frontiers in Plant Science

ORIGINAL RESEARCH
published: 22 October 2021
doi: 10.3389/fpls.2021.712179

Expression Elements Derived From Plant Sequences Provide Effective Gene Expression Regulation and New Opportunities for Plant Biotechnology Traits

Jennifer P. C. To[1,2,3]*, *Ian W. Davis*[1,2,3], *Matthew S. Marengo*[1,2,3], *Aabid Shariff*[2,3,4], *Catherine Baublite*[1], *Keith Decker*[1], *Rafaelo M. Galvão*[1,2,3], *Zhihuan Gao*[1,2,3], *Olivia Haragutchi*[1,2,3], *Jee W. Jung*[1,2,3,5], *Hong Li*[1], *Brent O'Brien*[1,2,3], *Anagha Sant*[1] and *Tedd D. Elich*[2,3,6]

[1] Bayer Crop Science, Chesterfield, MO, United States, [2] GrassRoots Biotechnology, Durham, NC, United States, [3] Monsanto Company, Research Triangle Park, Durham, NC, United States, [4] Pairwise Plants, Durham, NC, United States, [5] Duke University, Office for Translation and Commercialization, Durham, NC, United States, [6] LifeEDIT Therapeutics, Durham, NC, United States

OPEN ACCESS

Edited by:
Qi Chen,
Kunming University of Science and Technology, China

Reviewed by:
Tobias Jores,
University of Washington, United States
Xiaomin Deng,
Chinese Academy of Tropical Agricultural Sciences, China

***Correspondence:**
Jennifer P. C. To
jenn.to@bayer.com

Specialty section:
This article was submitted to Plant Biotechnology, a section of the journal Frontiers in Plant Science

Received: 20 May 2021
Accepted: 15 September 2021
Published: 22 October 2021

Citation:
To JPC, Davis IW, Marengo MS, Shariff A, Baublite C, Decker K, Galvão RM, Gao Z, Haragutchi O, Jung JW, Li H, O'Brien B, Sant A and Elich TD (2021) Expression Elements Derived From Plant Sequences Provide Effective Gene Expression Regulation and New Opportunities for Plant Biotechnology Traits. Front. Plant Sci. 12:712179. doi: 10.3389/fpls.2021.712179

Plant biotechnology traits provide a means to increase crop yields, manage weeds and pests, and sustainably contribute to addressing the needs of a growing population. One of the key challenges in developing new traits for plant biotechnology is the availability of expression elements for efficacious and predictable transgene regulation. Recent advances in genomics, transcriptomics, and computational tools have enabled the generation of new expression elements in a variety of model organisms. In this study, new expression element sequences were computationally generated for use in crops, starting from native Arabidopsis and maize sequences. These elements include promoters, 5′ untranslated regions (5′ UTRs), introns, and 3′ UTRs. The expression elements were demonstrated to drive effective transgene expression in stably transformed soybean plants across multiple tissues types and developmental stages. The expressed transcripts were characterized to demonstrate the molecular function of these expression elements. The data show that the promoters precisely initiate transcripts, the introns are effectively spliced, and the 3′ UTRs enable predictable processing of transcript 3′ ends. Overall, our results indicate that these new expression elements can recapitulate key functional properties of natural sequences and provide opportunities for optimizing the expression of genes in future plant biotechnology traits.

Keywords: gene expression, plant biotechnology, promoter, intron, 3′ UTR, transcription, expression elements, optimized

INTRODUCTION

Innovations in plant biotechnology have delivered ways to enhance agricultural productivity and sustainability, as well as improve crop quality to meet the farmer and consumer needs (Datta, 2013; Aldemita et al., 2015). With growing demands for productivity and quality for a growing world population, as well as growing pressures from insect pests, weeds, and climate change

on crop production (Tilman et al., 2011; FAO, 2017), new traits with increasing variety are critical to meet these increasing needs (Huang et al., 2015; Ricroch and Hénard-Damave, 2016; Li et al., 2020). Newer plant biotechnology products require trait combinations, also known as trait stacks, to provide multiple trait solutions within one crop (Que et al., 2010; Huang et al., 2015). Sequence redundancy among stacked traits has been identified as a potential risk factor in transgene expression instability (Vaucheret et al., 1998; Kooter et al., 1999; Fagard and Vaucheret, 2000). Hence, diversifying sequences to avoid redundancy, including in expression elements and transgene coding sequences, is important for reducing this risk and maintaining stability and efficacy in plant biotechnology traits. In addition, new biotechnology product concepts may require new modes of expression control to achieve trait efficacy. Altogether, these combined trends lead to an increasing need for diversified and optimized expression solutions.

The availability of efficacious and diverse gene expression elements has been identified as a key bottleneck for developing new biotechnology traits in plants (Que et al., 2010; Nuccio, 2018). For both protein-coding and noncoding transgenes, expression is conferred by a combination of key gene expression elements that are collectively called a gene expression cassette. The gene expression cassette requires the following key components: a promoter, a 5′ untranslated region (UTR), a 3′ UTR, and optionally, one or more introns. The promoter and 5′ UTR enable the assembly of the transcription initiation machinery and recruit RNA polymerase II to the transcription start site (TSS) (Smale and Kadonaga, 2003; Hetzel et al., 2016), thus directing the transcription of the intended coding or noncoding trait gene-of-interest. The 5′ UTR also recruits ribosomes to initiate translation of the coding sequence from transcribed mRNAs (Sonenberg and Hinnebusch, 2009). The 3′ UTR defines the cleavage and polyadenylation of the pre-mRNA, while also contributing to transcriptional initiation and elongation of coding or noncoding sequences (Proudfoot, 2004). The gene expression cassette can also include one or more introns that are transcribed as part of the pre-mRNA and are spliced out during pre-mRNA processing (Lorkovic et al., 2000). Introns can contribute to expression regulation (Le Hir et al., 2003), including increasing expression through a mechanism known as an intron-mediated enhancement (Rose and Beliakoff, 2000; Rose, 2018).

Novel expression elements have been generated for transgenes in plants using a variety of methods, including leveraging cisgenic or transgenic sequences from plants or other species, mutation of such sequences, combinatorial arrangement of fragments or motifs from these sequences, and *de novo* design methods (Venter, 2007; Peremarti et al., 2010; Nuccio, 2018). Outside of directly sourcing native sequences from plants or other species, most novel expression elements described in the literature have been generated from native sequences by using mutational or combinatorial methods (Liu et al., 2014; Rushton, 2016; Grant et al., 2017; Maruyama et al., 2017; Belcher et al., 2020). Recent advances in genomics, combined with machine learning and other computational tools, have offered new opportunities to learn from native genomic sequence datasets for applications in modulating gene expression (Camacho et al., 2018; Gilman et al., 2019; Hollerer et al., 2020). Recent publications have reported *de novo* design of promoters in various model systems, including bacteria and yeast (Kotopka and Smolke, 2020; Wang et al., 2020), and have primarily focused on short sequences, comprising the core promoter. The development of short core promoter sequences in plants, with demonstrated expression in transient reporter systems, has also been reported (Jores et al., 2021).

In this study, we have characterized a set of new expression elements that are computationally derived from native plant sequences. Our promoters include the core promoter and extend upstream to include sequences that confer unique expression profiles, and downstream to include the 5′ UTR. In addition to promoters, we have developed introns and 3′ UTRs and demonstrated *in planta* function for all three classes of elements. We present data from expression characterization of these elements in stably transformed plants, including a detailed analysis of the transcripts produced. Our data demonstrate that these expression elements effectively leverage motifs learned from native sequences to drive reporter gene expression across a variety of plant tissues. With these characteristics, our computationally derived expression elements show promise for delivering increased predictability and tunability for optimizing plant biotechnology traits.

