in the transcript localization with the WPRE sequence, indicating that the previously observed foci arise at higher levels of expression and potentially are specific to the copy number of the transgene (Supplementary Fig. S8). However, the HCR Flow-FISH signal approaches the limit of detection, even for the strongest promoters. Therefore, we may not be able to resolve differences between mRNA levels at this low copy number using HCR Flow-FISH. Nevertheless, PAS choice significantly impacts the levels of protein expression from this promoter set at low copy number. Together, these findings demonstrate that PAS and 3′ UTR sequences tune mRNA and protein expression with different effects in transfection and integration.

Gene coding sequence impacts effective translation rate but not mRNA levels

The identity of the transgene affects mRNA and protein levels via its specific sequence, which influences transcript stability, translation kinetics, and protein stability. In particular, coding sequences may impact transcription elongation and/or the secondary structure of the mRNA transcript, which affects translation [48]. To investigate how the identity of the target gene affects expression profiles, we exchanged mRuby2 for tagBFP. While the mRuby2 and tagBFP transgenes are of similar length (711 and 702 bp, respectively), tagBFP has a modestly higher GC-content (56% versus 49% for mRuby2), which may result in differential transcription elongation [49], mRNA nuclear export [50], stability [51], and translation efficiency [52]. Using our constitutive promoter panel to express tagBFP with a bGH PAS, we characterized the profiles of mRNA and protein expression in transfection of HEK293T cells (Fig. 4A and B, and Supplementary Fig. S10). While the slopes of the weak promoters show similar effective translation rates for tagBFP and mRuby2 (Figs 2C and 4B), tagBFP transcripts exhibit a different ordering of slopes for the set of strong promoters (Figs 2D and 4C).

To understand the differences between mRuby2 and tagBFP expression, we directly compared their RNA distributions (Fig. 4A and E). Since both transcripts are labeled using the same HCR Flow-FISH amplifiers, we can quantitatively compare mRNA profiles generated by HCR Flow-FISH (Supplementary Fig. S9). Strikingly, each constitutive promoter generates similar mean RNA levels regardless of the transgene expressed (Fig. 4E). Similar to mRuby2, tagBFP RNA levels show more variance than tagBFP protein levels (Supplementary Fig. S10). Thus, we find substitution of tagBFP for mRuby2 does not substantially affect the profiles of mRNA.

While the levels of mRNA are comparable between coding sequences for each promoter, protein levels differ substantially across the stronger promoters (Supplementary Fig. S10B). The strongest promoters (EF1α and CAG) display lower levels of tagBFP protein relative to CMV. Lower expression of tagBFP protein cannot be explained by potential differences in protein half-life because a reduction in protein half-life would result in lower protein expression across all promoters. Rather, these data suggest that the promoters affect the translation of tagBFP transcripts (Fig. 4D). Curiously, unlike mRuby2, we observe that tagBFP exhibits a lower effective translation rate for strong promoters than for weak promoters (Fig. 4F). The consistency in mRNA levels between mRuby2 and tagBFP suggests that these transgenes are transcribed at similar rates despite displaying different trends at the protein level. Rather, we suggest that the sequences of these coding regions may impact RNA processing, transcript localization, RNA stability, or translation that manifest as differences in protein levels. Further characterization of 5′ UTR architectures may allow for the identification of factors impacting translation [53].

Examining transcript isoforms highlights promoter-specific patterns of gene regulation

The UTRs of transcripts can have sequence-specific effects on mRNA transport, translation, and stability [54]. Additionally, introns within the 5′ UTRs associated with synthetic promoters, such as CAG, EF1α, and UbC, may affect transcription and mRNA processing kinetics via "intron-mediated enhancement" [55]. We sought to define these effects by precisely mapping TSSs and transcription end sites (TESs), as well as splice site positions, using long-read direct RNA sequencing (Fig. 5A). We sequenced transcripts from six cell lines randomly integrated with transgenes containing different constitutive promoters, which we previously analyzed with HCR Flow-FISH and RT-qPCR (Figs 1F and 5B, and Supplementary Fig. S5A). We find that transcript isoforms are highly uniform. For the set of constitutive promoters, only a small fraction of transcripts deviate from expected transcript start and end positions (Fig. 5C and Supplementary Fig. S11A).

Despite the possibility of substantial differences in local sequence, gene, or chromatin context due to random integration into the genome, we observe very little evidence for transcriptional read-through from an upstream promoter or for transcriptional differences caused by local effects. Additionally, transcript counts from sequencing agree with previous measures of mRuby2 mRNA levels via HCR Flow-FISH (Fig. 5D). Remarkably, even for synthetic promoters encoding introns, very few reads exhibit unexpected splicing patterns, where aberrant splicing results in the use of a cryptic splice site in the mRuby2 coding sequence (Fig. 5E). An even smaller fraction of reads display readthrough past the expected TES. However since direct RNA sequencing only captures polyadenylated molecules, we lack the ability to estimate any readthrough transcripts that are not polyadenylated(Fig. 5F). Moreover, we cannot rule out readthrough from adjacent endogenous genes without knowing the genomic sequences of regions flanking the integration sites.

The distribution of 5′ UTR lengths differs across the set of constitutive promoters (Fig. 5G). We defined the 5′ UTR as the distance between the observed TSS and the mRuby2 start codon. Specifically, CAG, UbC, and hPGK have the longest

