*Nucleic Acids Research*, 2025, **53**, gkaf097
https://doi.org/10.1093/nar/gkaf097
**Gene regulation, Chromatin and Epigenetics**

# Sequence-dependent and -independent effects of intron-mediated enhancement learned from thousands of random introns

**Emma J.K. Kowal, Yuta Sakai, Michael P. McGurk, Zoe J. Pasetsky, Christopher B. Burge***

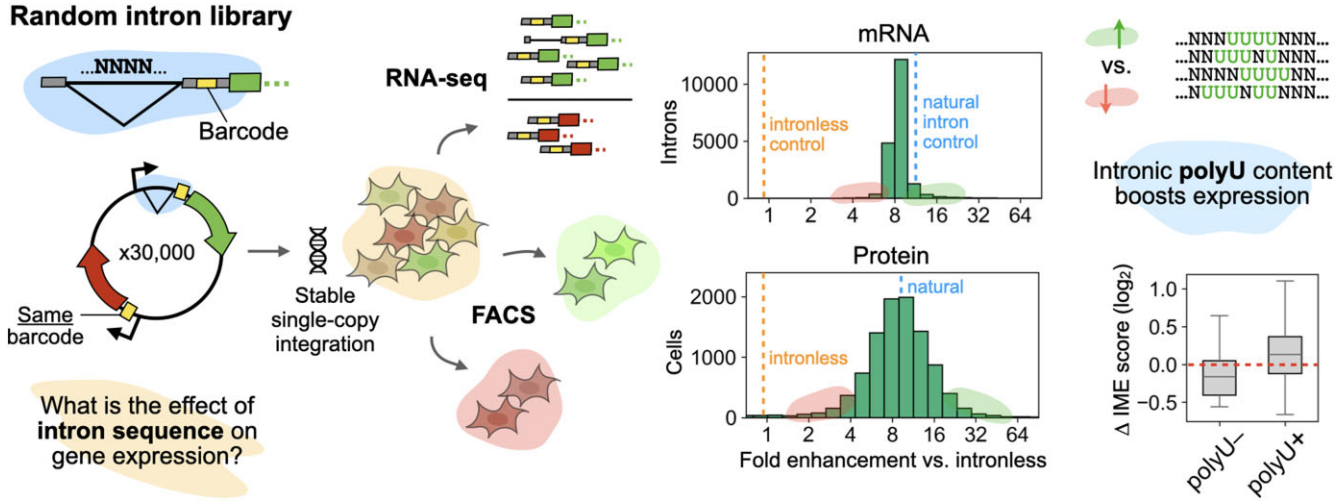
Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139, United States

*To whom correspondence should be addressed. Email: cburge@mit.edu

## Abstract

Spliceosomal introns are a ubiquitous feature of eukaryotic genes, whose presence often boosts the expression of their host gene, a phenomenon known as intron-mediated enhancement (IME). IME has been noted across diverse genes and organisms but remains mysterious in many respects. For example, how does intron sequence affect the magnitude of IME? In this study, we performed a massively parallel reporter assay (MPRA) to assess the effect of varying intron sequence on gene expression in a high-throughput manner, in human cells, using tens of thousands of synthetic introns with natural splice sites and randomized internal sequence. We observe that most random introns splice efficiently and enhance gene expression as well as or better than fully natural introns. Nearly all introns stimulate gene expression ~eight-fold above an intronless control, at both mRNA and protein levels, suggesting that the primary mechanism acts to increase mRNA levels. IME strength is positively associated with splicing efficiency and with the intronic content of poly-uridine stretches, which we confirm using reporter experiments. In sum, this work assesses the IME of a diverse library of introns and uncovers sequence-dependent aspects, but suggests that enhancement of gene expression is a general property of splicing, largely independent of intron sequence.

## Graphical abstract




## Introduction

Despite appearing to be unnecessary for gene expression, introns are ubiquitous in eukaryotic genomes, with more than 10 per protein-coding gene in humans on average [1, 2]. The complex process of splicing required to remove introns from pre-mRNA is performed by the spliceosome, a dynamic macromolecular machine comprising many dozens of proteins and five small nuclear RNA-protein complexes [3, 4]. The spliceosome and other splicing regulatory proteins parse an extremely subtle and complex set of signals in the primary sequence of the pre-mRNA to identify introns and exons, signals which can be alternatively interpreted under different conditions to generate many isoforms from the same gene [4, 5]. Each splicing reaction also results in the deposition of an exon junction complex, a large protein assembly which remains bound to the mRNA as it is exported from the nucleus into

**Received:** October 29, 2024. **Revised:** January 28, 2025. **Editorial Decision:** January 30, 2025. **Accepted:** February 7, 2025

the cytoplasm [6, 7]. Thus each intron, in the process of being spliced, orchestrates the interaction of the nascent transcript with a vast cohort of RNA-binding proteins (RBPs) and Ribonucleoprotein, some of which persist on the mRNA in its place long after its removal.

A curious aspect of intron splicing is that despite the extra time and effort required for their processing, introns generally enhance rather than inhibit gene expression (Fig. 1A). This was noticed very soon after the discovery of introns [8, 9] and has since been described in diverse genes in a wide variety of plant, animal and fungal systems (reviewed in [10]). Indeed, many promoters popularly used to drive expression of minigenes such as SV40, Cytomegalovirus (CMV), beta-globin, thymidine kinase, Ubiquitin C (UbC), and EF1a commonly include a portion of 5′UTR with an intron that has been shown to boost expression significantly compared to the promoter alone. The magnitude of this intron-mediated enhancement (IME) can range from negligible to hundreds of times more mRNA and protein produced from a spliced versus intronless gene [11, 12].

The mechanism of these effects is hypothesized to rely on each intron's recruitment of *trans*-acting RNA and protein factors. The IME literature contains examples of introns that can positively impact nearly all stages of mRNA processing, from transcription and splicing through export, translation and decay, with individual introns often exerting multiple distinct effects [13–19] reviewed in [10, 20, 21]. Certain general rules have emerged, such as that introns tend to enhance most strongly near the 5′ end of genes, with first introns often being especially potent, [22, 10, 23], but otherwise any trend seems to have as many exceptions as examples.

One of the many questions about IME that remain unanswered: what is the effect of intron sequence on the magnitude of enhancement? It has been observed that different introns tested in an identical context can boost expression to varying extents, sometimes by up to an order of magnitude [10, 22, 24–28]. For example, several introns, all well-spliced, exerted distinct effects in the GUS1 reporter in *Arabidopsis thaliana* [27]. Subsequent development of a computational method to predict the magnitude of an intron's IME in plants from its sequence—the IMEter—enabled the discovery of a motif, TTNGATYTG, which in reporter experiments transformed an intron with no IME into one that increased mRNA accumulation 24-fold and protein accumulation 40-fold relative to the intronless control [29]. However, relatively little is known about the sequence determinants of IME in animals.

In this work, we sought to understand the effect of intron sequence variation on the strength of IME in human cells. We designed and executed a massively parallel reporter assay (MPRA) to test the expression of tens of thousands of unique introns, using high-efficiency and low-background (HILO) recombination-mediated cassette exchange (RMCE) technology [30]. This allowed us to insert introns at single copy into a constant genetic context and to compare their expression at the mRNA and protein levels in a native, controlled genomic setting, revealing insights into the nature of IME.

## Materials and methods

### Mammalian cells, plasmids and reagents

HEK293T A2 and HeLa A12 HILO-RMCE cells and plasmids pEM689 and pEM784 were kindly provided by Eugene Makeyev at Nanyang Technological University, Singapore [30]. All oligos used in this research are listed in Supplementary Table S1 and were manufactured by Eton Bioscience or IDT. All plasmids used in this research are listed in Supplementary Table S2. All other reagents and their sources are listed in Supplementary Table S3.

### Cloning of individual intron reporters

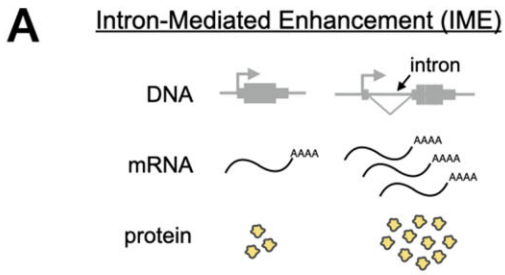
All plasmids used in this study are adapted from pEM689. The CMV enhancer, chicken ß-actin promoter and chimeric intron from dTomato was replaced to minimal CMV promoter to create the workhorse backbone plasmid pEK1 for the library and individual intron reporters. Enhanced green fluorescent protein (EGFP) gene with UbC promoter, UbC 5′UTR with intron, UbC 3′UTR and dTomato was inserted into pEK1 to construct the pilot construct pEK2 to insert the introns. The intron sequence in the UbC 5′UTR was variably replaced with alternative intron sequences or no intron. All assembled plasmids were transformed into *Mix & Go*! DH5 Alpha competent cells before plating on 2% LB-agar plates containing ampicillin (100 μg/mL) and incubated at 37°C overnight to obtain colonies. The colonies were picked and cultured in 2mL LB media containing ampicillin (100 μg/mL) and grown overnight to purify the plasmids using QIAprep Spin Miniprep kit. The complete sequence of the insert or the whole plasmid was confirmed by Sanger sequencing (performed by Eton Biosciences) or by full-plasmid sequencing (performed by Plasmidsaurus).

### Cloning of random intron library

See Supplementary methods for full details; see also Fig. 2B for schematic of procedure. The final dictionary of all introns detected in RNA-seq of the library and their associated barcodes is available in Supplementary Table S4.

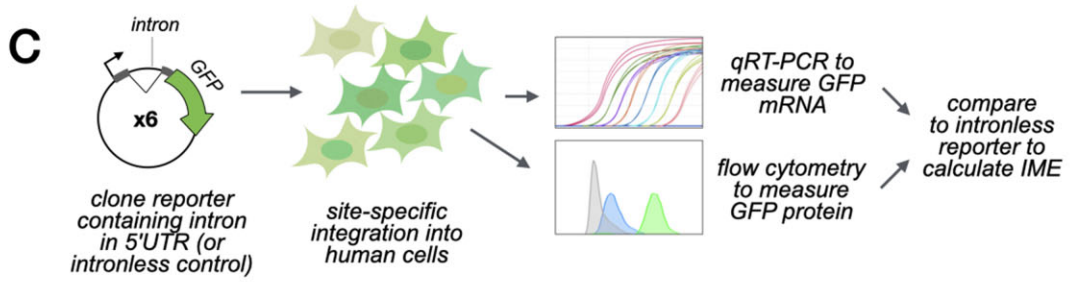
### Cell culture and generation of transgenic lines

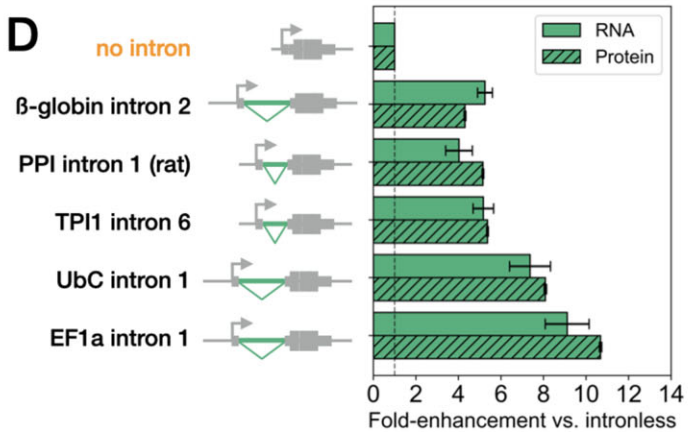
Both HEK293T A2 and HeLa A12 HILO-RMCE cells were grown in Dulbecco's modifiedEagle's medium (DMEM), with high glucose and pyruvate, supplemented with 10% fetal bovine serum and 10% penicillin/streptomycin. Cells were passaged 1:10 every 2–3 days by washing with PBS and treated with 0.25% Trypsin following standard procedures and tested for Mycoplasma by PCR periodically. The intron reporter cassettes were integrated by following authors' recommendations: briefly, we first plated $1.5 \times 10^5$ cells per well on a 12-well plate coated with 0.1mg/mL Poly-d-Lysine. After culturing the cells for 24 h in antibiotic-free media, the cells in each well were co-transfected using Lipofectamine 2000 with 500 ng of a reporter plasmid plus 0.5% wt/wt Cre recombinase vector pEM784 [30], and allowed them 24–48 h before selection with 4–8μg/mL Puromycin for 2 weeks to ensure genomic integration. Cells were monitored throughout for surviving colony formation and wells were split and pooled before becoming overconfluent. At this stage aliquots of 0.5 to $5 \times 10^6$ cells were frozen in media with 10% Dimethyl sulfoxide (DMSO) and later thawed again when necessary. For experiments with inducible promoters, cell lines with integrated dox-inducible reporters were treated with 0–16 μg/mL Doxycycline in the cell culture media and allowed to grow for 24 h before trypsinization and flow cytometry.