MATERIALS AND METHODS

Plant Sequence Training Set Identification

Expression elements were computationally derived from nucleotide sequences from native genes with desired expression properties (**Figure 1**). First, sets of co-expressed Arabidopsis or maize genes were identified from genome-wide transcriptome data across a variety of tissue types and developmental stages in *Arabidopsis thaliana* (Schmid et al., 2005; Brady et al., 2007) and *Zea mays* (Schnable et al., 2009). Gene expression values for Arabidopsis microarray and maize RNAseq were first calculated for each gene and each tissue by using established methods, gcRMA (Wu et al., 2004) and FPKM (Mortazavi et al., 2008), respectively. Co-expressed gene sets were then identified by unsupervised clustering of calculated gene expression values (Brady et al., 2007). The expression elements characterized in this study were derived from sequences of 102 *Arabidopsis* constitutively expressed genes (At.GSP442, At.GSI17, and At.GSI21), 129 Arabidopsis leaf-preferred genes (At.GSP571 and At.GSP576), and 1,461 maize-expressed genes (Zm.GST7). Sequence training sets from co-expressed native genes were extracted from Arabidopsis TAIR9 (Swarbreck et al., 2008) or Maize B73 RefGen_v1 (Schnable et al., 2009) genome assemblies with adjusted annotations based on EST mapping (Alexandrov et al., 2008; Schnable et al., 2009; Soderlund et al., 2009; Troukhan et al., 2009). Plant native regulatory sequences were extracted from these Arabidopsis and maize gene annotations to collect training sets as follows: promoters and 5′ UTRs (1,000 bp upstream of the transcription start site (TSS) to 50 bp downstream of TSS), introns (sequence between 5′ and 3′ splice

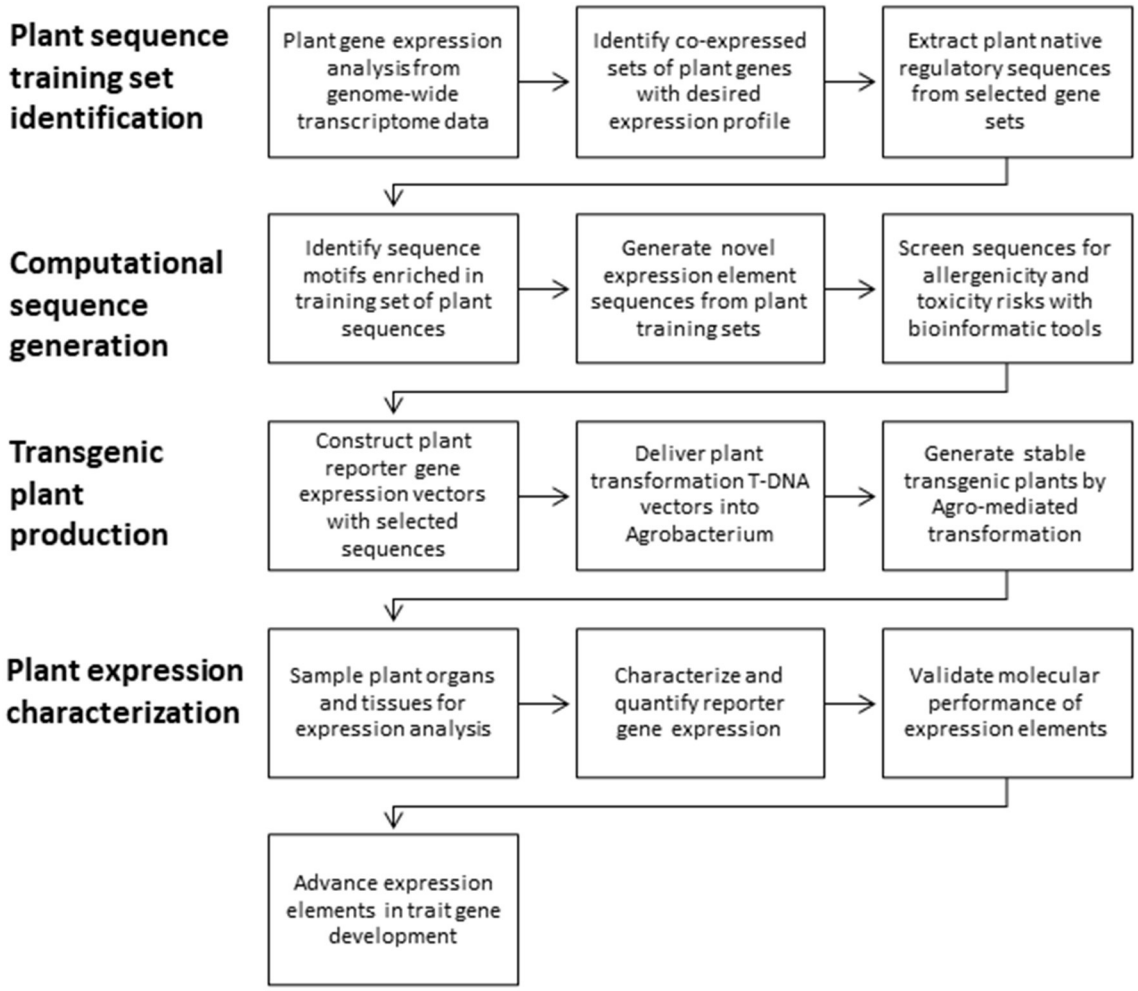

FIGURE 1 | Overview of expression element development and testing.

sites with flanking 5 bp), and 3′ UTRs (−400 to +200 bp relative to the polyadenylation site).

Computational Sequence Design

Position-specific enrichment of motifs in the sequence training sets was identified by POWRS as previously described (Davis et al., 2012). The identified putative motifs are predicted to contribute to the expression pattern and/or molecular function. The sequence training sets and putative motifs were used to train a proprietary machine learning algorithm to generate new expression element sequences. The promoters and 5′ UTRs were named GrassRootsPromoters (GSPs), the introns were named GrassRootsIntrons (GSIs), and the 3′ UTRs were named GrassRootsTerminators (GSTs). The GSPs, GSIs, and GSTs described in this study were approximately 500 bp, 300 bp, and 300 bp, respectively. These sizes are in general agreement with literature estimates of the sizes of these elements (Korkuc et al., 2014; Jafar et al., 2019). Additional design constraints were applied to reduce the risk of unintended molecular function. GSPs include a short leader sequence at the 3′ end to promote transcript processing and translation initiation of the resulting mRNA. To avoid unintended coding sequences, start codons (ATG) were avoided downstream of the predicted TSS. GSIs also include flanking exonic sequences for efficient splicing and avoiding consensus splice site sequences (Lorkovic et al., 2000) between the intended splice sites to reduce the risk of alternative splicing. As an additional precaution, bioinformatic analysis was performed to meet regulatory requirements for safety assessment of plant biotechnology products (Codex_Alimentarius, 2009), and only expression elements that met these requirements were advanced for the optimization of plant biotechnology traits.

All expression element nucleotide sequences characterized in this study are provided in **Supplementary Material S1**.

Expression Element Motif Variants

Known key motifs for molecular function of native expression elements were modified to transversions in GSPs, GSIs, and GSTs. These motifs included TATA boxes for GSPs (Zhu et al., 1995), 5′ and 3′ splice sites (Lorkovic et al., 2000), as well as intron-mediated enhancement (IME) motifs for GSIs (Rose, 2008), and

TABLE 1 | A list of expression cassettes.