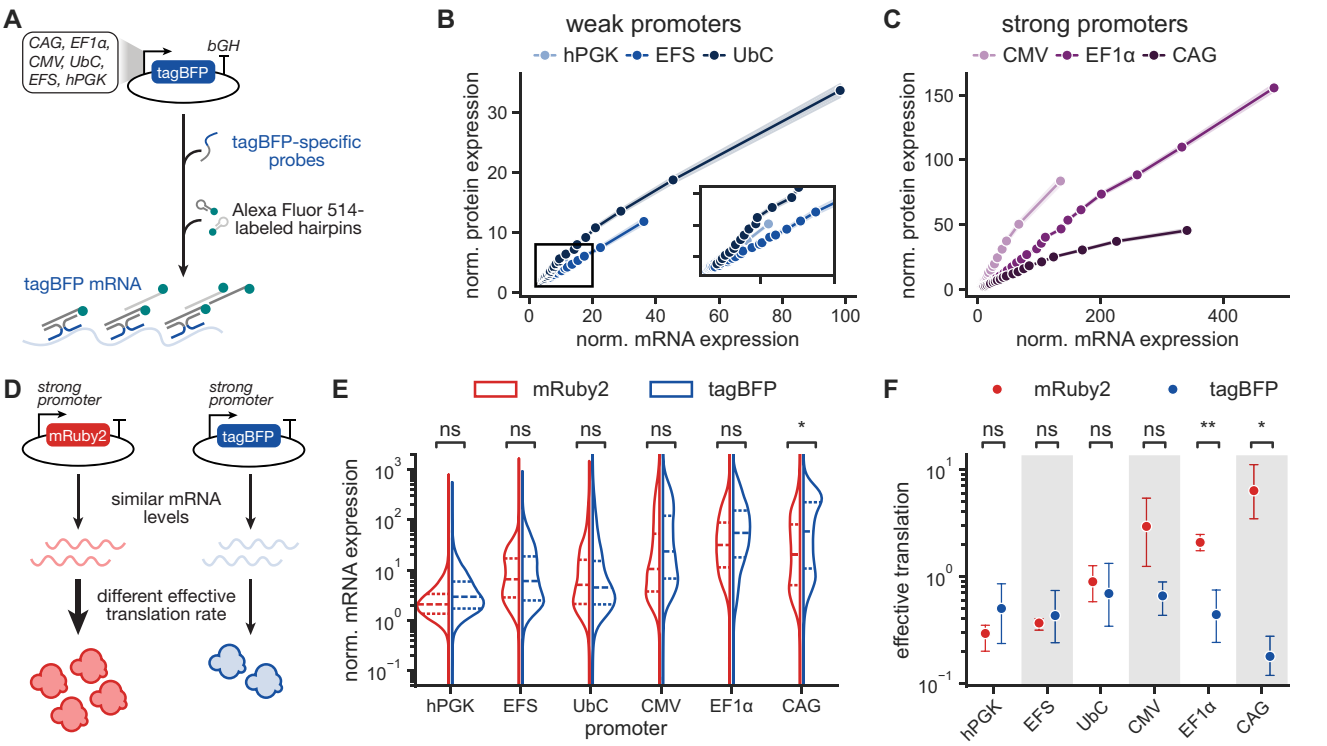


Figure 4. Gene coding sequence impacts effective translation rate but not mRNA levels. (**A**) HEK293T cells were transfected with a plasmid encoding tagBFP driven by one of six different constitutive promoters (CAG, EF1α, CMV, UbC, EFS, or hPGK). (**B**) Normalized protein expression as a function of normalized mRNA expression with three weak constitutive promoters. Inset axes show the indicated low expression domain. Data for one representative biological replicate are binned by marker level into 20 equal-quantile groups. Points represent geometric means of mRNA and protein levels for each bins. Shaded regions represent the 95% confidence interval. (**C**) Normalized protein expression as a function of normalized mRNA expression for three strong constitutive promoters for one representative biological replicate. (**D**) For genes with similar mRNA levels, differences in protein levels can indicate sequence-specific effects on the effective translation rate of RNA transcripts. (**E**) Normalized expression distributions for mRuby2 and tagBFP mRNA (Alexa Fluor™ 514) with six constitutive promoters as measured by flow cytometry. .05 $\geq$ *P, two-sided *t*-test. (**F**) Effective translation rate as calculated by the slope of a line fitted to the binned data. .05 $\geq$ *P, .01 $\geq$ **P > .001, one-sided *t*-test. Normalized expression is calculated as the fold change of fluorescence intensity relative to a nontransfected sample. Points represent means of three biological replicates, and error bars represent the 95% confidence interval. All data are in arbitrary units from a flow cytometer.

5′ UTRs at ∼50–100 nucleotides, while the 5′ UTRs of EF1α, CMV, and EFS are ∼20-30 nucleotides. Of the promoters tested, UbC and hPGK had the lowest number of mRuby2 reads (Fig. 5D). Potentially, the longer 5′ UTRs associated with these promoters may reduce transcript stability as has been observed natively [56].

Finally, we evaluated the relative impact of each promoter on endogenous gene expression by comparing to gene expression in the WT HEK293T cell line (Fig. 5H). Overall, expression of endogenous genes is highly correlated across cell lines. A small fraction of the annotated genes are differentially expressed, indicating that transgenes exert minimal impact on native genes (Fig. 5I, J and Supplementary Fig. S11B). We identified 89 differentially expressed transcripts common to all of the integrated cell lines (Supplementary Table S2), and many of these genes are associated with translation and protein stability (Supplementary Fig. S11C). It is possible that these differences are induced solely by the presence of the transgene or arise from the selection and sorting performed during cell line production. In addition to the minimal changes in relative gene expression levels, we observed minimal differences in total transcription as measured by 5-ethynyluridine (EU) labeling and total RNA yield (Supplementary Fig. S12). Overall, these results indicate that these promoters exert minimal burden across a range of expression levels in HEK293T cells.

The impact of canonical promoter sequence dominates the effective translation rate across a set of 5′ UTR sequences

Prior to analyzing transcript isoforms via long-read sequencing, the TSSs for the selected, functionally defined promoters were unknown, which prevented us from combinatorially interrogating the effects of the canonical promoters and their associated 5′ UTRs. Additionally, we did not know if transgenic mRNAs were diverse in isoforms or could be represented by a single dominant isoform. With the data from long-read isoform mapping of transcripts, we defined the TSS and dominant transcript isoforms for each synthetic promoter. With this sequence information, we could examine the processes impacting the effective translation rate with greater resolution (Fig. 6A). Specifically, we considered (1) the relative impacts of canonical promoter and 5′ UTR sequences on mRNA processing and transport and (2) the impact of the selected 5′ UTRs on translation. To study the relative impacts of the canonical promoter and 5′ UTR sequences on mRNA processing and transport, we separated the sequences for hPGK, EFS, and CMV into a canonical promoter upstream of the TSS and a 5′ UTR downstream of the TSS. We then cloned versions of the plasmids expressing identical mRNA transcripts (same 5′ UTRs) from different promoters (Fig. 6B). We excluded the promoters with introns to eliminate the potential impact of splicing.