B

| Intron | Organism | Length (nt) | Composition (%) | | | | Location in gene | Reported mode of IME | Reported increase in expression (protein) | Citation(s) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | A | C | G | T | | | | |
| ß-globin intron 2 | *H. sapiens* | 850 | 28 | 16 | 15 | 41 | ORF | stability, 3'end processing, translation | 5-fold | Lu & Cullen 2003 |
| PPI intron 1 | *R. norvegicus* | 119 | 17 | 29 | 24 | 29 | 5'UTR | translation | 23-fold | Wiegand 2003 |
| TPI1 intron 6 | *H. sapiens* | 128 | 16 | 26 | 34 | 24 | ORF | transcription, translation, nuclear export | 29-fold | Nott 2003 |
| UbC intron 1 | *H. sapiens* | 812 | 16 | 21 | 37 | 21 | 5'UTR | transcription | 7-fold | Bianchi 2009, 2013 |
| EF1a intron 1 | *H. sapiens* | 943 | 15 | 27 | 33 | 25 | 5'UTR | - | 18-fold | Kim 2002 |



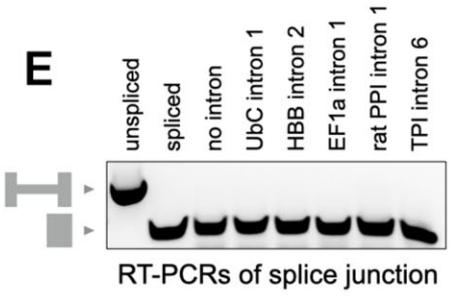




**Figure 1.** Different introns lead to different increases in gene expression. (**A**) Schematic of IME. (**B**) Table of introns selected for testing in pilot experiment along with their reported modes and degrees of enhancement. (**C**) Reporter experiment workflow. (**D**) Pilot experiment showing effects on GFP reporter expression stimulated by insertion of different introns, using transgenic human HEK293T cells. Numbers are normalized to show fold change with respect to intronless control (dotted line). mRNA was measured from $n = 5$ qRT-PCR replicates and protein from $n = \sim$20 000 cells using flow cytometry. Error bars denote standard error of the mean. (**E**) RT-PCR of RNA from cell lines in D showing efficient splicing of all introns. Lanes 1 and 2 are controls for unspliced and spliced band size, PCR of plasmid DNA for UbC intron-containing and intronless reporters, respectively.

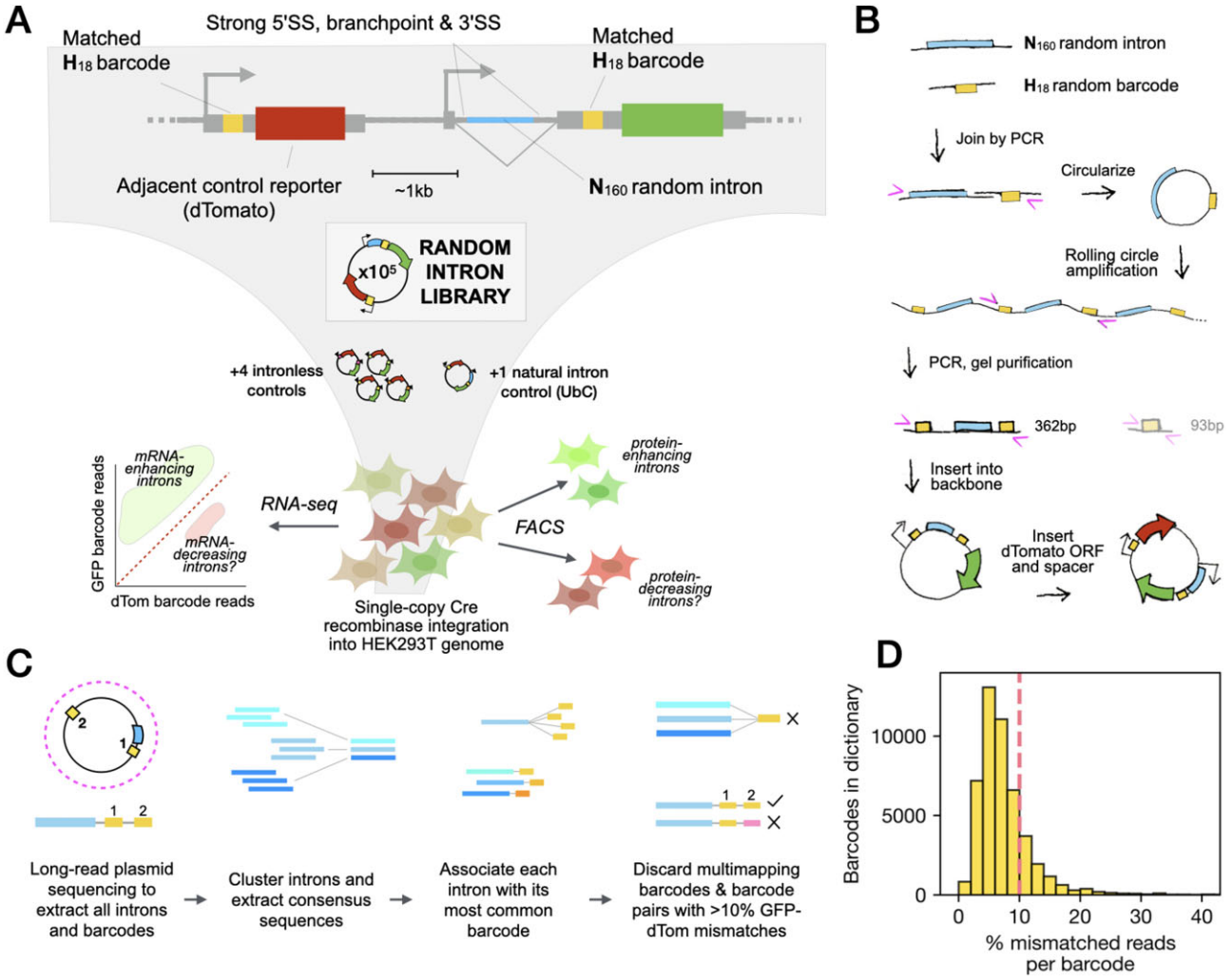



**Figure 2.** Development of a barcoded dual-reporter random intron library to measure IME effects of many introns in parallel. (**A**) Schematic of reporter design and workflow for random intron screen. (**B**) Cloning strategy for assembly of random library with random paired barcodes. H = A, C, or T. (**C**) Generation of barcode-to-intron map using reads from Oxford nanopore long-read sequencing of library plasmid DNA. (**D**) Distribution of mismatching rates between GFP and dTomato barcodes for all barcodes mapped to introns. Dotted line indicates cutoff for inclusion in further analyses (<10% mismatched).

## RT-PCR and qRT-PCR

For each RNA extraction, $10^5$–$10^6$ cells were washed once with PBS and either trypsinized and pelleted or lysed directly in the tissue culture well. RNA was subsequently extracted using either Zymo Quick RNA MiniPrep kit with in-situ DNase I treatment, or for higher throughput applications, with the Chemagic 360 RNA extraction protocol. Purified RNA was quantified by either Nanodrop or Synergy H1 plate reader and 300–1000 ng RNA per sample was reverse transcribed using SuperScript IV VILO For reversetranscriptase-polymerase chain reaction (RT-PCR) to verify correct splicing, 2 µL of each RT reaction was used as the template for a Polymerase-chain-reaction (PCR) using Phusion DNA polymerase PCR and exon-junction-spanning primers (UbC-splicecheck-3 and green fluorescent protein (GFP)-N-out). Products were separated and visualized using agarose gel electrophoresis with SYBRSAFE dye and relative band intensities were quantified using FIJI. For qRT-PCR, RT reactions and primers for genes of interested were arrayed in a 96-well plate on ice, mixed with PowerUp SYBR Green Master Mix, and then transferred in quadruplicate to a 384-well plate using the TECAN EVO. Plates were run on a QuantStudio 5 thermocycler (Thermo) using SYBR manufacturer's recommendations for cycling protocol. Cq values (mean of four technical replicates) were downloaded and data was analyzed using custom Python scripts. Each plate contained a serial dilution with at minimum five samples for plotting a standard curve and extracting primer efficiencies. Relative quantities for each target gene were calculated between samples of interest and the intronless control samples by exponentiating their deltaCq values using target-specific amplification efficiencies. GFP and dTomato values were then normalized to the geometric mean of 3 control targets (*GAPDH, RPL27, SRP14*) before computing their log ratio, which was finally again normalized to the intronless log ratio.

## Flow cytometry

Flow cytometry experiments were performed using LSR II HTS-1 and FACS Symphony A3 HTS-1 instruments (BD Biosciences) at the Koch Institute Flow Cytometry core. Cells intended for analysis were washed with PBS, trypsinized, pel-

leted, carefully resuspended, passed through a 35 μm filter lid to dissociate clumps (Corning) and transported on ice. Single live cells were identified by binning FSC-A/SSC-A, FSC-W/H, and SSC-W/H, with binning performed by flow core staff. GFP was measured in the FITC channel (Laser: 488nm, Filter: 515/20) and dTomato in the PE channel (Laser: 561nm, Filter: 586/15).