Cassette	Promoter	Intron	GOI	3′ UTR
1	At.Cyco_promoter_leader	At.Cyco_intron	Ec.uidA+St.LS1	Gb.Fbl2
2	At.GSP442	At.Cyco_intron	Ec.uidA+St.LS1	Gb.Fbl2
3	At.GSP571	At.Cyco_intron	Ec.uidA+St.LS1	Gb.Fbl2
4	At.GSP576	At.Cyco_intron	Ec.uidA+St.LS1	Gb.Fbl2
5	At.GSP442_TATA	At.Cyco_intron	Ec.uidA+St.LS1	Gb.Fbl2
6	At.GSP571_TATA	At.Cyco_intron	Ec.uidA+St.LS1	Gb.Fbl2
7	At.GSP576	At.GSI17	Ec.uidA+St.LS1	Gb.Fbl2
8	At.GSP571	At.GSI21	Ec.uidA+St.LS1	Gb.Fbl2
9	At.GSP576	At.GSI17_IME	Ec.uidA+St.LS1	Gb.Fbl2
10	At.GSP576	At.GSI17_splicesite	Ec.uidA+St.LS1	Gb.Fbl2
11	At.GSP571	-	Ec.uidA+St.LS1	Gb.Fbl2
12	At.GSP571	At.GSI21_IME	Ec.uidA+St.LS1	Gb.Fbl2
13	At.GSP571	At.GSI21_splicesite	Ec.uidA+St.LS1	Gb.Fbl2
14	At.GSP571	At.Cyco_intron	Ec.uidA+St.LS1	Zm.GST7
15	At.GSP571	At.Cyco_intron	Ec.uidA+St.LS1	Zm.GST7_NUE
16	At.GSP571	At.Cyco_intron	Ec.uidA+St.LS1	Zm.GST7_T-rich_tracts

the near-upstream element (NUE) and T-rich tracts for GSTs (Li and Hunt, 1995). The sequences of expression element variants with motif mutations characterized in this study are provided in **Supplementary Material S1**.

Transgenic Plant Generation

GSPs, GSIs, GSTs, and their motif variants were tested in the context of a transgenic β-Glucuronidase (*GUS*) reporter gene from *Escherichia coli*. The functional gene expression unit, as a combination of gene expression elements and reporter gene-coding sequence, is described as a gene expression cassette. A reference gene expression cassette (cassette 1 in **Table 1**) was generated to comprise a series of plant native expression elements operatively linked together with the GUS-coding sequence, including (from 5′ to 3′) promoter and leader sequence from *Arabidopsis thaliana CYTOCHROMEC OXIDASE* gene (AT4G37830) (*At.Cyco*_promoter_leader), the first intron of the same gene (*At.Cyco*_intron) inserted within the 5′ UTR, coding sequence from *Escherichia coli* GUS gene with an inserted intron from *Solanum tuberosum* light-inducible gene (Ec.uidA+St.LS1), and the 3′ UTR from *Gossypium barbadense* Fiber Late gene (*Gb.Fbl2*). DNA fragments of the GSPs, GSIs, and GSTs were generated by synthesis and sequence verified (Bio Basic, Markham, ON, Canada). To generate expression cassettes to test the novel sequences, the corresponding functional element(s) from the reference cassette were replaced with GSPs, GSIs, or GSTs fragments. The expression cassettes were inserted into a binary plant transformation vector and verified by sequencing. The T-DNA vectors were transformed into *Agrobacterium* and introduced into *Glycine max* (A3555 germplasm) by *Agrobacterium*-mediated transformation. Transformed plants were assayed for GUS insertion copy number by DNA TaqMan assays. Transformed plants that had a single copy of the GUS transgene and normal morphological characteristics were selected for further tissue sampling and analysis.

Details of the expression cassettes characterized in this study are listed in **Table 1**. Sequences of the reference expression cassette and component elements are provided in **Supplementary Material S1**.

Plant Expression Characterization

The following organs were sampled from plants in the T0 generation at vegetative (V) and reproductive (R) developmental stages: V3 stage leaf and root; V5 stage leaf (source and sink) and root; R1 stage leaf (source and sink), root, and flowers; R3 stage pod and immature seed; and R5 stage seed cotyledon. Plant developmental stages were identified as previously described (Licht, 2014).

GUS Reporter Analysis

To assay quantitative GUS enzymatic activity, approximately 50 mg of fresh weight tissue was lyophilized and powdered. Total protein was extracted from the powdered tissue using a 500- to 800-μl 100-mM KPO4 (pH 7.4) extraction buffer (supplemented with 1-mM NaEDTA, 0.1% lauryl sarcosine, 0.1% Triton 100 X, 0.05% glycerol, 2% PVP, 10-mM β-mercaptoethanol, and 0.2-mM PMSF). Total protein concentration was determined using the Bradford protein assay per instructions of the manufacturer (BIO-RAD Life Science, Hercules, CA, USA). To assay for GUS activity, 1-3-μg total protein extract was incubated with 50-nmol 4-methylumbelliferyl-β-D-glucopyranosiduronic acid (MUG) substrate (Sigma Aldrich, St Louis, MO, USA) in a 50-μl reaction at 37°C for 1 h. The reaction was stopped by the addition of 350-μL 0.2-M sodium carbonate. The fluorescence product was measured with excitation at 365 nm, emission at 445 nm using a FLUOstar Omega microplate reader (BMG Labtech, Cary, NC, USA), then converted to pmol 4 MU with a standard curve and normalized to total protein. GUS enzyme activity was reported as pmol 4 MU/μg total protein/h. Statistical analysis was performed by *t*-tests between sample groups to

determine *p*-values. A Bonferroni-type procedure (Benjamini and Hochberg, 1995) was used to determine *p*-value cutoffs for multiple comparisons and control for a false discovery rate <0.05. Comparisons that met the significance criteria were reported in the results with the adjusted *p*-value thresholds.

To conduct qualitative expression analysis of the transformed plants, fully expanded leaves, roots, and flowers were collected from plants at the R1 stage. Leaf cross sections were cut to 90–120-micron thickness using a sliding microtome (Leica Biosystems, Buffalo Grove, IL, USA). Roots were cut manually to collect 0.5–1-mm thick sections in the mature zone. Flowers were bisected to enable staining buffer access. Leaf sections, root sections, and bisected flowers, were submerged in GUS staining solution: 1-mg/ml X-Gluc (5-bromo-4-chloro-3-indolyl-b-glucuronide) (Sigma Aldrich, St Louis, MO, USA), 25-μM potassium ferricyanide, 2.5-μM potassium ferrocyanide, 0.05% Triton X-100 (v/v) in a 50-mM potassium phosphate buffer (pH 7.4). The tissues were incubated in the staining solution at 37°C for 5 h. Destaining was performed by incubating in 70% EtOH: glacial acetic acid (1:1 v/v) overnight, followed by 70% EtOH wash. The tissues were imaged under a stereodissecting microscope (Nikon Instruments, Melville, NY, USA) for flowers, or a compound microscope (Nikon Instruments, Melville, NY, USA) for leaf and root cross sections to detect cell-type-specific expression patterns.