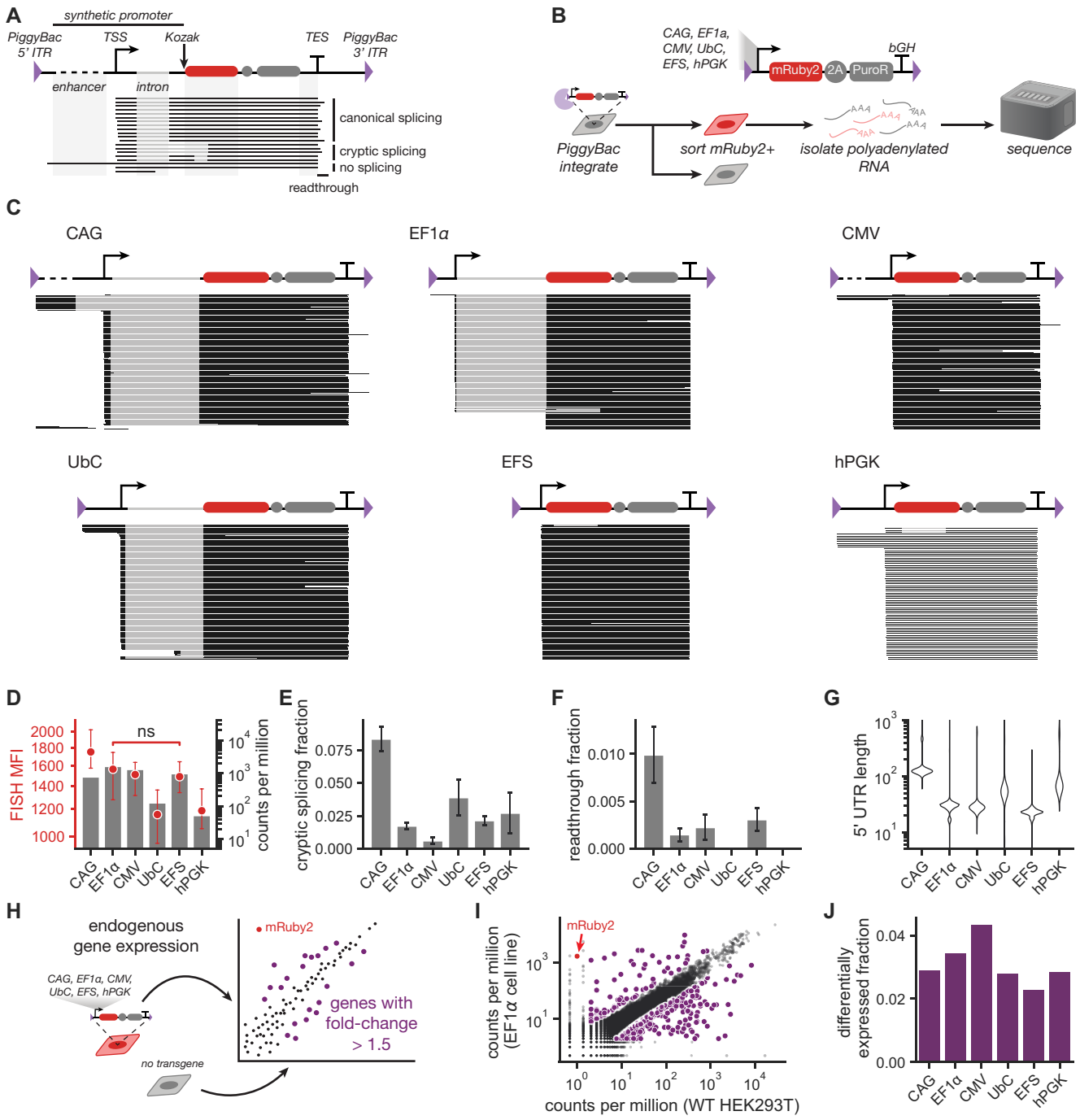


Figure 5. Examining transcript isoforms highlights promoter-specific patterns of gene regulation. (**A**) Example transcripts depicting a synthetic promoter sequence containing an intron and enhancer. Each line represents a separate sequencing read. The TSS and TES are the locations where the majority of reads begin and stop, respectively. Reads are separated by splicing status, where canonical splicing refers to reads with the annotated intron in the promoter region and cryptic splicing refers to reads using a cryptic intron in the coding region. Readthrough transcripts extend past the annotated TES. (**B**) Genetic cargoes were integrated randomly into HEK293T cells using PiggyBac transposase. Cells were sorted to yield polyclonal, mRuby2+ cell lines. RNA was isolated from each cell line and subjected to long-read sequencing. (**C**) RNA transcript maps for the six constitutive promoters tested. All samples were downsampled to display 144 reads for consistency, aside from hPGK, for which there are only 77 total reads. All reads are shown in Supplementary Fig. S11A. (**D**) HCR Flow-FISH MFI (left axis, points) and mRuby2 sequencing counts per million (right axis, bars) for the six cell lines. MFI is displayed in arbitrary units as the mean of three biological replicates ± the 95% confidence interval. ns: $P > .05$, two-sided *t*-test. Fraction of reads exhibiting cryptic splicing (**E**) or readthrough (**F**) error bars represent the standard deviation assuming counts follow a binomial distribution. (**G**) Distribution of 5′ UTR lengths across reads. (**H**) Endogenous genes are considered "differentially expressed" between the cell lines if the absolute value of the fold change is >1.5. These genes are indicated in purple. mRuby2 expression is indicated in red. (**I**) Endogenous gene expression levels for the cell line integrated with the EF1α (y-axis) promoter and the WT HEK293T cell line (x-axis). Data for the rest of the promoters are shown in Supplementary Fig. S11B. Lists of common differentially expressed genes are reported in Supplementary Table S2. (**J**) Fraction of annotated endogenous genes considered differentially expressed in each cell line.