### RNA sequencing and analysis: sample preparation and sequencing run

For each RNA-seq sample, 2μg RNA extracted from the pooled random intron library was treated with Turbo DNase (Thermo) to remove any residual genomic DNA before further processing. We removed DNase using Zymo RNA clean & concentrate kit and next performed gene-specific RT using SuperScript IV (Thermo) and two primers, complementary to the GFP and dTomato N-termini, respectively, in order to generate cDNA of the two reporter gene 5′UTRs only. After first strand synthesis we treated the samples with RNase cocktail (Thermo) to avoid interference with downstream PCR. The cDNA was amplified with 16 cycles of PCR using Q5 polymerase (NEB) and custom sample-specific indexing primers. The PCR product was purified by a double-sided SPRI purification, first removing genomic DNA with 0.55x volumes of beads and then adding 85 μL beads for a final 1.4x volume. The amplicon was sequenced with either 50 + 50 nt PE reads for replicates 1–5, on an Illumina HiSeq at the Whitehead institute sequencing core, or 105 + 45 nt PE reads for replicates 6–10 using the Element AVITI instrument at the MIT BioMicroCenter.

### Read processing

Reads were first scanned for a correct sample index and for a match or close match to the constant sequence distinguishing the GFP from dTomato 5′UTR, and discarded if they were missing either of these. GFP reads were next classified into spliced, unspliced, or other by searching for a match to the expected exon-exon junction sequence. Unspliced reads with the expected intron sequence (constant 5′SS and 3′SS) were counted for that barcode as unspliced, while reads matching neither the expected spliced sequence nor the expected unspliced sequence were set aside for further analysis. Finally, classified reads were filtered for perfect or close matches to known and trusted barcodes and otherwise discarded. This yielded ∼125 million RNA-seq reads across 10 biological replicates. The total read counts per intron are available in Supplementary Table S5.

### Splicing and cryptic splicing analysis

Reads identified as originating from GFP were classified into spliced or unspliced based on the presence or absence of the expected exon 2 sequence TCGTGAA 3 nt downstream of the exon-exon junction. Unspliced reads were further categorized into canonical unspliced if they instead had the predicted sequence of the unspliced intron (AGTAGCG), or unknown isoform if they had a significantly different sequence. The splicing efficiency of each intron was obtained by tallying the fraction of correctly spliced reads per barcode, per sample, taking the median across all samples with at least 100 GFP reads. To identify instances of cryptic splicing, reads with unusual intronic sequence were partitioned into those with mismatches at the 5′-end or 3′-end (some reads belonging to both sets) and a window of 16 or 20 nt (for 5′ or 3′ respectively) from the read was used to scan along the predicted intron sequence, with Levenshtein distance calculated at each position. If all reads for a given barcode had an optimal match at the same position of the intron, and the Levenshtein distance between the read and intron at that position was ≤2, this position was identified as a cryptic splice site.

### Iterative FACS analysis

Fluorescence-assisted cell sorting (FACS) experiments were performed at the Koch Institute Flow Cytometry core using the BD FACSAria III instrument (BD Biosciences). Cells were prepared and sorted as described in Flow Cytometry methods section. For sort 1, 60 million pooled integrated random intron library cells were sorted in 2 replicate batches. Bins were drawn to select the top and bottom 10% of cells along the diagonal FITC versus PE signal axis and 1.8–2.6 million cells were collected per bin per replicate, into tubes containing 50% DMEM and 50% fetal bovine serum (FBS). These cells were plated in 4 × 15 cm dishes and allowed to grow for 4 days. After trypsinizing each plate from the first sort, cells were counted and 1 million cells from each plate were pelleted and snap-frozen in an aluminium block cooled to −80°C for later RNA extraction. Bins were re-drawn to capture the top 10% of the green-shifted populations and the bottom 10% of the red-shifted populations, and 0.6–0.7 million cells were collected per bin per replicate. These were plated in T25 flasks and allowed to grow for an additional 3 days before the final sort, at which point 0.15 million cells from each flask were pelleted and snap-frozen. The remainder of each flask was sorted into bins were re-drawn to enrich for the most extreme 10% once again, yielding 130–180 thousand cells per bin per replicate. These were immediately pelleted and snap-frozen such that downstream processing of all 12 samples from the iterative FACS experiment occurred in parallel. The total read counts per intron from RNA-seq of these bins are available in Supplementary Table S6.

### IME analyses

For bulk RNA-seq, we filtered all barcodes to consider only those detected in three or more samples, then ran DESeq2 to get intron-specific log2fold change estimates, using Ashr shrinkage [31,32]. For the iterative FACS libraries, we identified bins of introns detected in both trajectories, and designated “green” and “red” sets as those present in every stage of green sorting and no stage of red sorting, and vice versa. These sets in combination were used with mRNA-level IME scores to study the splicing and sequence features of introns as a function of their magnitude of gene expression enhancement.

## Data availability statement

The RNA-seq data underlying this article are available in the NCBI GEO Database, under accession code GSE278584. The primary analysis code is available at https://doi.org/10.5281/zenodo.14728464 as well as https://github.com/ejkk0/IME.

For methods concerning library cloning, long-read sequencing and analysis, barcode pairing analysis, and simulations, see Supplementary Methods.

## Results

### Different introns enhance gene expression to different extents

We first sought to reproduce the observation from plants that different well-spliced introns can stimulate different levels of expression from otherwise identical reporter genes. We used the HILO-RMCE system [30] for rapid generation of human cell lines with single-copy genomic integration of our reporter constructs. We selected five introns which had been reported in previous literature to enhance gene expression (Fig. 1B) and cloned each separately into the same position in the 5′UTR of an EGFP gene driven by the human UbC promoter. Each of these vectors, plus an otherwise identical intronless EGFP vector, was co-transfected with a Cre recombinase plasmid into HILO-RMCE HEK293T A2 cells (Fig. 1C), which were then maintained under puromycin selection for two weeks to ensure integration of the transfected DNA.

Each intron-containing cell line yielded higher steady-state levels of EGFP mRNA and protein than the intronless version, as measured by qRT-PCR and flow cytometry (Fig. 1C and D), and RT-PCR analysis of EGFP RNA from these transgenic cell lines showed that all introns were spliced with comparable high efficiency, also at steady-state (Fig. 1E). The magnitude of IME—defined as the fold increase in expression over intronless—varied from about four-fold to more than 10-fold at both mRNA and protein levels. Interestingly, across introns, the increase in protein yield was highly correlated to the increase in mRNA. The variability in IME attributable to these introns suggests that intron identity (i.e. sequence) contributes to differences in enhancement. For the purposes of this study, we defined IME as the increase in expression of a host gene attributable to the splicing of an intron. Thus, if an intron were to enhance expression by containing a transcriptional enhancer, or by stabilizing unspliced transcripts, for example, this would not be considered IME.

### Design and construction of a barcoded dual-reporter system to study IME

In order to investigate the relationship between intron sequence and IME, we conceived a large-scale reporter experiment to measure the enhancing activity of many introns in parallel (Fig. 2A). We reasoned that sequence-specific differences in IME may be driven by the recruitment of sequence-specific RBPs, which typically bind short, 4–8 nucleotide (nt) motifs within introns. Using a library of comparatively long (160 nt) random sequences allows exploration of the sequence-activity relationship of IME in an unbiased manner, with a large diversity of intron sequences and coverage of all short motifs in different relative arrangements.

We generated a DNA library of introns with 160 nt of fully random internal sequence, flanked by strong splicing signals (UbC 5′SS and EF1a branch point and 3′SS) for an intron length of 212 nt in total. We inserted these introns into the UbC promoter and 5′UTR in place of the endogenous UbC first intron, upstream of the EGFP coding sequence. Each intron was randomly paired with a random $H_{18}$ barcode (18 nt comprising A, C or T, but no G) located in the 5′UTR downstream of the intron. Locating the barcode in a non-coding exonic region of the transcript enables recovery of the intron sequence from RNA sequencing reads of the spliced mRNA, and the exclusion of Gs precludes the formation of alternative splice sites as well as upstream AUG start codons, which are known to be a major 5′UTR sequence determinant of translational efficiency and mRNA stability [33,34].

A key feature of our design is that we included a second minigene in the reporter system, the red fluorescent protein dTomato (dTom), roughly 1-kb upstream of the transcription start site of EGFP (Fig. 2A). We constructed the library so that the dTom and EGFP 5′UTRs in each plasmid bear identical $H_{18}$ barcodes, and therefore the ratio of read counts originating from EGFP versus dTomato for a particular barcode serves as a measure of the associated intron's impact on EGFP expression. This design controls for: (i) the abundance of the plasmid in the library; (ii) the transcriptional activity of the reporter cassette in the particular cell(s) that integrated it; and (iii) any impact of the barcode on mRNA or protein expression from the transcript.

Accurate duplication of randomly generated barcodes in a pooled fashion had not been accomplished previously to our knowledge. We ultimately achieved this through circular ligation of the intron-barcode pairs, followed by rolling circle amplification, fragmentation, PCR, gel electrophoresis and subsequent size selection for fragments containing two copies of each barcode flanking a single intron (Fig. 2B; Materials and methods). We optimized this procedure to reduce barcode mismatching from PCR-induced chimerism [35] and generated a highly diverse vector library containing over 300 000 unique introns detectable by short-read sequencing of plasmid DNA. We proceeded to investigate IME using both this full library as well as a bottlenecked (sub-sampled) library ∼10% of the original size, in order to increase the read depth associated with each individual intron.

We assayed the reduced plasmid library via Oxford Nanopore long-read sequencing to construct the barcode-to-intron map and to validate the integrity of dTom-GFP barcode matching. For each intron, we analyzed the set of associated barcodes and excluded promiscuous barcodes which could not be confidently assigned to a single intron (Fig. 2C). This procedure yielded a dictionary of 49 737 barcode-intron pairs, whose intron sequences were supported by clusters of 116 reads on average (median 78). We additionally confirmed whether, in each read, the dTom barcode matched the GFP barcode as intended. We expected some degree of mis-pairing to have occurred in the library cloning procedure and sought to restrict our analysis to only those barcodes that were correctly paired. We found that 38 158 GFP barcodes (76.7%) were paired with matching dTom barcode in ≥90% of plasmid reads (Fig. 2D). All further analyses focused on this set of confident barcodes and introns only.