Transcript Characterization

Transcription Start Site Mapping by 5′ RACE

RNA was extracted from flash frozen soybean V3 or R1 leaf tissue *via* RNeasy Plant Mini kit (QIAGEN, Germantown, MD, USA), followed by RNase-free DNaseI (Ambion) treatment and cleanup by RNA Clean and Concentrator-5 kit (Zymo Research, Irvine, CA, USA) per instructions of the manufacturers. 5′ Rapid Amplification of cDNA Ends (RACE) was performed by using the First Choice RLM-RACE kit (Ambion-Thermo Fisher Scientific, Waltham, MA, USA). 5′ ends of the target cDNA were amplified by nested PCR with two pairs of the adaptor and gene-specific primers. Gene-specific primers were designed for the GUS reporter-coding sequence. PCR products were TA-cloned *via* Topo TA Cloning kit (pCRIII) (Invitrogen-Thermo Fisher Scientific, Waltham, MA, USA) and sequenced using an M13 reverse sequencing primer. Transcription start sites were identified based on the alignment of reads, containing the 5′ adapter to the expression cassette sequence.

Intron Splicing and 3′ Polyadenylation Characterization by Sequencing

Ribonucleic acid was extracted and purified as described for 5′ RACE above. Amplicons were generated by using the SMARTer® RACE 5′/3′ Kit (Takara Bio USA, Mountain View, CA, USA). In brief, cDNA was generated from total RNA by using modified oligo(dT) primers. Amplicons were created using 25 cycles of touch-down PCR with a gene-specific primer and Universal Primer A Mix. The PCR product was purified by using SeqPurebeads (Biochain, Newark, CA, USA) and verified by using a fragment analyzer (Agilent Technologies, Santa Clara, CA, USA). Final libraries were created using the Nextera DNA Flex Library Prep Kit (Illumina, San Diego, CA, USA) for tagmentation and sequencing primer addition. The final amplification and adapter addition were performed with Kappa HiFi HotStart ReadyMix (Roche Sequencing and Life Science, Wilmington, MA, USA). Libraries were pooled and sequenced with a NextSeq 500/550 mid-output kit v2.5, 300 cycles (Illumina, San Diego, CA, USA). The reads were adaptor trimmed (Trim Galore) (Krueger, 2012) and mapped to the expression cassette sequence by using a splice-aware aligner (STAR) (Dobin et al., 2013). For each expression cassette, transcripts from at least three independent transgenic events were characterized, and all uniquely mapped reads were pooled across events for the analysis. For intron-splicing analysis, exon-exon junctions were identified, and the number of reads spanning exon-exon junctions and exon-intron junctions was quantified and compared to evaluate frequencies of expected splicing, unexpected splicing, and unspliced transcripts. 3′ polyadenylation sites were identified based on the alignment of reads containing the 3′ sequencing adapter. Primer sequences used for amplicon generation are provided in **Supplementary Material S2**.

Reverse Transcriptase Quantitative Polymerase Chain Reaction (RT-qPCR) Characterization of GUS Transcript and Read-Through

Ribonucleic acid was extracted from flash frozen soy V3 or R1 leaf tissue with Tri-reagent (Sigma Aldrich, St Louis, MO, USA) and purified by Zymo Direct-zol 96 RNA kits (Zymo Research, Irvine, CA, USA) with Turbo DNase treatment (Thermo Fisher Scientific, Waltham, MA, USA). RT-qPCR assays were performed using ABI Fast 1-Step Mix (Thermo Fisher Scientific, Waltham, MA, USA) on Applied Biosystems 7900 HT instrument per instructions of the manufacturer. TaqMan primer-probe sets were designed to the GUS-coding sequence, normalizing genes, and a read through amplicon downstream of the 3′ UTR (**Supplementary Material S2**). Relative expression of the GUS and read-through transcripts were calculated and normalized by using the 2[-ΔΔC(T)] method (Schmittgen and Livak, 2008). The % read through was calculated as a percentage of read-through transcripts as compared to GUS. Statistical analysis was performed by *t*-tests with adjustments to control for false discovery in multiple testing as described above for GUS quantitative analysis. Primer and probe sequences are provided in **Supplementary Material S2**.

RESULTS

New Expression Elements Were Generated From Native Sequence Training Sets

New expression elements were generated and advanced for trait gene optimization by the following framework in four main parts (**Figure 1**). First, training sets of native plant sequences are collected. These training sets are nucleotide sequences from plant genes that demonstrate the desired expression profile (e.g., constitutive or leaf preferred) and the intended expression regulatory function (e.g., promoters, introns, or 3′ UTRs). Second, using these nucleotide sequence training sets, new

expression element sequences are generated using computational tools to recapitulate the properties represented in the training set. Third, the novel sequences are introduced into plants in the context of transgenic expression cassettes to test for function in stably transformed plants. Fourth, these new expression elements are characterized in detail *in planta* to evaluate the expression profiles and molecular function. Finally, the expression elements that meet the criteria for effective expression and molecular function are then advanced to enable trait gene optimization and development.

New expression elements characterized in this study include GrassRootsPromoters and 5′ UTRs (At.GSP442, At.GSP571, and At.GSP576) and GrassRootsIntrons (At.GSI17 and At.GSI21) that are derived from Arabidopsis training sets, as well as a GrassRootsTerminator (Zm.GST7) that is derived from a maize training set.

To show that the novel sequences are diversified from the training sets, both new and plant expression elements in the test cassettes were searched against the Arabidopsis and maize genomes by BLAST (**Supplementary Material S3**). As expected, both the native Arabidopsis sequences *At.Cyco*_promoter_leader and *At.Cyco*_intron aligned with their respective genomic source sequences with nearly complete and identical matches to the TAIR9 reference genome (E-score 0 and 1e-176). In contrast, BLAST searches with the novel sequences, as well as the native cotton sequence Gb.Fbl2_3′ UTR, only generated matches to the Arabidopsis genomic sequence with low alignment scores ($\leq$46.4) and low significance (*E*-score > 1e-5). The sequence alignments only gave fragmented matches with short stretches of continuous sequence identity $\leq$22 bp. Similarly, the GSPs, GSIs, and GSTs, as well as all of the native Arabidopsis and cotton expression elements, also only generated matches to the maize genomic sequence with low alignment ($\leq$51.8) and significance scores (*E*-score >1e-4). These results indicate that the novel sequences bear no significant sequence identity to Arabidopsis or maize genomic sequences that were used in their development. The BLAST analysis further demonstrated that the native plant expression elements analyzed here also bear no significant sequence identity to the genomic sequences of other plant species analyzed here.

GSPs Deliver Effective Gene Expression Profiles and Predictable Transcript Initiation

To evaluate the function of the new promoters *in planta*, stably transformed transgenic soybean plants with β-Glucuronidase (*GUS*) reporter gene from *Escherichia coli* were generated. All promoter and 5′ UTR leader sequences were tested in the context of the same expression cassette, where each promoter and 5′ UTR leader sequence were operatively linked to the *At.Cyco*_intron, Ec.uidA+St.LS1, and the Gb.Fbl2_3′ UTR (**Table 1**). The *At.Cyco* promoter_leader (**Figure 2A**) and CaMV.35S_promoter_leader (**Supplementary Material S7**) were used as references for comparison with the GSPs.

To evaluate the overall performance of these computationally derived promoters, a larger set of GSPs was generated from the training set of Arabidopsis leaf-preferred genes, including At.GSP571 and At.GSP576, and was tested with the GUS reporter in stable soy transformants. Overall, 43% of the 156 GSPs tested demonstrated medium to super-high levels of average leaf tissue expression as intended. The highest expression levels detected were comparable to *CaMV.35S*, while low-leaf expression was detected in the rest of the GSPs, indicating that this computational approach can generate a useful range of transgene expression that can be utilized for different plant biotechnology traits. We focused the detailed characterization efforts on a subset of expression elements.