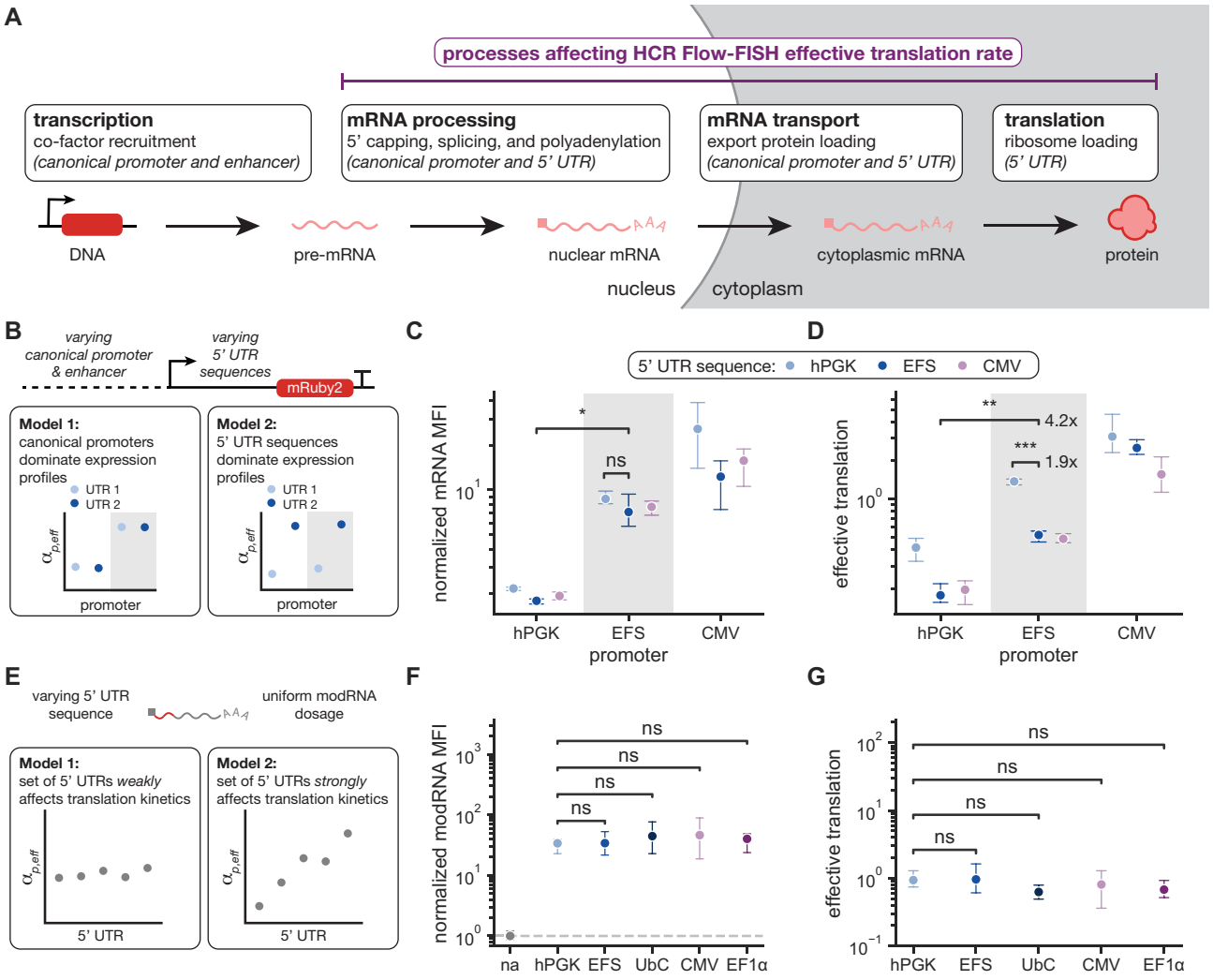


Figure 6. The impact of canonical promoter sequence dominates the effective translation rate across a set of 5′ UTR sequences. (**A**) The effective translation rate as determined by HCR Flow-FISH encompasses mRNA processing and transport as well as translation itself. This parameter can be impacted both by the canonical promoter sequence and the promoter's associated 5′ UTR. Using TSS locations determined by long-read sequencing, these genetic parts can be varied independently to investigate their relative impacts on effective translation rate. (**B**) HEK293T cells were transfected with plasmids expressing transcripts from varying canonical promoters with different 5′ UTRs. If canonical promoter sequences impact effective translation of identical transcripts, then transcriptional regulation can impact translation beyond just the sequence of the 5′ UTR. (**C**) Normalized geometric mean of mRNA fluorescence for mRuby2 plasmid transfection with varying 5′ UTR and promoter sequence. (**D**) Effective translation rate as calculated by the slope of a line fitted to data binned by fluorescence of a co-transfected marker plasmid. (**E**) HEK293T cells were transfected with equal amounts of modRNA encoding the 5′ UTRs of various constitutive promoters to investigate the impact of 5′ UTR sequence on translation. (**F**) Normalized geometric mean of modRNA fluorescence for mRuby2 modRNA transfection with varying 5′ UTR sequence. (**G**) Effective translation rate as calculated by the slope of a line fitted to data binned by fluorescence of a co-transfected marker modRNA. Normalized expression is calculated as the fold change of fluorescence intensity relative to a non-transfected sample. Points represent means of three biological replicates, and error bars represent the 95% confidence interval. All data are in arbitrary units from a flow cytometer. *.05 ≥ *P > .01, .01 ≥ **P > .001, .001 ≥ ***P > .0001, two-sided t-test with Bonferroni correction.

Since none of these 5′ UTRs contain introns, differences in effective translation rate between constructs may indicate differences in nuclear export and mRNA processing such as 5′ capping. We found that the choice of canonical promoter sequence has a larger impact on mRNA expression (Fig. 6C), protein expression (Supplementary Fig. S13A), and effective translation rate (Fig. 6D) than changing the 5′ UTR sequence. Nevertheless, the hPGK 5′ UTR does lead to higher effective translation rates than the EFS and CMV 5′ UTRs, possibly due to more efficient nuclear export mediated by higher GC content (68% for hPGK versus 60% and 62% for EFS and CMV, respectively) [57]. Thus, synthetic promoter choice impacts mRNA kinetics downstream of transcription. Promoters retain a dominant impact on effective translation rate even when combined with 5′ UTR sequences beyond those canonically associated with each promoter. Our results suggest that effective translation rate can be tuned by selection of the promoter sequence across a range of 5′ UTRs.

To probe the effects of the 5′ UTR sequences on translation only, we synthesized modRNA encoding the mature, spliced transcript sequences corresponding to the hPGK, EFS, UbC, CMV, and EF1α promoters (Fig. 6E). If these 5′ UTRs only weakly impact loading and translation, we would observe small variation in protein expression (Model 1). Alternatively, if the 5′ UTR significantly influences ribosome loading and translation kinetics, we would expect to observe a large

range of protein expression across 5′ UTR variants (Model 2). When co-transfected with a marker modRNA, we found that these 5′ UTR sequences did not significantly impact RNA levels (Fig. 6F), protein expression (Supplementary Fig. S13B), or effective translation rate (Fig. 6G). Therefore, we conclude that this set of 5′ UTR sequences weakly influences the rate of ribosome loading and translation kinetics of the transgene. Instead, the differences in effective translation rate observed with plasmid transfection reflect differences in mRNA processing and transport.