### Short random introns are efficiently spliced in human cells

We co-transfected this plasmid library with a Cre recombinase vector into the HILO-RMCE HEK293T A2 cell line [30] in order to create a pooled cellular library with single-copy genomic integration of the random-intron-containing EGFP. As a readout of intron effects on steady-state RNA level, we extracted total RNA from these cells and sequenced an RT-PCR amplicon of the GFP and dTom 5′UTRs. In each sample, this experiment yields three types of reads per barcode: dTom reads, spliced GFP reads, and unspliced GFP reads (Fig. 3A, Supplementary Table S5). In total, we performed ten biological replicates of RNA extraction from the cellular library for

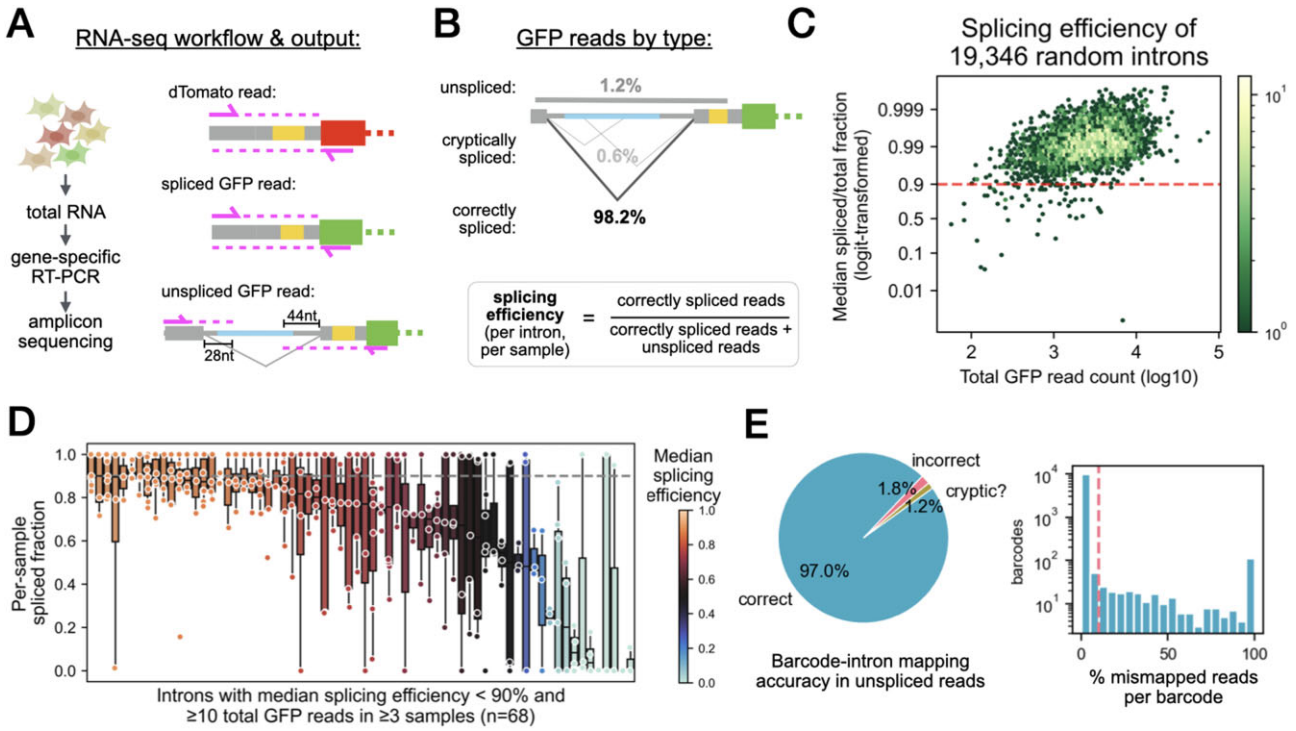



**Figure 3.** Almost all short random introns splice efficiently. (**A**) Schematic of RNA-seq workflow and resulting read types. (**B**) Of all the GFP reads detected, the vast majority are spliced as expected between the intended library splice sites. Splicing efficiency calculations per intron, per sample take into account correctly spliced reads and unspliced reads that look as expected, disregarding reads of unknown splicing status. (**C**) Median splicing efficiencies of all introns in the random library, across all samples in which that intron has at least 10 GFP reads, as a function of total GFP read count for that intron. (**D**) Per-sample splicing efficiencies of introns with median splicing efficiency <0.9. Boxplot centers represent medians, box edges represent interquartile range (IQR) or middle 50% of data, whiskers extend to 1.5x IQR past each box edge. (**E**) Fraction of barcodes for which intron random regions in unspliced reads match the expected sequence. To be classified as "correct," the intron must match in ≥90% of reads with the associated barcode.

amplicon sequencing, five with the full plasmid library and five with the bottlenecked version.

It was essential to determine whether random introns would be efficiently spliced in these cells, despite their internal sequence being random and thus lacking natural splicing regulatory elements other than the core 5′SS, branch point and polypyrimidine tract/3′SS. Across the ten replicates of RNA-seq, over 98% of GFP reads were spliced as expected, using the constant 5′SS and 3′SS present in all introns (Fig. 3B). Unspliced reads were associated with 10 507 introns, but usually accounted for a small fraction of the GFP reads from any individual intron. Of 19 346 introns that met our minimum read count threshold for analysis, requiring at least 10 GFP and 10 dTom reads in at least three samples, the vast majority (99.7%) were well-spliced, with median splicing efficiency 0.9 or higher across all samples with minimum 10 GFP reads (Fig. 3C). Only 68 introns had median splicing efficiency below 0.9, and the majority of these were still above 0.8 (Fig. 3D, Supplementary Table S4). Because the PCR step in our amplicon-sequencing approach likely favors shorter (spliced) products over unspliced, we do not report percent spliced in (PSI) values here; however, when directly compared to a natural intron control which is known to be well spliced (UbC), the random introns exhibited similar observed splicing efficiencies (Supplementary Fig. S2F). These observations suggest that presence of strong splice sites and a branch point, as in our reporter, may be sufficient for splicing of short (~200 nt) introns, more or less independently of the intervening sequence [36].

The unspliced reads also serve a valuable quality control role for this system: in the latter five of the ten RNA-seq replicates performed, our sequencing scheme read 28 nt past the fixed 5′SS and 44 nt upstream of the fixed 3′SS in each transcript (Fig. 3A). Accounting for constant regions, this yields 16 and 4 nt, respectively, from the random region of the intron on either end, allowing us to verify that the intron sequence is that expected from the barcode in the read. In a first pass analysis, just over 97% of barcodes observed in unspliced reads (10 194 out of 10 507) correctly predicted the first 16 nt of intron sequence with ≥90% accuracy (Fig. 3E), confirming high accuracy of barcode-intron mapping in our system. Of the remaining 3% of "mismapped" reads, we noticed that most appeared to be truly unspliced and mismapped, while a subset aligned partially to the correct intron sequence and appear to be spliced to a cryptic 5′SS or 3′SS created by chance in the random region of the intron. These reads were therefore reclassified as cryptically spliced rather than mismapped, increasing our estimate of the accuracy of barcode-intron mapping to almost 98%.

## Usage of cryptic 5′ and 3′SSs

To explore the distribution of cryptic splicing in the random library, we investigated the set of ~250K GFP reads that were neither unspliced nor canonically spliced. The constant 5′ and 3′SS are relatively strong, with MaxEnt scores [37] of 10.7 and 13.0 bits, respectively (Fig. 4A). Assuming cryptic splicing from one of these SS to an internal cryptic site, we reasoned

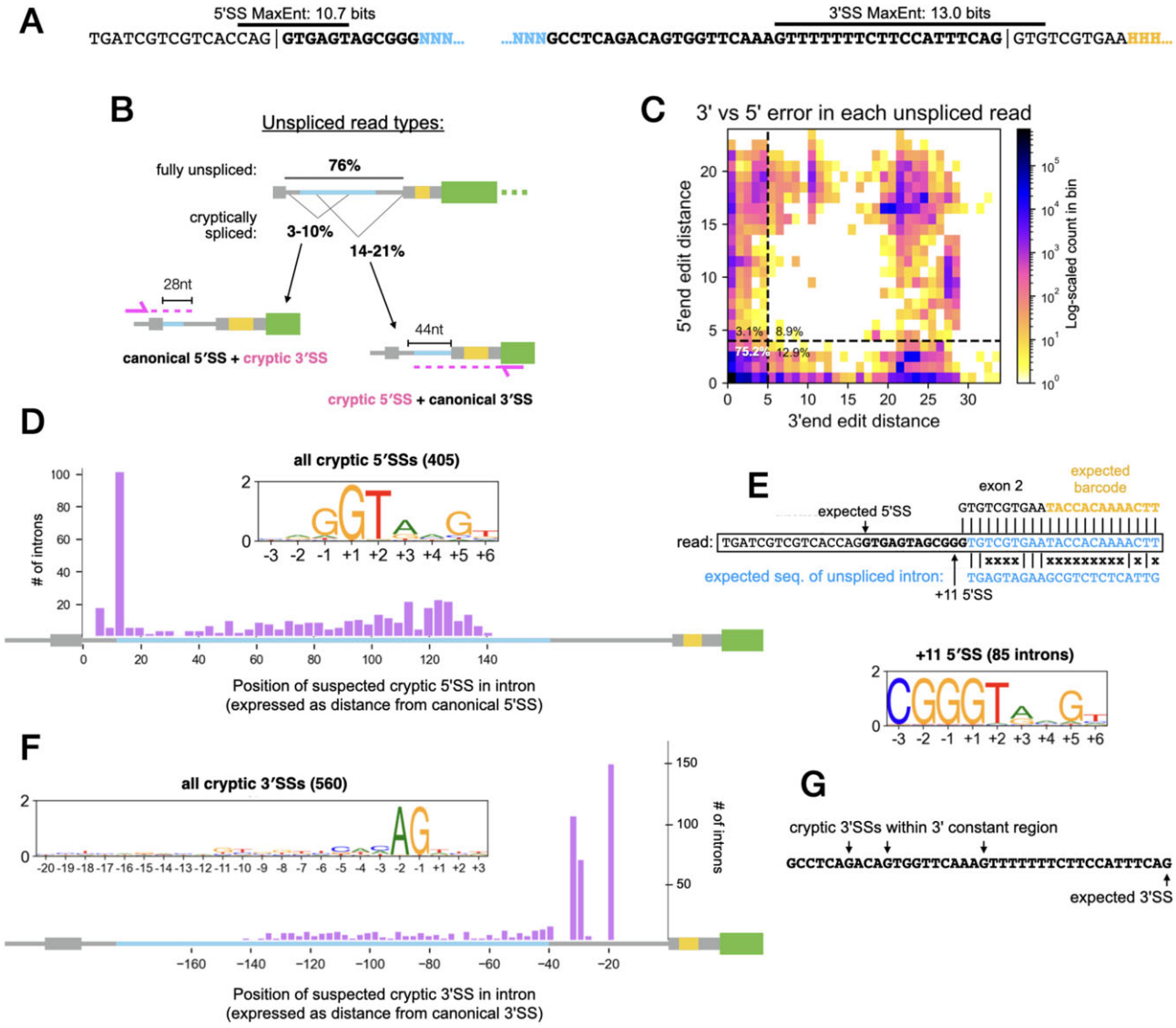



**Figure 4.** Cryptic splicing in random introns. (**A**) Splice sites used in random library, annotated with respective MaxEnt scores. Intron-exon boundaries denoted by |, constant 5′ and 3′ SS sequences flanking random region in bold, intron random nucleotides in blue, barcode random nucleotides in yellow. (**B**) Schematic of unspliced GFP read types, i.e. completely unspliced versus 5′ or 3′ end cryptic splicing. (**C**) 2D histogram of edit distance from expected sequence at 5′-end and 3′-end of all unspliced reads. Percentage of reads in each quadrant are annotated at inside corners. (**D**) Distribution of locations within intron where evidence is seen for use of cryptic 5′SSs. Inset: sequence logo of nucleotides [-3,+9] at these positions. (**E**) Example of an unspliced read inferred to have used the + 11G as a cryptic 5′SS, and sequence logo for the subset of 5′ cryptic splicing events mapping to this position. (**F**) As in D, for cryptic 3′SS. (**G**) Constant sequence at 3′end of random library with positions of suspected cryptic 3′ splicing indicated. See also Supplementary Fig. S1.

that we should be able to infer the exact location of the cryptic site from the read sequence. Splicing to an internal cryptic 5′SS would generate reads that included exon 1 and a portion of the 5′ end of the intron, followed by exon 2, and vice versa for splicing to an internal cryptic 3′SS (Fig. 4B). Reads with unexpected sequence at both ends were largely derived from cryptic sites in the intron constant regions, i.e. close enough to each end to be captured in both reads (Supplementary Fig. S1A). Altogether, about 1 in 4 unspliced reads had some unexpected sequence at either or both ends (Fig. 4C).