The expression profile of expression cassettes with *At.Cyco* promoter_leader, At.GSP442, At.GSP571, and At.GSP576 (**Table 1**, cassettes 1, 2, 3, and 4) was characterized in detail across vegetative and reproductive stages in leaf, root, flower, pod, and seed. All four promoter cassettes showed measurable GUS reporter activity across multiple tissue types and developmental stages (**Figures 2A–E**), with the exception that the *At.Cyco*_promoter cassette expression at V5 stage leaf was below the limit of quantification (< 20-pmol 4-MU/μg total protein/h). GUS-staining images at the R1 developmental stage in cross-sections of source leaf and root, as well as whole mount flowers, corroborated the quantitative GUS activity detected. Moreover, this staining showed broad GUS expression across multiple cell types in the leaf and root, although *At. Cyco* promoter expression in the leaf was concentrated in vascular tissues. Overall, the range of GUS activity detected across tissue types from the GSP cassettes was comparable or higher than *At. Cyco*, indicating that these computationally derived promoters can effectively drive gene expression.

While all four promoters tested showed detectable activity, the expression profile of each promoter was unique. The *At. Cyco* promoter showed a developmentally regulated profile (**Figure 2A**). GUS expression was significantly higher in both source leaf and root in R1 than in V5 ($p < 0.05$), where the GUS activity increased from the limit of quantification at V5 to 69.7 ± 15.6-pmol/μg total protein/h and 191. ± 41.4-pmol/μg total protein/h in R1 source leaf and root, respectively. At.GSP442 demonstrated low to medium expression with a broad profile that was root enhanced (**Figure 2B**). GUS activity in roots was measured at 81.2 ± 8.8-pmol/μg total protein/h and 109.2 ± 16.6-pmol/μg total protein/h at V5 and R1 stages, which was significantly higher than in source leaf at both stages ($p < 0.05$) by 48 and 58%, respectively. At.GSP571 and At.GSP576 both showed high-expression levels that were leaf and pod enhanced (**Figures 2C,D**). For both At.GSP571 and At.GSP576, the highest GUS activity was found in the pod wall, measured at 2,141.4 ± 616.3-pmol/μg total protein/h and 701.1 ± 141.9-pmol/μg total protein/h, respectively. The detected GUS expression in pod walls was significantly higher than immature seed dissected out of the pod at the same R3 stage and the R5 stage ($p < 0.05$), with a greater than 3-fold difference in GUS activity. At.GSP571 showed generally above ground-preferred expression, with significantly higher expression in various leaf and flower tissues than in root tissues at both V5 and R1 stages ($p < 0.05$). At.GSP576 showed a similar above ground-preferred expression similar to At.GSP571, albeit with overall

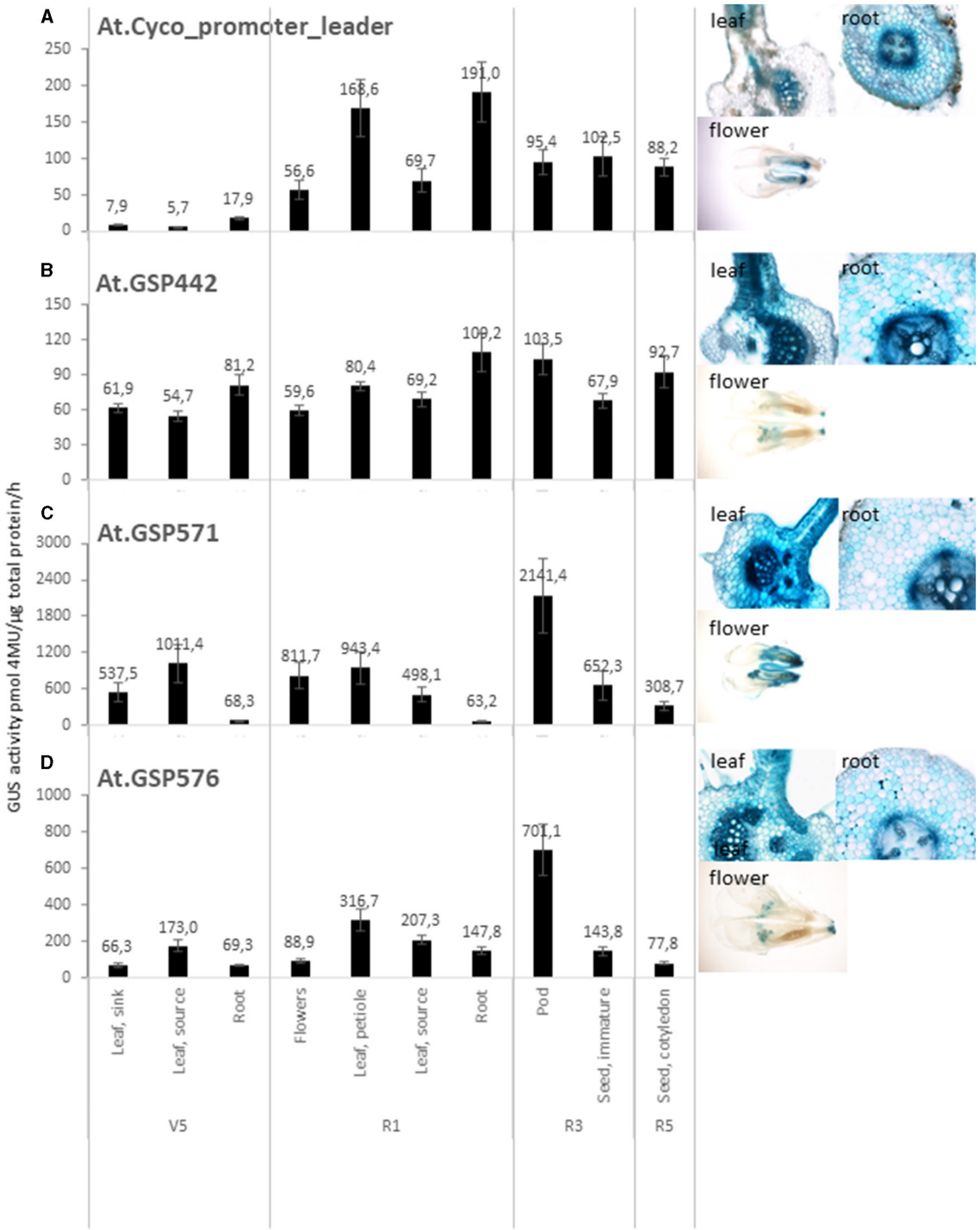

FIGURE 2 | GSPs drive effective gene expression. Quantitative and qualitative analysis of GUS reporter genes driven by promoter and leaders from At.Cyco **(A)**, At.GSP442 **(B)**, At.GSP571 **(C)**, and At.GSP576 **(D)**. Left panels: Quantitative analysis of GUS reporter gene activity was performed on transgenic soybean plants across multiple tissues, including leaf, root, flowers, pods, and seeds over vegetative (V5) and reproductive (R1, R3, and R5) stages. GUS enzyme activity on MUG substrate was normalized to total protein and reported as pmol 4-MU/µg total protein/h. At least six independent transgenic events were analyzed, and the data are reported as the mean with standard error. Right images: Qualitative analysis was performed on leaf, root, and flowers at the R1 stage. Tissue cross-sections (leaf and root) or whole mount (flower) were incubated with X-Gluc substrate to produce blue staining where GUS enzyme activity was detected. At least five independent transformation events were imaged, and one representative image is shown.