Discussion

Properly tuning transgene expression levels is essential to synthetic circuit design. Selecting a promoter—commonly, either a strong native promoter, a viral-derived promoter, or a synthetic promoter that recruits transcriptional activator proteins [27, 28, 58]—is often the method of choice to change the rate of transcription. However, promoter choice can be a blunt tool, as mRNA processing, mRNA transport, translational initiation, and protein stability can all contribute to levels of transgene expression. In this work, we simultaneously measured mRNA and protein levels in order to assess impacts on both transcription and translation at steady state (Fig. 1). Indeed, we find that strong promoters both transcribe more mRNA and have higher effective translation rates (Fig. 2). To explore how other experimentally-accessible, genomically-encoded variables affect expression, we investigated how the PAS in the 3′ UTR (Fig. 3), coding sequence identity (Fig. 4), and 5′ UTR identity (Figs 5 and 6) jointly determine transgene levels.

In prokaryotes, steady-state mRNA transcript levels strongly correlate with translation initiation rate, though RNA secondary structures such as hairpins, G-quadruplexes, and i-motifs in the 5′ and 3′ UTRs all affect translation [59]. In eukaryotes, the transcriptional landscape is more complicated, with processes such as splicing, polymerase termination, polyadenylation, and transcription degradation driven from relatively opaque sequences with no discernible secondary structure. In fact, the sequence of the 5′ UTR encoded by a functionally defined synthetic promoter may impact both the translation rate and mRNA stability of the transgenic transcripts in human cells [54, 60, 61]. Nevertheless, within the set of promoters analyzed here, we observe that the choice of the canonical promoter has a larger impact on the effective translation rate of a transcript than the 5′ UTR sequence of the transcript (Fig. 6), pointing to the role of promoters in directing mRNA maturation and transport [57, 62].

Within our panel of promoters, splicing within the 5′ UTR increases effective translation rates, consistent with previous observations of splicing-mediated enhancement of gene expression [2, 40, 55, 63–65]. Specifically, the effective translation rate of EF1α is six times higher than that of EFS, which has the EF1α intron removed (Fig. 2). As confirmed by long-read sequencing, the post-spliced 5′ UTRs of transcripts expressed from these promoters differ by only 10 nucleotides. This expression difference may result from the recruitment of splicing factors that aid in not only splicing but also nuclear export of the mRNA transcripts, as has been observed with native genes [55, 63].

A growing number of synthetic circuit designs place synthetic target sites for post-transcriptional control in the 5′ or 3′ UTR or utilize synthetic introns to encode for circuit components [8, 11, 66]. Understanding how the transcript isoform influences transcript processing and translation will inform design of robust circuits. Here, we find that mature RNA isoforms are highly uniform (Fig. 5). Remarkably, only a small fraction of mature transcripts have non-canonical TSS or TES usage, splicing patterns, or readthrough, suggesting that it is reasonable to conceptually model each gene as its dominant isoform, instead of having to rely on a more complicated population model that accounts for isoform diversity. Rather, variability in transcription and translation rates may explain the variation in RNA and protein levels within a design. Since direct RNA sequencing quantifies only mature transcripts—those that are polyadenylated—we cannot assess the extent of immature, unprocessed transcripts that are produced. However, given the inherent instability of transcripts lacking a polyA tail, these transcripts are likely quickly degraded and would not contribute substantially to steady-state mRNA or protein levels.

As PAS putatively affects mRNA stability, we investigated three commonly used 3′ UTR sequences used in synthetic biology; two viral-derived sequences (WPRE, SV40) and one native sequence (bGH). All three are commonly used and are often arbitrarily paired to reduce the probability of plasmid recombination. While the SV40 and bGH PASs behave similarly,(Fig. 3B and D), we find that WPRE, a virally-derived sequence that is necessary for lentiviral production (Supplementary Fig. S5), localizes mRNA transcripts to puncta in the cytoplasm and decreases the effective translation rate (Fig. 3). Potentially, these puncta may represent specific subcellular structures such as stress granules or P bodies [67, 68]. For gene circuits that act at the post-transcriptional level (such as a microRNA-based iFFL [11] or an antisense integral controller [9]), the localization of mRNA transcripts may limit performance by altering the local concentrations of circuit components. Further study of mRNA localization can better inform selection of 3′ UTR elements to maximize circuit response and minimize variability. Nevertheless, we observed that these PAS effects may depend on copy number or integration context (Fig. 3).

When comparing the expression profiles of promoters between transfection and integration systems, we identified cases where the transfection profile offers reasonable predictability of expression within the integrated context. While complex coupling can occur on multi-gene plasmids [41], single-gene plasmids are free from coupling to adjacent genes and may offer a rapid prototyping platform for transcriptional units. Particularly with the same pairings of promoter and PAS, we found that transfection experiments could predict the relative expression profiles for integrated transgenes, albeit with different absolute levels of expression (Supplementary Fig. S5). Interestingly, we found that weak promoters—particularly EFS—exhibit higher mRNA and protein expression in PiggyBac integration than predicted by transfection (Supplementary Fig. S5A). The selection pressure applied during cell line creation may cause the expression from weaker promoters to skew higher due to the co-transcriptional expression of an antibiotic resistance gene. Taking into account the time-intensive workflow for cell line generation and the limitations of HCR Flow-FISH detection at low mRNA copy copy number, characterization in transfection offers a way to quickly assess mRNA distributions with higher resolution than in low-copy number cell lines. Therefore, transfection characterization results have predictive power as long as promoter, gene, and PAS sequences remain the same (Fig. 4 and

Supplementary Fig. S5). This predictability is important for the development of gene circuits that function reliably in engineered cell lines.

Our study also allowed us to consider how transgene expression interacts with endogenous gene expression in engineered cell lines. First, we observed very consistent transcription start and end site usage across transgenes randomly integrated in the genome (Fig. 5C), demonstrating that the local sequence context does not heavily impact transgene expression through readthrough from upstream promoters and/or genes. Second, we find that there is very little variability in endogenous gene expression across cells with integrated transgenes driven by different promoters of different strengths, suggesting that transgene expression does not cause competition for cellular resources (Fig. 5, Supplementary Fig. S11C, and Supplementary Table S2). Together, our two findings suggest that transgenes are often self-contained gene regulatory units that exert minimal impact on native gene regulatory mechanisms and networks. Further investigation of the durability of transgene expression over time and characterization of effects on endogenous genes in more cell types could aid in identifying genetic elements with reliable performance in different contexts.

With the increasing number of studies using library screening to identify genetic elements [12, 53, 69], HCR Flow-FISH might be adapted to screen larger libraries of genetic sequences for desired profiles of RNA expression. However, while we have demonstrated the value of high-resolution profiling with HCR Flow-FISH and direct RNA sequencing via long-reads for a handful of transgenes, scaling these methods to hundreds or thousands of designs remains unfeasible. In its current embodiment, HCR Flow-FISH characterization only requires basic biochemistry and access to a flow cytometer. However, the handling time over a three-day workflow limits throughput to ∼100 transgene variants. Integration of HCR Flow-FISH with a microfluidic platform with automated wash steps would greatly increase capacity and could reduce variability in mRNA labeling, supporting broader adoption into characterization workflows [38].