By aligning the 3′ end of each of these reads with the expected intron sequence, we were able to identify over 400 likely cryptic 5′SS (Fig. 4D, Supplementary Fig. S1B, Supplementary Table S7). Overall, aside from one exceptional position, the frequency of cryptic 5′SS increased with distance from the canonical 5′SS up to 150 nt, and the inferred 5′SS motifs resembled the canonical human 5′SS motif, as expected (Fig. 4D). The exception was the proximal +11G position in the 12 nt of 5′ constant sequence, which was a hotspot of cryptic splicing, with 85 introns using the +11G in at least 10 reads. This site is likely preferred because it already has consensus bases at the –3, –1, and +1 positions of the 5′SS motif for a 5′SS at +11 in the intron (Fig. 4E). The absence of cryptic 5′SS past 150 nt implies a minimum intron length of ~70 nt, consistent with previous studies in mammals [38].

Repeating this procedure with reads that had matching to the 3′ end of their intron, we identified instances of introns using internal cryptic 3′SS (Fig. 4F, Supplementary Fig. S1C, Supplementary Table S8). These, too, were distributed throughout the intron no further than ~150-nt upstream of

the canonical 3′SS, again implying a minimum intron length of ∼70 nt. These sites were enriched for positions within the constant region that already matched the minimal 3′SS consensus NAG (Fig. 4F and G). Curiously, we observed a negative correlation between the scores of cryptic splice sites and their positions within the random region of the intron, with weaker cryptic 3′SS and 5′SS motifs being observed only near the 3′ end of the intron (Supplementary Fig. S1D and E), even when splice sites overlapping constant regions were excluded. This suggests that the constant regions at the 3′ end of the intron and the 5′ end of exon 2 may represent a particularly strong context for splicing.

## Random introns enhance gene expression at mRNA and protein levels

Having confirmed that the introns in our library predominantly splice as expected, we next examined the GFP and dTom read counts per barcode to study each intron's effect on GFP mRNA expression. For this analysis we used all ten replicates of amplicon sequencing, across which the mean GFP and dTom read counts of the different barcodes varied over 5 orders of magnitude (Fig. 5A). Negative (intronless) and positive (UbC intron) controls were spiked into the random intron plasmid library at a 1:5000 mass ratio prior to transfection and, as expected, yielded total read counts roughly ten-fold higher than the library average. Batch variation in the global average ratio of GFP to dTom reads was observed, corresponding to the two sets of replicates performed pre- and post-bottleneck (Supplementary Fig. S2A). To control for these batch effects and remove noisier (lower-count) barcodes, we applied a minimum read count filter and normalized read counts across replicates using DESeq2 [31]. We reasoned that the question of intron effects on GFP expression relative to dTom can be treated as a differential expression analysis, where GFP and dTomato counts from the same replicate are handled as paired "treatment" and "control" samples, and we model the variation across replicates to determine which "genes" (distinct introns in our analysis) are significantly altered in the treatment versus control.

The filtered and normalized read count data suggested that the presence of an intron alone, independently of its sequence, stimulates GFP mRNA expression (Fig. 5B, Supplementary Table S4). Since the GFP and dTomato counts are treated as separate libraries during DESeq2 modeling and shrinkage, their mean ratio across all barcodes is set to zero during normalization. To estimate the magnitude of each intron's effect on expression compared to an intronless version of the same gene, we subtracted the mean log ratio of the four intronless controls, yielding a library average $\log_2$ fold change of ∼3.1, or an absolute fold increase in GFP mRNA of ∼8.5-fold compared to intronless GFP. To validate this observation, we also measured the difference in GFP/dTom ratio by qRT-PCR of RNA extracted from a pool of cells with the whole library integrated compared to cells with integration of the intronless plasmid, yielding an estimated average enhancement of five-fold (Supplementary Fig. S2B). These two approaches for measuring the aggregate IME of the library have distinct strengths and weaknesses, as discussed below. Nearly all intron-containing reporters exhibited a higher GFP/dTom ratio than the four intronless reporters, indicating that considerable enhancement of expression is a general property of splicing in this context.

Some variation in enhancement levels occurred within the library: 244 out of 19 346 introns analyzed had significantly stronger IME than other introns, while 65 had significantly weaker IME, including three of the four intronless control barcodes (Fig. 5B). Prior to application of DESeq2, we observed that GFP/dTom read count ratios were not always well correlated between replicates: specifically, replicates from separate transfections conducted in parallel had minimal correlation, while replicates originating from the same transfection were highly correlated (Supplementary Fig. S2C). Nevertheless, the intronless control barcodes consistently yielded significantly lower GFP/dTom read count ratios than the rest of the library, across all replicates, whether between or within transfections. This observation implies that our setup can reveal reproducible variation in IME as long as it is of sufficient magnitude.

To estimate the range of IME values represented by the data, we performed various simulations of our library experiments. These simulations modeled both the sampling noise in estimation of individual IME values from RNA-seq and the variability resulting from different transfections. This includes any experimental variability in the precise transfection and subsequent antibiotic selection conditions, as well as epigenetic/cell state differences between cells that integrate a particular intron across different transfections (Supplementary Note 1). A simulation which visually recapitulated the spread and density of points in the observed scatter plots was obtained (Supplementary Fig. S2D). With this simulation, the correlations observed between replicates matched best to simulated values with a total relative IME score log ratio range of 0.1–0.3 (Supplementary Fig. S2E). This is equivalent to ∼95% of introns in the library conferring $\log_2$ enhancement between 2.7 and 3.5 (i.e. IME between 6.5-fold and 11-fold). Taken together, these data suggest that virtually any well-spliced intron can confer considerable IME, independent of sequence.

To assess whether this effect also manifested at the level of protein expression, we used flow cytometry to measure the GFP and dTomato fluorescence in the pooled cellular library, alongside our UbC intron and intronless control lines, and negative control cells expressing neither protein (Fig. 5C). The random library and the UbC intron control both displayed similar dTom expression and significantly higher GFP expression than the intronless control lines. We observed a small population of dTom-negative cells in the library and verified that these represent carryover of a cloning intermediate; these barcodes are excluded from subsequent analyses via removal from the barcode-to-intron dictionary, and from this analysis by considering only dTom + cells (Fig. 5C, red dotted line) in calculating the average expression ratio of the library.

We computed the per-cell ratio of GFP to dTom fluorescence after subtracting autofluorescence (median intensity of negative control cells) and compared the distribution of ratios in the library to that of ratios in the intronless control cell lines. Consistent with the RNA-seq-derived estimates, we found that cells with intron-containing GFP produce on average 8.9-fold more GFP than intronless cells when each is normalized to its paired dTom measurement (Fig. 5D). In both cases, the UbC intron falls close to or just above the mean of the random introns. Though the distribution of mRNA-level effects appears highly concentrated compared to the protein-level measurements, this difference in spread likely reflects differences in measurement noise and data processing (e.g. the

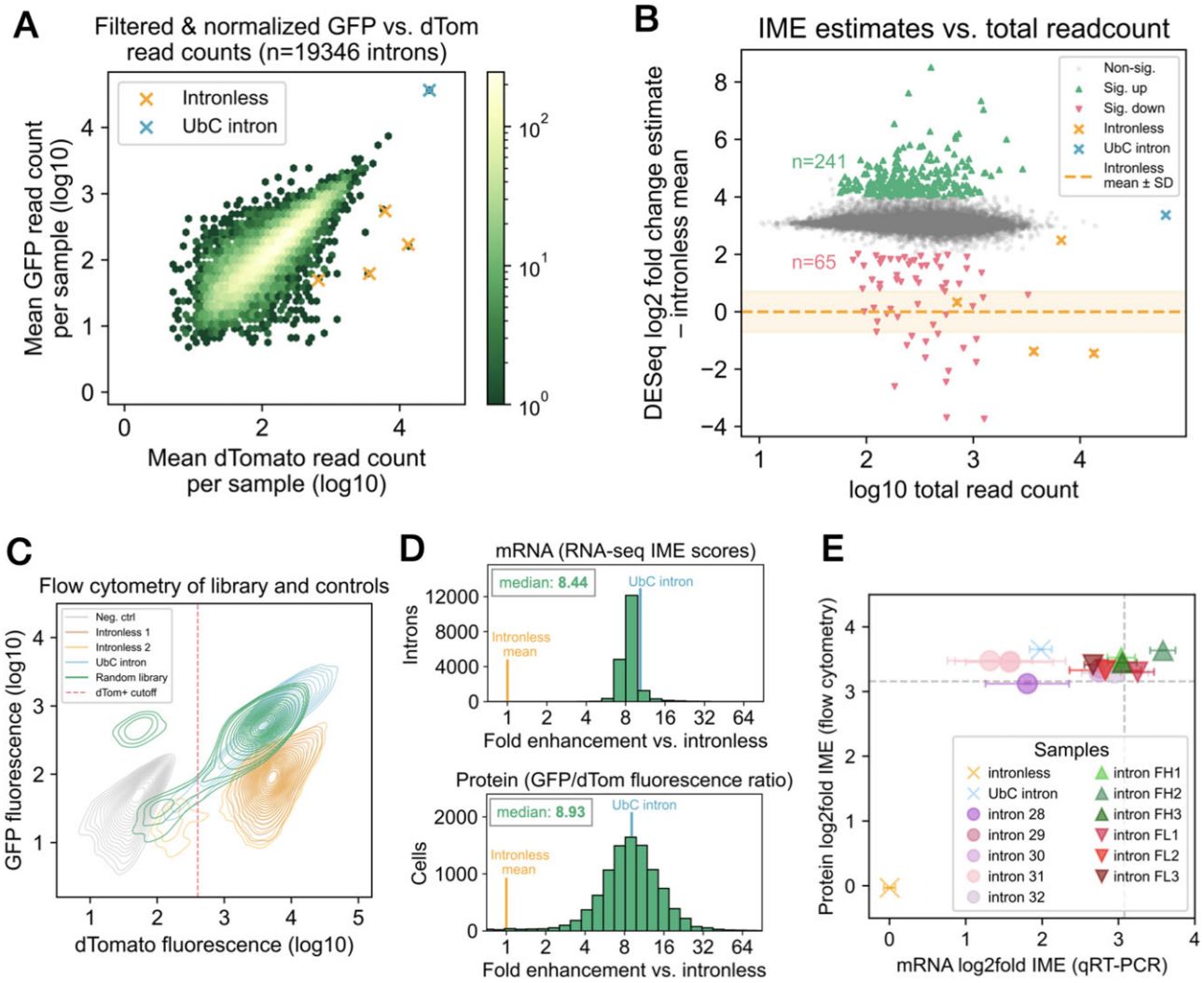



**Figure 5.** Random introns enhance GFP mRNA and protein expression. (**A**) GFP and dTomato RNA-seq read counts per intron, mean across ten samples. (**B**) DESeq2 $\log_2$ fold change estimates versus total read count for each intron. Estimates are normalized by subtracting the mean $\log_2$ fold change of the four intronless control barcodes. Introns called significantly different from the mean of the library are highlighted ($P \leq 0.05$, absolute $\log_2$ fold change $\geq 1$). (**C**) Flow cytometry of cells with integrated library overlayed with various controls. For calculation in D, library cells not expressing dTomato (left of dotted line) were excluded. (**D**) Fold enhancement at mRNA and protein level relative to intronless are compared on the same scale. Steady-state GFP mRNA and protein levels are each ~eight-fold higher with a short random intron than with no intron. (**E**) qRT-PCR and flow cytometry of transgenic cell lines expressing GFP with individual selected introns for validation. Error bars denote standard error of the mean. See also Supplementary Fig. S2.

shrinkage and normalization performed at the mRNA level) rather than biology. Furthermore, the flow cytometry experiment does not assess which intron is in which cell, only capturing the bulk population, so the two distributions are not directly comparable.