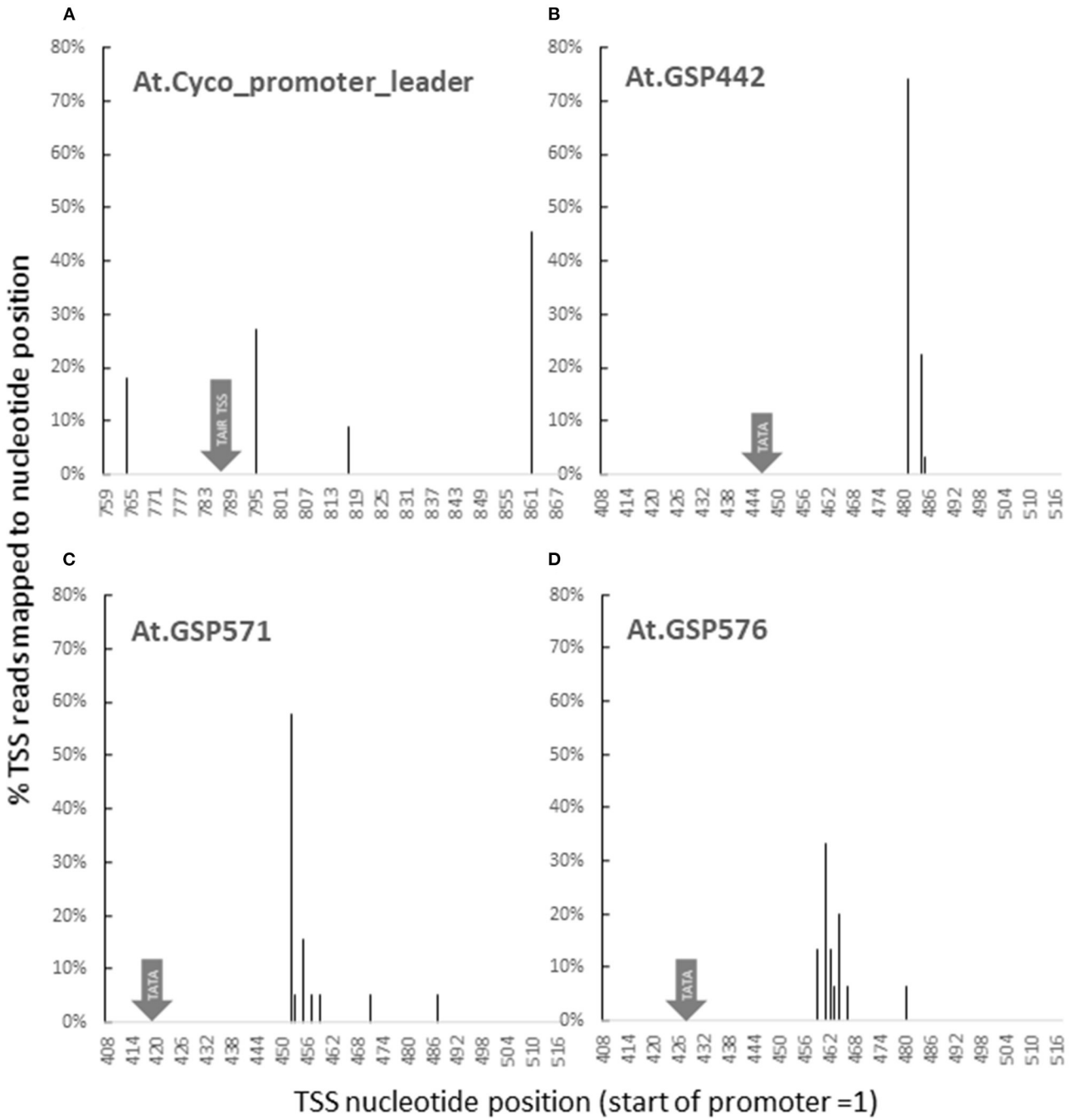

FIGURE 3 | GSPs initiate transcription as predicted. Transcription start sites (TSS) for GUS reporter genes driven by promoters and leaders from At.Cyco **(A)**, At.GSP442 **(B)**, At.GSP571 **(C)**, and At.GSP576 **(D)** were identified by 5′ RACE. 5′ ends of GUS cDNA fragments generated from leaf RNA were captured and sequenced. At least two independent transgenic events were characterized to produce > 20 reads. Sequencing reads were mapped to promoters to identify the TSS. The resulting sequencing reads were pooled to report the TSS distribution as % reads mapped to each nucleotide position. Nucleotide positions are reported with the 5′ most position of the promoter = 1.

lower levels of expression across above ground tissues compared with At.GSP571. Significant differences were also observed in V5 leaf, R1 flowers, and R3 and R5 seed tissues ($p < 0.025$), ranging from a 2- to 9-fold difference. In particular, expression in flowers for At.GSP576 is significantly lower than leaf tissues sampled at the same developmental stage ($p < 0.05$), indicating a difference in tissue-specific expression profiles between At.GSP571 and At.GSP576. Overall, our data demonstrate that

these new promoters can direct diverse expression levels and tissue specificity.

To assess whether the key molecular function of promoters followed predictions, 5′ RACE was used to map transcription start sites (TSS) for *At.Cyco*_promoter and the three GSPs. For *At. Cyco*, transcription initiation was detected at four sites across a 100-bp region (**Figure 3A**). Interestingly, none of the four detected sites coincided with the TSS annotated in TAIR9 or TAIR10 at 787 (position numbered with 5′ end of promoter = 1). The closest TSS detected was at 795, which was similar to the plant Initiator element (Inr) (Nakamura et al., 2002; Hetzel et al., 2016). The dominant transcription start site detected was the most downstream of the four at 861. All four TSS's for *At.Cyco*_promoter_intron were located ∼30-bp downstream of an AT-rich region, which is consistent with previous reports of transcription initiation in relation to TATA-like sequences (Zuo and Li, 2011; Hetzel et al., 2016). For all three GSPs, 90% transcription initiation was detected within a 10-bp window, with dominant transcription start sites detected at positions 481, 452, and 462, for At.GSP442, At.GSP571, and At.GSP576, respectively (**Figures 3B–D**). The sequences at all three GSP TSSs were similar to the Arabidopsis Inr-like consensus sequence (Hetzel et al., 2016). These transcription initiation sites were also predictably located ∼30 bp downstream of the intended TATA box, similar to endogenous plant promoters. These dominant and narrow TSS peaks were also found to be consistent with results from high throughput sequencing of 5′ transcript ends (**Supplementary Material S4**). Overall, these TSS mapping results for At.GSP442, At.GSP571, and At.GSP576 are consistent with the function of this intentionally placed TATA box in directing transcription initiation through interaction with RNAPol II and other cellular machinery.

To test whether the TATA box is necessary for promoter-driven expression in these promoters, mutant variants of both At.GSP442 and At.GSP571 were generated by introducing transversions in the intended TATA box motif to replace the TATA box sequence with GC-rich sequences. The resulting TATA box variants were tested in the context of the same reference expression cassette as the original promoters to compare the GUS activity in the variants to the original versions of the respective promoters. Each variant showed significant decreases compared with the original promoters ($p < 0.025$; **Figures 4A,B**). The At.GSP442_TATA variant GUS activity in the V3 root was 43% lower than the original promoter and was reduced to near the assay quantification limit similar to the expression in the V3 leaf. The At.GSP571_TATA variant GUS activity in V3 leaf decreased by 72% but still maintained a low level of activity, suggesting that alternative but less effective transcription initiation sequences may be utilized in the variant. These results indicate that the TATA box is necessary for the appropriate function of these promoters.