Through single-cell profiling of RNA and protein expression as well as analysis of RNA isoforms, we identify determinants of transgene expression levels in engineered cells and interactions of these transgenes with endogenous gene expression. We find that promoter choice for transgene expression influences both RNA transcript abundance and effective translation rate. The effective translation rate is further affected by the transgene coding sequence and 3′ UTR sequence. Differences in effective translation rates from constructs may be larger across cell types and show cell-type specific profiles of expression, potentially explaining the heuristic choice. Our framework of profiling RNA and protein levels simultaneously in single cells can be expanded to additional cell types and genetic elements to identify new sequences as well as tune expression for diverse functions. Increasing knowledge of how genetic elements contribute to profiles of expression will support predictive design of programmable gene circuits with controlled functions in diverse cell types.

Acknowledgements

The authors thank Adam Beitz, Nat Wang, Brittany Lende-Dorn, Mary Ehmann, Jane Atkinson, and Maria Castellanos for their helpful feedback on the manuscript.

Author contributions: Emma L. Peterman (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing—original draft, Writing—review & editing), Deon S. Ploessl (Conceptualization, Data curation, Investigation, Writing—review & editing), Kasey S. Love (Conceptualization, Data curation, Formal analysis, Investigation, Writing—review & editing), Valeria Sanabria (Data curation, Formal analysis, Investigation), Rachel F. Daniels (Data curation, Formal analysis, Investigation, Writing—review & editing), Christopher P. Johnstone (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Writing—review & editing), Diya R. Godavarti (Data curation, Investigation, Writing—review & editing), Sneha R. Kabaria (Conceptualization, Data curation, Investigation, Writing—review & editing), Conrad G. Oakes (Methodology, Validation), Athma A. Pai (Funding acquisition, Project administration, Supervision, Writing—review & editing), and Kate E. Galloway (Conceptulatization, Funding acquisition, Project administration, Supervision, Writing—original draft, Writing—review & editing).

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health [R35-GM143033 to K.E.G., R35-GM133762 to A.A.P.]; the National Science Foundation CAREER [2339986 to K.E.G., 2237568 to A.A.P.]; the Institute for Collaborative Biotechnologies; the Air Force Research Laboratory MURI [FA9550-22-1-0316 to K.E.G.]; and the National Science Foundation GRFP [1745302 to E.L.P., K.S.L., and S.R.K.]. Work completed at the MIT BioMicro Center was supported in part by the Koch Institute Support (core) Grant P30-CA014051 from the National Cancer Institute. Funding to pay the Open Access publication charges for this article was provided by MURI (FA9550-22-1-0316).

Data availability

Raw flow cytometry data and fluorescence images are deposited at Zenodo (10.5281/zenodo.15175222). Long-read sequencing data are deposited with the National Library of Medicine under BioProject accession PRJNA1191811. Data and code have been uploaded to Zenodo (10.5281/zenodo.15358310 and 10.5281/zenodo.15175222). Plasmids listed in tables S4 and S5 will be deposited at Addgene. All other plasmids are available upon request.