We sought to validate these high-throughput measurements by cloning and testing the IME activity of individual random introns from the library. We selected 11 introns and constructed 11 transgenic lines, each harboring GFP containing one of these introns and an intronless dTom, without barcodes. We assayed their mRNA- and protein-level enhancement by qRT-PCR and flow cytometry and compared it to the expected values from the high-throughput measurements as well as the intronless and UbC control lines (Fig. 5E). We selected introns with a range of predicted IME strengths, and observed a range of mRNA-level IME values from 2.5- to 12-fold, similar to the ranges predicted by DESeq2 and by our simulations. For a subset of these introns, efficient splicing was confirmed using RT-PCR (Supplementary Fig. S2F). For one specific intron, we conducted three separate transfections to produce three independent cell lines, verifying that the measured IME value is reproducible (Intron 31, Fig. 5E). Discrepancies in the relative ranking of introns measured by qRT-PCR versus RNA-seq suggest the presence of biological variability between the pooled and individual transgenic lines.

## Selection of introns with stronger and weaker IME by iterative FACS

In order to identify the specific random introns with stronger and weaker protein-level IME, we devised an iterative FACS procedure to enrich for introns with extreme enhancement values. We collected cells from the library with the highest and lowest GFP/dTom fluorescence ratios by empirically slicing the top and bottom 10% of cells along the $y = x$ diagonal (Fig. 6A, Supplementary Fig. S3A). After each sort, we reserved some sorted cells for RNA extraction and sequencing, and cultured the remainder until re-sorting several days later. Three successive sorts in two independent trajectories led to smaller and smaller sets of barcodes being captured, with fluorescence

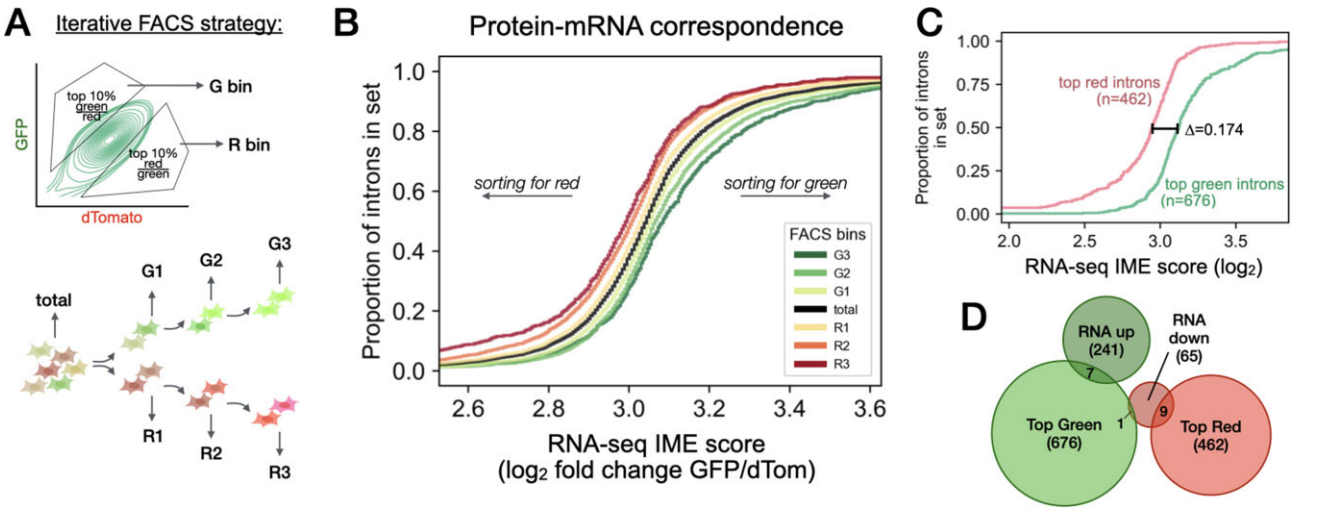



**Figure 6.** Iterative enrichment of cells with extreme GFP/dTomato fluorescence ratios allows study of protein-level IME. (**A**) Schematic of iterative FACS experimental design. (**B**) CDFs of mRNA-level IME scores (DESeq l2fc estimates) for the set of introns detected at each sort stage. Shown are the union of both replicate trajectories, using 100 reads as cutoff for detection. (**C**) Designated sets of introns with strongest and weakest protein-level IME for downstream analysis. Each set comprises introns detected at all stages of one color and none of the other color. The difference between the median log2 IME scores of the two sets is 0.174, or ∼12%. (**D**) Overlaps of iterative FACS-derived sets with RNA-seq-derived sets (DESeq significant introns). See also supplementary Fig. S3.

intensity distributions shifted higher or lower, as expected (Supplementary Fig. S3B and C, Supplementary Table S6).

We next explored the relationship between the measurements of RNA and protein: specifically, whether introns in these FACS bins were shifted in their mRNA-level IME estimates as well. Taking the union of all introns seen in a given sort stage (requiring at least 100 reads in either replicate trajectory), we observed that red sorts (R1, R2, R3) had progressively lower and green sorts (G1, G2, G3) had progressively higher mRNA-level IME distributions, indicating consistency between RNA and protein measurements (Fig. 6B). This pattern was independently true for each replicate, and for the intersection of introns detected in both replicates (Supplementary Fig. S3D). These observations provide orthogonal support for the integrity of our RNA-seq measurements, and for our conclusion that different random introns have distinct effects on gene expression. They further suggest that effects on mRNA levels drive most effects observed at the protein level.

For use in further analysis, we defined a set of "top green introns" as those detected (≥100 reads) in either replicate of all three green sorting stages, and not detected in any red sort; we likewise defined a set of "top red introns" with the opposite requirements. This yielded sets of 676 top green and 462 top red introns, which were well separated in their distributions of RNA-seq-derived IME scores (Fig. 6C, Supplementary Table S4). Relatively low overlap was observed with the corresponding significant up or down sets from the DESeq analysis (Fig. 6D), suggesting some unknown source of variability in one or both experimental approaches, at least for introns with more extreme IME values.

## PolyU motifs are enriched in highly enhancing introns

Having sets of relatively stronger and weaker introns in hand, at both the mRNA and protein levels, we sought to identify features enriched in the strongest-enhancing introns. We reasoned that differences in enhancement could be attributable to differences in splicing efficiency, in RNA secondary structure, or in primary sequence, and investigated each possibility accordingly.

First, using a conservative approach that normalizes the GFP:dTom ratio to the proportion spliced, we observed a small but significant positive correlation between splicing efficiency and IME (Fig. 7A, Spearman's rho = 0.072, $P = 2.2e\text{-}18$; see also Supplementary Note 2): introns with higher proportions of unspliced reads were less likely to be strongly enhancing. However, most introns in the library were efficiently spliced, so other sources of IME variation are clearly present.

We next explored whether RNA secondary structures were enriched or depleted in introns with higher/lower IME using RNAfold [39, 40]. However, no significant patterns were observed, except for a slight bias against secondary structure downstream of the barcode, in the primer binding region for the RNA-seq amplicon, likely a purely technical effect.

We then interrogated the sequence content of highly versus lowly enhancing introns, considering that short RNA motifs often have biological activity and could conceivably mediate increased IME through, e.g. recruiting *trans*-acting factors to the pre-mRNA. We computed the frequencies of 4-mers and 5-mers in each intron in the library and compared the average frequency of each *k*-mer in the significantly up sets to the corresponding significantly down sets, at mRNA and protein levels (Fig. 7B). We noticed that U-rich *k*-mers were enriched in introns with stronger IME activity, while C-rich *k*-mers tended to be depleted. In particular, polyU motifs (stretches of three or more consecutive Us) were enriched in the introns with the strongest IME from both the FACS and RNA-seq analyses, and the polyU count per intron was slightly but significantly correlated with RNA-seq IME score across the whole library (Supplementary Fig. S4A, Spearman's rho = 0.055, $p = 2e\text{-}14$, for $U_3$).

Comparing the distribution of mRNA-level IME scores between the set of introns lacking polyU and those containing one or more polyU motifs (Supplementary Fig. S4B), we