GSIs Can Enhance Expression and Demonstrate Predictable Splicing

Building on the GSP cassettes, GSIs were substituted for *At.Cyco*_intron within the 5′ UTR of the expression test cassettes and were evaluated for their ability to drive effective expression and direct effective splicing. At.GSI21 was tested in combination with At.GSP571, and At.GSI17 was tested in combination with At.GSP576 (**Table 1**, cassettes 7 and 8). Both GSIs in the context of the relevant GSP cassette produced GUS expression levels at least comparable to those observed when using the *At.Cyco*_intron (**Table 1**, cassettes 3 and 4) and further enhanced expression in some tissues (**Figure 5**). At.GSI21 showed a significant enhancement of expression relative to the comparable At.GSP571 cassette with *At.Cyco*_intron in R1 flowers, leaf petiole, leaf source, root, and R3 pod ($p < 0.042$). At.GSI17 modified the expression profile of the comparable At.GSP576 cassette with At.Cyco_intron, with significant enhancement of expression, observed in R1 flowers and R3 seed ($p < 0.025$). The expression also appeared to be increased in leaf petiole and reduced in a pod but was not statistically significant.

To test whether the key molecular function of the introns followed predictions, transcript characterization was conducted to elucidate splicing patterns in *At.Cyco*_intron, At.GSI17, and At.GSI21. Detailed molecular characterization was performed by generating cDNA libraries from leaf RNA by reverse transcription, followed by sequencing library generation for high throughput sequencing. Sequencing reads were trimmed and mapped to the expression cassette sequences to obtain both sequence-specific information on splice junctions, as well as quantitative information on the frequency of splice site usage (**Table 2**). For each cassette, a minimum of 10,000 reads that uniquely mapped to the expression cassette were generated (**Supplementary Material S4**). Mapped reads were further analyzed to identify splice junctions across the expression cassette. Reads that demonstrated expected splicing at the predicted 5′ and 3′ sites, unspliced transcripts at the predicted 5′ and 3′ sites, as well as any unexpected splice junctions were quantified (**Table 2**). Splice junctions observed for the 5′ UTR and *St.LS1* introns indicate that the intended splice sites enable efficient splicing. To assess the predictability of the splicing at the intended 5′ and 3′ sites, read counts mapped to each of the intended splicing nucleotide positions were categorized as expected splicing, unspliced, or unexpected splicing, and the resulting number of reads in each category was normalized to the total number of reads mapped to the position. The resulting usage efficiency of the intended splicing position was reported as % observed at the expected 5′ or 3′ splicing position. The majority of mapped transcripts demonstrated expected splicing, with expected splice site usage ranging from 82.6 to 98.9% across both 5′ and 3′ splicing nucleotide positions for all three cassettes. Minor amounts of unspliced introns and alternative splicing were detected for all three 5′ UTR introns. For the *At.Cyco*_intron, 1.2 and 1.5% unspliced reads were detected at the intended 5′ and 3′ splice sites, respectively. In At.GSI17, 2.8 and 3.2% of reads at the 5′ and 3′ splice sites were unspliced, respectively, and 4.2 % of reads at the 5′ splicing nucleotide positions were found to be spliced to an alternative 3′ site 13 bases downstream of the intron. In At.GSI21, reads that mapped to the intended 5′ and 3′ splice sites were found to be unspliced in 10.7 and 10.8% of reads, respectively, and alternative splicing

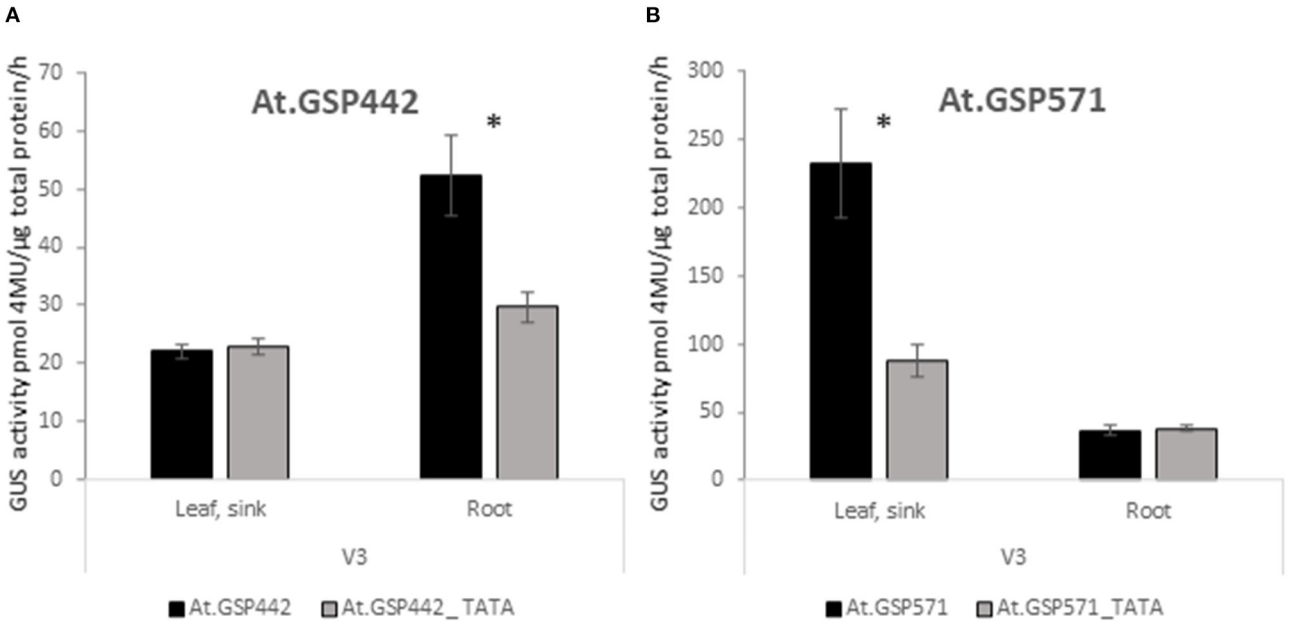

FIGURE 4 | GSPs require the TATA box for function. Expression of GUS reporter genes with TATA box variants of GSPs was compared to original At.GSP442 **(A)** and At.GSP571 **(B)** promoters. At.GSP442_TATA and At.GSP571_TATA promoter variants were generated by introducing transversions in the TATA boxes of the respective promoters. Promoter variants were tested in the same expression cassette context as the original promoters in **Figure 1**. GUS reporter gene activity was performed on leaf and root tissues from transgenic soybean plants at the V3 stage. GUS enzyme activity on MUG substrate was normalized to total protein and reported as pmol 4MU/µg total protein/h. At least 12 independent transgenic events were analyzed, and the data are reported as the mean with SE. *Indicates a statistically significant difference between original and TATA variants when compared in the same tissue type with *p*-value threshold corrections to control for false discovery < 0.05 in multiple testing.

was observed in 6.6% of reads at both positions. Two alternative 5′ splice sites and one alternative 3′ splice site were detected in these alternatively spliced reads. These results are within the range of intron-splicing efficiencies reported in Arabidopsis and soybean studies, where detectable levels of alternative splicing (including intron retention and alternative 5′ and 3′ splice sites) are found in the majority of Arabidopsis and soybean genes (Lorkovic et al., 2000; Filichkin et al., 2010; Marquez et al., 2012; Iñiguez et al., 2017; Chaudhary et al., 2019; Song et al., 2020). In summary, the GSIs demonstrated predictable splicing, with all detected splice junctions in the reporter gene found to be using at least one of the intended splice sites. Overall > 82% of the reads at the intended splice sites of GSIs were spliced as expected.