References

1. Ietswaart R, Smalec BM, Xu A *et al.* Genome-wide quantification of RNA flow across subcellular compartments reveals

determinants of the mammalian transcript life cycle. *Mol Cell* 2024;**84**:2765–84. https://doi.org/10.1016/j.molcel.2024.06.008
2. Choquet K, Patop IL, Churchman LS. The regulation and function of post-transcriptional RNA splicing. *Nat Rev Genet* 2025;**26**:378–94. https://doi.org/10.1038/s41576-025-00836-z
3. Schwanhäusser B, Busse D, Li N *et al.* Global quantification of mammalian gene expression control. *Nature* 2011;**473**:337–42. https://doi.org/10.1038/nature10098
4. Abreu RdS, Penalva LO, Marcotte EM *et al.* Global signatures of protein and mRNA expression levels. *Mol Biosyst* 2009;**5**:1512–26. https://doi.org/10.1039/b908315d
5. Richards AL, Watza D, Findley A *et al.* Environmental perturbations lead to extensive directional shifts in RNA processing. *PLOS Genet* 2017;**13**:e1006995. https://doi.org/10.1371/journal.pgen.1006995
6. Pai AA, Luca F. Environmental influences on RNA processing: biochemical, molecular and genetic regulators of cellular response. *Wiley Interdiscip Rev RNA* 2019;**10**:e1503. https://doi.org/10.1002/wrna.1503
7. Zhu R, del Rio-Salgado JM, Garcia-Ojalvo J *et al.* Synthetic multistability in mammalian cells. *Science* 2022;**375**:eabg9765. https://doi.org/10.1126/science.abg9765
8. Jones RD, Qian Y, Siciliano V *et al.* An endoribonuclease-based feedforward controller for decoupling resource-limited genetic modules in mammalian cells. *Nat Commun* 2020;**11**:5690. https://doi.org/10.1038/s41467-020-19126-9
9. Frei T, Chang CH, Filo M *et al.* A genetic mammalian proportional–integral feedback control circuit for robust and precise gene regulation. *Proc Natl Acad Sci* 2022;**119**:e2122132119. https://doi.org/10.1073/pnas.2122132119
10. Peterman EL, Ploessl DS, Galloway KE. Accelerating diverse cell-based therapies through scalable design. *Annu Rev Chem Biom Eng* 2024;**15**:267–92. https://doi.org/10.1146/annurev-chembioeng-100722-121610
11. Love KS, Johnstone CP, Peterman EL *et al.* Model-guided design of microRNA-based gene circuits supports precise dosage of transgenic cargoes into diverse primary cells. *Cell Syst* 2025;101269. https://doi.org/10.1016/j.cels.2025.101269
12. O'Connell RW, Rai K, Piepergerdes TC *et al.* Ultra-high throughput mapping of genetic design space. bioRxiv, https://doi.org/10.1101/2023.03.16.532704, 5 May 2025, preprint: not peer reviewed.
13. Dunlop MJ, Cox RS, Levine JH *et al.* Regulatory activity revealed by dynamic correlations in gene expression noise. *Nat Genet* 2008;**40**:1493–8. https://doi.org/10.1038/ng.281
14. Swain PS, Elowitz MB, Siggia ED. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* 2002;**99**:12795–800. https://doi.org/10.1073/pnas.162041399
15. Quarton T, Kang T, Papakis V *et al.* Uncoupling gene expression noise along the central dogma using genome engineered human cell lines. *Nucleic Acids Res* 2020;**48**:9406–13. https://doi.org/10.1093/nar/gkaa668
16. Qin JY, Zhang L, Clift KL *et al.* Systematic comparison of constitutive promoters and the doxycycline-inducible promoter. *PLoS One* 2010;**5**:e10611. https://doi.org/10.1371/journal.pone.0010611
17. Ede C, Chen X, Lin MY *et al.* Quantitative analyses of core promoters enable precise engineering of regulated gene expression in mammalian cells. *ACS Synth Biol* 2016;**5**:395–404. https://doi.org/10.1021/acssynbio.5b00266
18. Wen S, Zhang H, Li Y *et al.* Characterization of constitutive promoters for piggyBac transposon-mediated stable transgene expression in mesenchymal stem cells (MSCs). *PLoS One* 2014;**9**:e94397. https://doi.org/10.1371/journal.pone.0094397
19. Takahashi K, Galloway KE. RNA-based controllers for engineering gene and cell therapies. *Curr Opin Biotechnol* 2024;**85**:103026. https://doi.org/10.1016/j.copbio.2023.103026
20. Dou Y, Lin Y, Wang T *et al.* The CAG promoter maintains high-level transgene expression in HEK293 cells. *FEBS Open Bio* 2020;**11**:95–104. https://doi.org/10.1002/2211-5463.13029
21. Foreman R, Wollman R. Mammalian gene expression variability is explained by underlying cell state. *Mol Syst Biol* 2020;**16**:e9146. https://doi.org/10.15252/msb.20199146
22. Mamrak NE, Alerasool N, Griffith D *et al.* The kinetic landscape of human transcription factors. bioRxiv, https://doi.org/10.1101/2022.06.01.494187, 2 June 2022, preprint: not peer reviewed.
23. Popp AP, Hettich J, Gebhardt J. Altering transcription factor binding reveals comprehensive transcriptional kinetics of a basic gene. *Nucleic Acids Res* 2021;**49**:6249–66. https://doi.org/10.1093/nar/gkab443
24. Fu X, Patel HP, Coppola S *et al.* Quantifying how post-transcriptional noise and gene copy number variation bias transcriptional parameter inference from mRNA distributions. *eLife* 2022;**11**:e82493. https://doi.org/10.7554/eLife.82493
25. Choi HMT, Beck VA, Pierce NA. Next-generation *in situ* hybridization chain reaction: higher gain, lower cost, greater durability. *ACS Nano* 2014;**8**:4284–94. https://doi.org/10.1021/nn405717p
26. Choi HMT, Schwarzkopf M, Fornace ME *et al.* Third-generation *in situ* hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* 2018;**145**:dev165753. https://doi.org/10.1242/dev.165753
27. Donahue PS, Draut JW, Muldoon JJ *et al.* The COMET toolkit for composing customizable genetic programs in mammalian cells. *Nat Commun* 2020;**11**:779. https://doi.org/10.1038/s41467-019-14147-5
28. Li HS, Israni DV, Gagnon KA *et al.* Multidimensional control of therapeutic human cell function with synthetic gene circuits. *Science* 2022;**378**:1227–34. https://doi.org/10.1126/science.ade0156
29. Wang Q, Liu J, Janssen JM *et al.* Precise homology-directed installation of large genomic edits in human cells with cleaving and nicking high-specificity Cas9 variants. *Nucleic Acids Res* 2023;**51**:3465–84. https://doi.org/10.1093/nar/gkad165
30. Blanch-Asensio A, van der Vaart B, Vinagre M *et al.* Generation of AAVS1 and CLYBL STRAIGHT-IN v2 acceptor human iPSC lines for integrating DNA payloads. *Stem Cell Res* 2023;**66**:102991. https://doi.org/10.1016/j.scr.2022.102991
31. Blanch-Asensio A, Grandela C, Mummery CL *et al.* STRAIGHT-IN: a platform for rapidly generating panels of genetically modified human pluripotent stem cell lines. *Nat Protoc* 2024;1–44. https://doi.org/10.1038/s41596-024-01039-2
32. Aznauryan E, Yermanos A, Kinzina E *et al.* Discovery and validation of human genomic safe harbor sites for gene and cell therapies. *Cell Rep Methods* 2022;**2**:100154. https://doi.org/10.1016/j.crmeth.2021.100154
33. Chavez M, Rane DA, Chen X *et al.* Stable expression of large transgenes via the knock-in of an integrase-deficient lentivirus. *Nat Biomed Eng* 2023;**7**:661–71. https://doi.org/10.1038/s41551-023-01037-x
34. Green MR, Sambrook J. Nested polymerase chain reaction (PCR). *Cold Spring Harb Protoc* 2019;**2019**:436–56. https://doi.org/10.1101/pdb.prot095182
35. Korbie DJ, Mattick JS. Touchdown PCR for increased specificity and sensitivity in PCR amplification. *Nat Protoc* 2008;**3**:1452–6. https://doi.org/10.1038/nprot.2008.133
36. Jia Z, Dong Y, Xu H *et al.* Optimizing the hybridization chain reaction-fluorescence in situ hybridization (HCR-FISH) protocol for detection of microbes in sediments. *Mar Life Sci Technol* 2021;**3**:529–41. https://doi.org/10.1007/s42995-021-00098-8
37. Carbon S, Mungall C. AGene Ontology Data Archive. Geneva, Switzerland: Zenodo, 2025. https://doi.org/10.5281/zenodo.15066566
38. Arrigucci R, Bushkin Y, Radford F *et al.* FISH-Flow, a protocol for the concurrent detection of mRNA and protein in single cells using