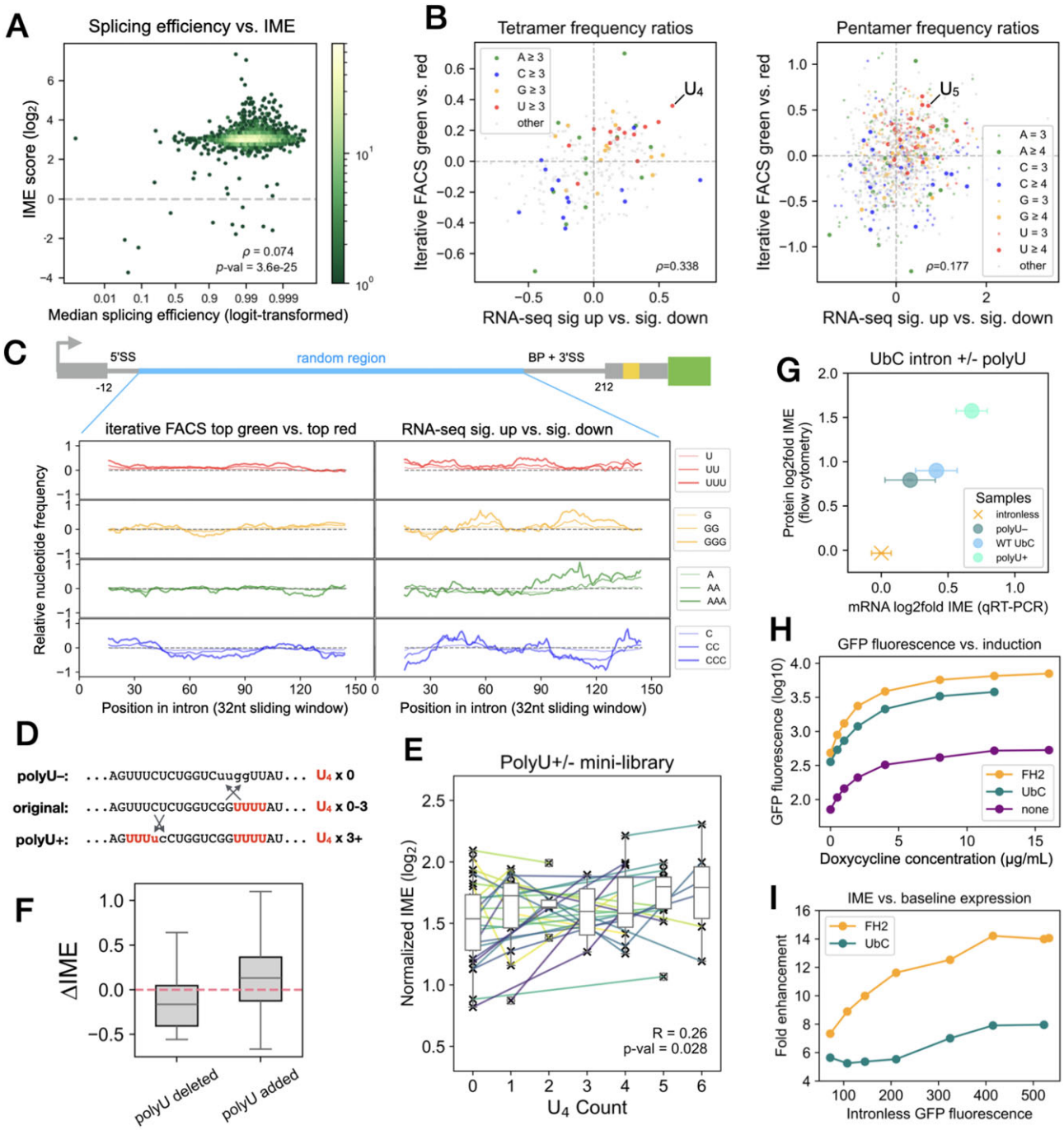



**Figure 7.** PolyU motifs are enriched in highly-enhancing introns. (**A**) Splicing efficiency is weakly but significantly negatively correlated with IME strength. (**B**) *K*-mer frequency ratios between highly enhancing and lowly enhancing intron sets for both iterative FACS and RNA-seq. Each dot is coloured according to the nucleotide composition of the *k*-mer and polyU is indicated on both plots (tetramer UUUU, pentamer UUUUU). (**C**) Metaplots showing enrichment of single, di- and tri-nucleotide runs in highly versus lowly enhancing introns, for both iterative FACS and RNA-seq, as a function of position in the intron random region. (**D**) Permutation strategy for polyU ± mini-library. (**E**) IME of introns in the second library is significantly correlated with polyU count. Lines indicate permuted versions of the same intron. Boxplot centers represent medians, box edges represent interquartile range (IQR) or middle 50% of data, whiskers extend to 1.5x IQR past each box edge. (**F**) Distributions of differences in IME after permutation. (**G**) qRT-PCR and flow cytometry of UbC intron polyU series reporters compared to intronless controls. Error bars denote standard error of the mean. (**H**) GFP protein expression from a doxycycline-inducible promoter with 5′UTR intron indicated in legend. (**I**) IME (GFP normalized to dTomato, normalized to intronless) as a function of transcription induction. See also Supplementary Fig. S4.

observed that for all three motif lengths considered ($U_3$, $U_4$, or $U_5$), increased polyU count is generally associated with increased IME (Wilcoxon's rank-sum test, $p < 0.05$). The differences in the medians of these distributions, taken as a proxy for effect size, are on the order of 0.006–0.018 on a $\log_2$ scale, or 0.4–1.3% change in observed enhancement. We also broadened this analysis to ask, for all 4-mer and 5-mer motifs, which have significant differences in IME when comparing introns containing them versus introns without them (Supplementary Fig. S4C). While U-rich *k*-mers again emerged as the strongest differentiators, we also observed that some motifs are associated with a significant decrease in expression—for example the pentamer CGTCA—and that these tended to be either C-rich or A-rich motifs or complex motifs containing three or four distinct nucleotides.

Meta-intron plots of nucleotide enrichment in highly versus lowly enhancing sets also indicate that U-richness is correlated with enhancement (Fig. 7C). In particular, mono-, di- and tri-nucleotide runs of Us are slightly but consistently more frequent across the entire random region of introns with stronger IME, both by iterative FACS and RNA-seq measurements. Though there are local regions of enrichment of other nucleotides—in particular G and A—in the RNA-seq assessment, no base other than U is globally enriched across all positions. Enrichments between iterative FACS bins had generally lower magnitude than those from RNA-seq, perhaps due to differences in the criteria used to define up- and down-regulated sets.

Based on these observations, we designed and queried a second, smaller intron library to directly test the influence of polyU content on IME. We selected 30 random introns from the original library and for each one made "polyU+" and "polyU–" variants, in which we minimally permuted the intron sequence to either bring together or disperse Us, querying 90 sequences in total containing between 0 and 6 instances of $U_4$ (Fig. 7D, Supplementary Table S9). We sequenced RNA from cells transiently transfected with this pool and found that the IME from these introns was significantly correlated with polyU count (Fig. 7E, Pearson's $R = 0.26$, $P = 0.028$). Though neither distribution was significantly different from zero in a one-sample T-test, deleting $U_4$s generally led to a slight drop in enhancement, while adding $U_4$s led to an increase in enhancement (Fig. 7F).

To further assess the polyU effect in a natural intron context using individual reporter lines, we also created shuffled versions of the human UbC intron sequence. This 812 nt intron natively contains 9 $U_4$s (including overlapping motifs, i.e. counting $U_5$ as two $U_4$s), with our polyU– and polyU + versions containing 0 and 24 instances of $U_4$, respectively. These reporters were integrated into HEK293T cells alongside a matching intronless reporter and observed modest but consistent differences in the IME of these introns (Fig. 7G). The polyU + version of the UbC intron was stronger in enhancing GFP expression at both the mRNA and the protein level than the original intron, which was in turn stronger than the polyU– intron, confirming that intron polyU content positively modulates IME in a natural sequence context as well as in short random introns.

We noted that the magnitude of fold enhancement versus intronless as well as the baseline expression from this plasmid was lower than for typical introns in the screen, though the plasmid differed in only minor ways from the original backbone (Supplementary Fig. S4D). Re-cloning these introns into the exact vector used for the screen recovered the magnitude of IME previously observed, but the differences between the polyU+, original and polyU– introns were less pronounced (Supplementary Fig. S4E). This difference between the vectors suggests that flanking sequence context and/or basal transcriptional output may impact the magnitude of IME and its modulators.

### Other sequence and contextual modifiers of IME

Intrigued by the observation that minor differences in the plasmid backbone altered the IME of introns in our reporter experiments, we decided to explore the effect of promoter strength as a context feature influencing IME. To this end we constructed and integrated one intronless and two intron-containing reporters driven by doxycycline-inducible promoters, and measured the enhancement by each intron compared to the intronless vector at various levels of induction (Fig. 7H). Surprisingly, higher levels of dox induction yielded stronger relative enhancement from both introns (Fig. 7I). Perhaps, in a given genomic context, higher rates of transcription promote more efficient splicing [41], which promotes more efficient IME, creating a positive feedback loop between transcription and splicing. This observation also suggests that features of the cellular context (e.g. growth rate, cell cycle phase, etc.) that impact the basal level of transcription of a locus may also impact the level of IME that occurs at that locus, suggesting that IME may act as a magnifier of transcriptionally-driven gene expression programs.

## Discussion

IME has been recognized for decades, yet our understanding of why different introns can exert different effects in the same context is limited. Here we developed an approach to address this question at an unprecedented scale. Screening tens of thousands of introns with a unique reporter design revealed insights into the sequence determinants of IME, and aspects of the splicing of short introns in human cells. In our view, the main takeaways from our work are: (i) the presence of any well-spliced intron exerts enhancement of expression; (ii) that sequence motifs such as polyUs can modulate IME to an extent; (iii) that short introns with strong 5′ and 3′ ends are generally well-spliced, even with random internal sequence; (iv) that expression enhancement is predominantly at the mRNA level; and (v) that the magnitude of IME is sensitive to local gene context. While many animal introns are longer than the tested 212 nt, a substantial proportion of human introns (15.2%) are ≤250 nt in length, so these conclusions may generalize reasonably well to natural introns.

Our aim was to understand sequence-dependent and -independent contributors to IME. To estimate intron-specific IME levels while accounting for the variability of the data and batch effects, we used DESeq2 to normalize the ratio of GFP to dTom expression per barcode across samples. We found that reporters with introns were expressed on average eight-fold higher than intronless controls, or five-fold based on qRT-PCR of the pooled library compared to intronless. A likely contributor to this difference is that the RNA-seq mean is the average of per-intron estimates for introns meeting our stringent analysis criteria, whereas the qRT-PCR mean includes many more introns, potentially including a higher proportion of weaker or less efficiently spliced introns.

Some individual introns showed stronger or weaker IME than average, suggesting that sequence features can modulate the magnitude of IME. These intron-specific effects were reproducible in our iterative FACS experiment, further supporting that they reflect properties of particular intron sequences. Investigating the sequence composition of these introns revealed an enrichment for polyU motifs, and depletion of polyC, in more strongly enhancing introns. PolyU motifs have been reported as enhancing IME in plants [27, 42–44] but not in animals. In plants, U-richness (or AU-richness) is an important determinant of intron splicing, and depletion of Us can impair splicing. However, deletion studies showed that removal of U-rich motifs can decrease IME without inhibiting splicing [44].

Our data support a functional impact of polyU sequences specifically rather than overall U content, as the polyU reporter experiments involve mutated introns where the nucleotides are shuffled to either disperse Us or gather Us into contiguous stretches, without changing base composition, similar to [29]. Many families of human RBPs are known to bind short U-rich sequences [45]. Proteins of the PTB family, for example, are known to bind introns and have been reported to enhance mRNA levels by stimulating 3′ end processing [16, 46]. It would therefore be interesting to explore whether the IME-promoting activity of polyU or other motifs is dependent on the activities of particular RBPs. Ultimately, the sequence contributions we observed were modest compared to the sequence-independent effect of simply having a well-spliced intron.

We were also interested to observe that the vast majority of introns in the library were well-spliced as, *a priori*, it was not clear whether introns with 160 nt of random internal sequence would splice or not. We found that the constant 5′ and 3′ flanking sequences, containing strong SS and a branch point, were sufficient to induce efficient splicing with most internal sequences, as confirmed by RNA-seq and RT-PCR. In rare cases where random introns were not spliced at the expected junction, we could map sites of cryptic splicing internal to the intron. The observed cryptic sites largely matched known 5′ and 3′ SS motifs and obeyed known constraints on minimum intron length in mammals.

Still, the relationship between splicing and IME is complex. Other groups have reported that some IME motifs enhance expression whether located in an intron or exon [47]. Transcriptional enhancers can, of course, enhance from outside a gene or from within an intron, where their own transcription may attenuate expression [48]. We observed a small but significant positive correlation between splicing efficiency and IME from random introns, suggesting that the speed of spliceosome assembly or the speed of progression through the spliceosome cycle may contribute to IME.

Curiously, our experiments with an inducible promoter indicated that higher basal levels of transcription from the same locus yields higher IME. Higher transcriptional activity of a locus is known to promote proximity to nuclear speckles, which in turn promotes more efficient splicing [41], which we observed to enhance IME. In this context, IME can be considered as one component of a positive feedback loop between transcription and splicing.

Notably, in all our observations of IME, GFP protein levels appeared to be enhanced to a similar degree as GFP mRNA. This suggests that IME acts primarily at the mRNA level in the context studied, e.g. via effects on transcription, 3′ end processing, or nuclear export, with protein levels increased as a result of higher cytoplasmic mRNA levels. Though not explored here, the intron library we have constructed could be used to directly interrogate the impact of each intron on each stage of gene expression, via experiments such as cellular fractionation to assess export, or metabolic labeling to measure rates of mRNA synthesis and stability

All our experiments, except for the polyU validation library in Fig. 7E, were performed with reporter plasmids integrated at single copy into the genomes of HEK293T cells. This approach was intended to reduce the large variation in plasmid and mRNA levels that occurs with transient transfection and to study IME in a native chromosomal context. Despite this design, we observed large differences in the magnitude of IME for integration of the same intron between different integrations of the same plasmid, and for the same intron in distinct but highly similar contexts. These observations suggest that IME is highly sensitive to both the flanking genomic context and to aspects of cell state that may differ between replicate transfection/antibiotic selection regimes, or between the individual cells where plasmids integrated. These properties make it harder to define the IME associated with an individual intron, but suggest that systematic investigation of the impacts of genomic and cellular context on IME might be quite fruitful in uncovering additional contributors to this phenomenon.