fluorescence in situ hybridization and flow cytometry. *Nat Protoc* 2017;**12**:1245–60. https://doi.org/10.1038/nprot.2017.039
39. Gam JJ, DiAndreth B, Jones RD *et al.* A 'poly-transfection' method for rapid, one-pot characterization and optimization of genetic systems. *Nucleic Acids Res* 2019;**47**:e106. https://doi.org/10.1093/nar/gkz623
40. Cabrera A, Edelstein HI, Glykofrydis F *et al.* The sound of silence: transgene silencing in mammalian cell engineering. *Cell Syst* 2022;**13**:950–73. https://doi.org/10.1016/j.cels.2022.11.005
41. Johnstone CP, Galloway KE. Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Rep* 2022;**41**:111492. https://doi.org/10.1016/j.celrep.2022.111492
42. Johnstone CP, Love KS, Kabaria SR *et al.* Gene syntax defines supercoiling-mediated transcriptional feedback. bioRxiv, https://doi.org/10.1101/2025.01.19.633652, 19 January 2025, preprint: not peer reviewed.
43. Blanch-Asensio A, Ploessl DS, Wang NB *et al.* STRAIGHT-IN Dual: a platform for dual, single-copy integrations of DNA payloads and gene circuits into human induced pluripotent stem cells. bioRxiv, https://doi.org/10.1101/2024.10.17.616637, 17 October 2024, preprint: not peer reviewed.
44. Hong D, Jeong S. 3′UTR diversity: expanding repertoire of rna alterations in human mRNAs. *Mol Cells* 2023;**46**:48–56. https://doi.org/10.14348/molcells.2023.0003
45. Mayr C. Regulation by 3′-untranslated regions. *Annu Rev Genet* 2017;**51**:171–94. https://doi.org/10.1146/annurev-genet-120116-024704
46. Mayr C. What are 3′ UTRs doing? *Cold Spring Harb Perspect Biol* 2019;**11**:a034728. https://doi.org/10.1101/cshperspect.a034728
47. Szostak E, Gebauer F. Translational control by 3′-UTR-binding proteins. *Brief Funct Genomics* 2013;**12**:58–65. https://doi.org/10.1093/bfgp/els056
48. Zamft B, Bintu L, Ishibashi T *et al.* Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc Natl Acad Sci USA* 2012;**109**:8948–53. https://doi.org/10.1073/pnas.1205063109
49. Cohen E, Zafrir Z, Tuller T. A code for transcription elongation speed. *RNA Biol* 2018;**15**:81–94. https://doi.org/10.1080/15476286.2017.1384118
50. Qiu Y, Kang YM, Korfmann C *et al.* The GC-content at the 5′ ends of human protein-coding genes is undergoing mutational decay. *Genome Biol* 2024;**25**:219. https://doi.org/10.1186/s13059-024-03364-x
51. Kudla G, Lipinski L, Caffin F *et al.* High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* 2006;**4**:e180. https://doi.org/10.1371/journal.pbio.0040180
52. Courel M, Clément Y, Bossevain C *et al.* GC content shapes mRNA storage and decay in human cells. *eLife* 2019;**8**:e49708. https://doi.org/10.7554/eLife.49708
53. Castillo-Hair S, Fedak S, Wang B *et al.* Optimizing 5′UTRs for mRNA-delivered gene editing using deep learning. *Nat Commun* 2024;**15**:5284. https://doi.org/10.1038/s41467-024-49508-2
54. Jia L, Mao Y, Ji Q *et al.* Decoding mRNA translatability and stability from the 5′ UTR. *Nat Struct Mol Biol* 2020;**27**:814–21. https://doi.org/10.1038/s41594-020-0465-x
55. Shaul O. How introns enhance gene expression. *Int J Biochem Cell Biol* 2017;**91**:145–55. https://doi.org/10.1016/j.biocel.2017.06.016
56. Chen M, Lyu G, Han M *et al.* 3′ UTR lengthening as a novel mechanism in regulating cellular senescence. *Genome Res* 2018;**28**:285–94. https://doi.org/10.1101/gr.224451.117
57. Palazzo AF, Qiu Y, Kang YM. mRNA nuclear export: how mRNA identity features distinguish functional RNAs from junk transcripts. *RNA Biol* 2024;**21**:145–56. https://doi.org/10.1080/15476286.2023.2293339
58. Kabaria SR, Bae Y, Ehmann ME *et al.* Programmable promoter editing for precise control of transgene expression. bioRxiv, https://doi.org/10.1101/2024.06.19.599813, 14 July 2024, preprint: not peer reviewed.
59. Cetnar DP, Hossain A, Vezeau GE *et al.* Predicting synthetic mRNA stability using massively parallel kinetic measurements, biophysical modeling, and machine learning. *Nat Commun* 2024;**15**:9601. https://doi.org/10.1038/s41467-024-54059-7
60. Hinnebusch AG, Ivanov IP, Sonenberg N. Translational control by 5′-untranslated regions of eukaryotic mRNAs. *Science* 2016;**352**:1413–6. https://doi.org/10.1126/science.aad9868
61. Araujo PR, Yoon K, Ko D *et al.* Before it gets started: regulating translation at the 5′ UTR. *Int J Genom* 2012;**2012**:475731. https://doi.org/10.1155/2012/475731
62. Schubert T, Köhler A. Mediator and TREX-2: emerging links between transcription initiation and mRNA export. *Nucleus* 2016;**7**:126–31. https://doi.org/10.1080/19491034.2016.1169352
63. Fiszbein A, Krick KS, Begg BE *et al.* Exon-mediated activation of transcription starts. *Cell* 2019;**179**:1551–65. https://doi.org/10.1016/j.cell.2019.11.002
64. Kowal EJK, Sakai Y, McGurk MP *et al.* Sequence determinants of intron-mediated enhancement learned from thousands of random introns. bioRxiv, https://doi.org/10.1101/2024.10.29.620880, 29 October 2024, preprint: not peer reviewed.
65. Seczynska M, Bloor S, Cuesta SM *et al.* Genome surveillance by HUSH-mediated silencing of intronless mobile elements. *Nature* 2022;**601**:440–45. https://doi.org/10.1038/s41586-021-04228-1
66. DiAndreth B, Wauford N, Hu E *et al.* PERSIST platform provides programmable RNA regulation using CRISPR endoRNases. *Nat Commun* 2022;**13**:2582. https://doi.org/10.1038/s41467-022-30172-3
67. Anderson P, Kedersha N. Stress granules: the Tao of RNA triage. *Trends Biochem Sci* 2008;**33**:141–50. https://doi.org/10.1016/j.tibs.2007.12.003
68. Ren Z, Tang W, Peng L *et al.* Profiling stress-triggered RNA condensation with photocatalytic proximity labeling. *Nat Commun* 2023;**14**:7390. https://doi.org/10.1038/s41467-023-43194-2
69. Wu MR, Nissim L, Stupp D *et al.* A high-throughput screening and computation platform for identifying synthetic promoters with enhanced cell-state specificity (SPECS). *Nat Commun* 2019;**10**:2880. https://doi.org/10.1038/s41467-019-10912-8

Received: December 3, 2024. **Revised:** April 24, 2025. **Editorial Decision:** May 7, 2025. **Accepted:** May 30, 2025