## Acknowledgements

We thank Craig Hunter, Sergei Ovchinnikov, Peter Reddien, Phillip Sharp, Seychelle Vos, Chris McAndrew, Dima Ter-Ovanesyan, Conor McMann, and members of the Burge lab for helpful discussions.

*Author contributions*: E.J.K.K. conceived the study, designed and performed experiments, analyzed data, and wrote the manuscript. Y.S. performed experiments, analyzed data, and wrote the manuscript. M.P.M. assisted with data analysis. Z.J.P. assisted with reporter experiments. C.B.B. conceived and supervised the study and wrote the manuscript.

## Supplementary data

Supplementary data is available at NAR online.

## Conflict of interest

None declared.

## Funding

This work was supported by National Institutes of Health (NIH) grant 5-R01-HG002439 to C.B.B. Funding to pay the Open Access publication charges for this article was provided by NIH.

## Data availability

The RNA-seq data underlying this article are available in the NCBI GEO Database, under accession code GSE278584. The primary analysis code is available at https://doi.org/10.5281/zenodo.14728464 as well as https://github.com/ejkk0/IME.

## References


1. Koonin EV, Csuros M, Rogozin IB. Whence genes in pieces: reconstruction of the exon-intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. *WIREs RNA* 2013;**4**:93–105. https://doi.org/10.1002/wrna.1143
2. Morales J, Pujar S, Loveland JE *et al.* A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature* 2022;**604**:310–5. https://doi.org/10.1038/s41586-022-04558-8
3. Maniatis T, Reed R. The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature* 1987;**325**:673–8. https://doi.org/10.1038/325673a0
4. Papasaikas P, Valcárcel J. The spliceosome: the ultimate RNA chaperone and sculptor. *Trends Biochem Sci* 2016;**41**:33–45. https://doi.org/10.1016/j.tibs.2015.11.003
5. Irimia M, Roy SW. Origin of spliceosomal introns and alternative splicing. *Cold Spring Harb Perspect Biol* 2014;**6**:a016071. https://doi.org/10.1101/cshperspect.a016071
6. Le Hir H, Izaurralde E, Maquat LE *et al.* The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions. *EMBO J* 2000;**19**:6860–9. https://doi.org/10.1093/emboj/19.24.6860
7. Saulière J, Murigneux V, Wang Z *et al.* CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat Struct Mol Biol* 2012;**19**:1124–31. https://doi.org/10.1038/nsmb.2420
8. Gruss P, Lai CJ, Dhar R *et al.* Splicing as a requirement for biogenesis of functional 16S mRNA of simian virus 40. *Proc Natl Acad Sci USA* 1979;**76**:4317–21. https://doi.org/10.1073/pnas.76.9.4317
9. Hamer DH, Leder P. Splicing and the formation of stable RNA. *Cell* 1979;**18**:1299–302. https://doi.org/10.1016/0092-8674(79)90240-X
10. Shaul O. How introns enhance gene expression. *Int J Biochem Cell Biol* 2017;**91**(Pt B):145–55. https://doi.org/10.1016/j.biocel.2017.06.016
11. Choi T, Huang M, Gorman C *et al.* A generic intron increases gene expression in transgenic mice. *Mol Cell Biol* 1991;**11**:145–55.
12. Mascarenhas D, Mettler IJ, Pierce DA *et al.* Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol Biol* 1990;**15**:913–20. https://doi.org/10.1007/BF00039430
13. Chiou HC, Dabrowski C, Alwine JC. Simian virus 40 late mRNA leader sequences involved in augmenting mRNA accumulation via multiple mechanisms, including increased polyadenylation efficiency. *J Virol* 1991;**65**:6677–85. https://doi.org/10.1128/jvi.65.12.6677-6685.1991
14. Damgaard CK, Kahns S, Lykke-Andersen S *et al.* A 5' splice site enhances the recruitment of basal transcription initiation factors *in vivo*. *Mol Cell* 2008;**29**:271–8. https://doi.org/10.1016/j.molcel.2007.11.035
15. Lu S, Cullen BR. Analysis of the stimulatory effect of splicing on mRNA production and utilization in mammalian cells. *RNA* 2003;**9**:618–30. https://doi.org/10.1261/rna.5260303
16. Millevoi S, Decorsière A, Loulergue C *et al.* A physical and functional link between splicing factors promotes pre-mRNA 3′ end processing. *Nucleic Acids Res* 2009;**37**:4672–83. https://doi.org/10.1093/nar/gkp470
17. Nott A, Meislin SH, Moore MJ. A quantitative analysis of intron effects on mammalian gene expression. *RNA* 2003;**9**:607–17. https://doi.org/10.1261/rna.5250403
18. Valencia P, Dias AP, Reed R. Splicing promotes rapid and efficient mRNA export in mammalian cells. *Proc Natl Acad Sci USA* 2008;**105**:3386–91. https://doi.org/10.1073/pnas.0800250105
19. Zhao C, Hamilton T. Introns regulate the rate of unstable mRNA decay. *J Biol Chem* 2007;**282**:20230–7. https://doi.org/10.1074/jbc.M700180200
20. Le Hir H, Nott A, Moore MJ . 2003; How introns influence and enhance eukaryotic gene expression. https://doi.org/10.1016/S0968-0004(03)00052-5
21. Rose AB. Introns as gene regulators: a brick on the accelerator. *Front Genet* 2018;**9**:215–20. https://doi.org/10.3389/fgene.2018.00672
22. Rose AB. The effect of intron location on intron-mediated enhancement of gene expression in Arabidopsis. *Plant J* 2004;**40**:744–51. https://doi.org/10.1111/j.1365-313X.2004.02247.x
23. Dwyer K, Agarwal N, Gega A *et al.* Proximity to the promoter and terminator regions regulates the transcription enhancement potential of an intron. *Front Mol Biosci* 2021;**8**:712639. https://doi.org/10.3389/fmolb.2021.712639
24. Bartlett JG, Snape JW, Harwood WA. Intron-mediated enhancement as a method for increasing transgene expression levels in barley. *Plant Biotechnol J* 2009;**7**:856–66. https://doi.org/10.1111/j.1467-7652.2009.00448.x
25. Bourdon V, Harvey A, Lonsdale DM. Introns and their positions affect the translational activity of mRNA in plant cells. *EMBO Rep* 2001;**2**:394–8. https://doi.org/10.1093/embo-reports/kve090
26. Gallegos JE, Rose AB. The enduring mystery of intron-mediated enhancement. *Plant Sci Int J Exp Plant Biol* 2015;**237**:8–15.
27. Rose AB. Requirements for intron-mediated enhancement of gene expression in Arabidopsis. *RNA* 2002;**8**:1444–53. https://doi.org/10.1017/S1355838202020551
28. Yuan L, Janes L, Beeler D *et al.* Role of RNA splicing in mediating lineage-specific expression of the von Willebrand factor gene in the endothelium. *Blood* 2013;**121**:4404–12. https://doi.org/10.1182/blood-2012-12-473785
29. Rose AB, Carter A, Korf I *et al.* Intron sequences that stimulate gene expression in Arabidopsis. *Plant Mol Biol* 2016;**92**:337–46. https://doi.org/10.1007/s11103-016-0516-1
30. Khandelia P, Yap K, Makeyev EV. Streamlined platform for short hairpin RNA interference and transgenesis in cultured mammalian cells. *Proc Natl Acad Sci USA* 2011;**108**:12799–804. https://doi.org/10.1073/pnas.1103532108
31. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550. https://doi.org/10.1186/s13059-014-0550-8
32. Stephens M. False discovery rates: a new deal. *Biostat Oxf Engl* 2017;**18**:275–94.
33. Hurt JA, Robertson AD, Burge CB. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* 2013;**23**:1636–50. https://doi.org/10.1101/gr.157354.113
34. Sample PJ, Wang B, Reid DW *et al.* Human 5′ UTR design and variant effect prediction from a massively parallel translation assay. *Nat Biotechnol* 2019;**37**:803–9. https://doi.org/10.1038/s41587-019-0164-5
35. Hegde M, Strand C, Hanna RE *et al.* Uncoupling of sgRNAs from their associated barcodes during PCR amplification of combinatorial CRISPR screens. *PLoS One* 2018;**13**:e0197547. https://doi.org/10.1371/journal.pone.0197547
36. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *WIREs RNA* 2013;**4**:49–60. https://doi.org/10.1002/wrna.1140
37. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004;**11**:377–94. https://doi.org/10.1089/1066527041410418
38. Wieringa B, Hofer E, Weissmann C. A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. *Cell* 1984;**37**:915–25. https://doi.org/10.1016/0092-8674(84)90426-4
39. Hofacker IL, Fontana W, Stadler PF *et al.* Fast folding and comparison of RNA secondary structures. *Monatsh Chem* 1994;**125**:167–88. https://doi.org/10.1007/BF00818163
40. Lorenz R, Bernhart SH, Höner zu Siederdissen C *et al.* ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;**6**:26. https://doi.org/10.1186/1748-7188-6-26

41. Ding F, Elowitz MB. Constitutive splicing and economies of scale in gene expression. *Nat Struct Mol Biol* 2019;**26**:424–32. https://doi.org/10.1038/s41594-019-0226-x
42. Luehrsen KR, Walbot V. Addition of A- and U-rich sequence increases the splicing efficiency of a deleted form of a maize intron. *Plant Mol Biol* 1994;**24**:449–63. https://doi.org/10.1007/BF00024113
43. Rose AB, Beliakoff JA. Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. *Plant Physiol* 2000;**122**:535–42. https://doi.org/10.1104/pp.122.2.535
44. Clancy M, Hannah LC. Splicing of the Maize Sh1 first intron is essential for enhancement of gene expression, and a T-rich motif increases expression without affecting Splicing. *Plant Physiol* 2002;**130**:918–29. https://doi.org/10.1104/pp.008235
45. Dominguez D, Freese P, Alexis MS *et al.* Sequence, structure, and context preferences of Human RNA binding proteins. *Mol Cell* 2018;**70**:854–67. https://doi.org/10.1016/j.molcel.2018.05.001
46. Martinson HG. An active role for splicing in 3′-end formation. *WIREs RNA* 2011;**2**:459–70. https://doi.org/10.1002/wrna.68
47. Cinghu S, Yang P, Kosak JP *et al.* Intragenic enhancers attenuate host gene expression. *Mol Cell* 2017;**68**:104–17. https://doi.org/10.1016/j.molcel.2017.09.010
48. Gallegos JE, Rose AB. An intron-derived motif strongly increases gene expression from transcribed sequences through a splicing independent mechanism in Arabidopsis thaliana. *Sci Rep* 2019;**9**:13777. https://doi.org/10.1038/s41598-019-50389-5

**Received:** October 29, 2024. **Revised:** January 28, 2025. **Editorial Decision:** January 30, 2025. **Accepted:** February 7, 2025