To test whether known motifs that are necessary for the function of native introns are also required for GSIs, we generated variants of At.GSI17 and At.GSI21 with transversion mutations in both predicted splice sites (At.GSI17_splice site and At.GSI21_splice site), and also in motifs previously identified to be required for intron-mediated enhancement (IME) of expression (Rose, 2008) (At.GSI17_IME and At.GSI21_IME). The resulting intron variants were tested in the context of the same expression cassettes as above (**Table 1**, cassettes 9, 10, 12, and 13) and evaluated for both GUS reporter gene expression and splicing. Mutation of the IME motifs in both GSIs reduced GUS activity compared to the original intron in the same tissue type, most notably in pod and leaf petiole for At.GSI21, and in leaf petiole for At.GSI17 (**Figure 6**); however, neither comparison met statistical significance criteria in multiple testing. At.GSI21_IME retained substantial GUS activity across all tissues assayed that is comparable to GSP571 without intron (**Table 1**, cassette 11), suggesting that expression enhancement can largely be attributed to the IME motifs (**Figure 6B**). The IME mutations did not substantially alter the splice site usage, consistent with the idea that these IME motifs are not necessary for splicing (**Table 2**). In contrast, mutations of both splice sites in At.GSI17 and At.GSI21 essentially abolished splicing at the original splice sites (**Table 2**). About 100% of reads detected at the original splice sites were unspliced in both At.GSI17_splice site and At.GSI21_splice site cassettes. Both At.GSI17_splice site and At.GSI21_splice site introns are expected to generate transcripts with short upstream open reading frames (ORFs) in the intron that would likely not produce protein. This disruption of splicing and protein expression was consistent with the large reductions in GUS reporter gene expression when compared to the corresponding original GSI across all tissues for At.GSI17 ($p < 0.05$), and across all tissues except root for At.GSI21 ($p < 0.04$) (**Figure 6**). In the expression cassette with At.GSI17_splice site, GUS activity was reduced to near the limit of quantitation across tissue types. Interestingly, for At.GSI21_splice site, the expression levels were reduced below the no intron control ($p < 0.04$), but

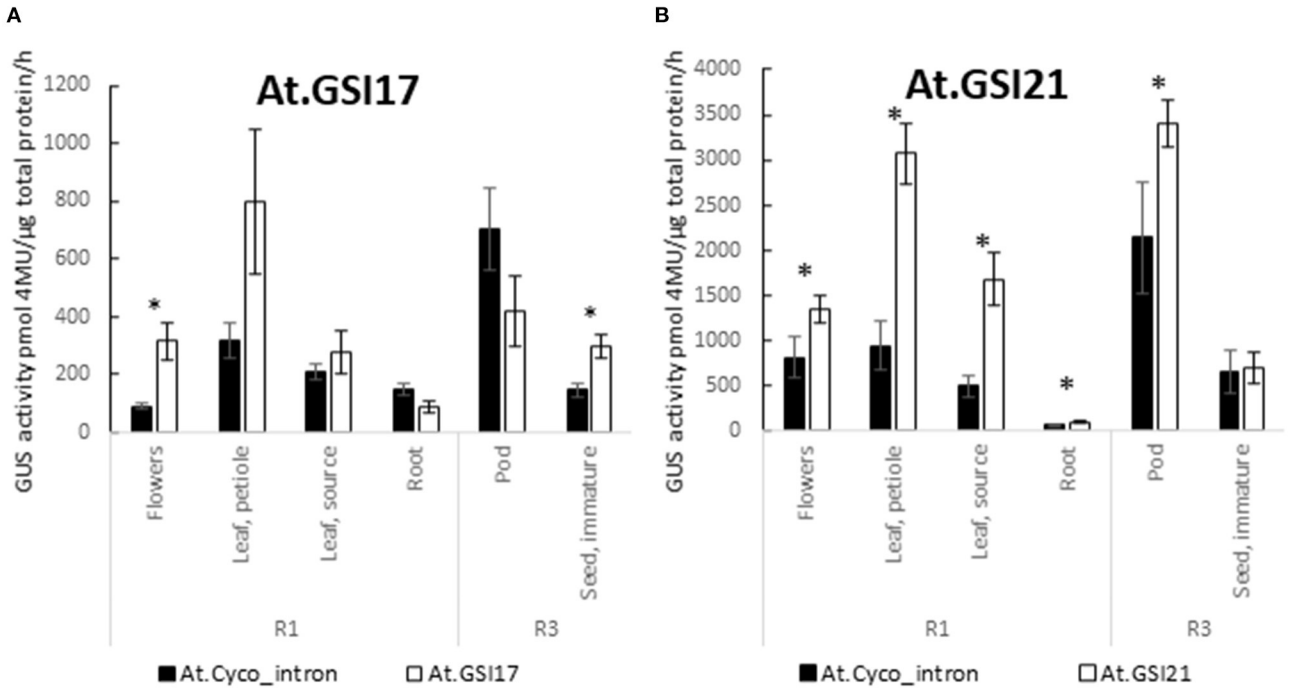

FIGURE 5 | GSIs enable effective expression. Introns At.GSI17 **(A)** and At.GSI21 **(B)** were tested in the context of At.GSP576 and At.GSP571, respectively, and the expression from cassettes with GSIs was compared with At.Cyco_intron when paired with the same promoter. GUS reporter gene activity was performed on leaf, root, flowers, pods, and seeds from transgenic soybean plants at R1 and R3 stages. GUS enzyme activity on MUG substrate was normalized to total protein and reported as pmol 4-MU/µg total protein/h. At least six independent transgenic events were analyzed, and the data are reported as the mean with standard error. *Indicates a statistically significant difference between At.Cyco_intron and the GSI when compared in the same tissue-type with *p*-value threshold corrections to control for false discovery < 0.05 in multiple testing.

the At.GSI21_splice site expression cassette still retained a low level of activity. Consistent with this observation, unexpected splicing was detected in At.GSI21_splice site with increased usage of alternative splice sites (**Table 2**). One of the alternatively spliced transcripts (5′ splice site at 580 and 3′ splice site at 673) results in an alternative translation start that is upstream and in-frame with the GUS-coding sequence and may explain the low level of GUS protein activity observed. Altogether, these results indicate that effective splicing and function of the GSIs are dependent on the splice sites that determine interactions with spliceosome machinery, whereas the IME motifs contribute to the expression enhancement.

GSTs Can Drive Effective Expression With Transcript Termination

To test for *in planta* function, Zm.GST7 was substituted for Gb.Fbl2_3′ UTR in the context of the At.GSP571 promoter testing cassette (**Table 1**, cassettes 3, 14, 15, and 16), and the resulting cassette was transformed into soybean. Compared with the At.GSP571 cassette with Gb.Fbl2_3′ UTR, the cassette with Zm.GST7 showed significantly enhanced expression in both V3 leaf and root (p <0.05; **Figure 7**). We also tested the expression activity of Zm.GST7 and another computationally derived 3′ UTR, Zm.GST43, in maize leaf protoplasts and found that both GSTs showed high-expression activity (**Supplementary Material S5**). Furthermore, Zm.GST43 demonstrated effective expression in stably transformed maize (**Supplementary Material S5**). These results show that GSTs can drive effective gene expression *in planta*.

Zm.GST7 was generated with two polyadenylation sites, similar to the known polyadenylation pattern in effective 3′ UTRs used in current commercialized plant biotech traits, such as the Nopaline synthase 3′ UTR from *Agrobacterium tumefaciens* (Depicker et al., 1982). To assess whether the molecular function of Zm.GST7 followed predictions for transcript polyadenylation, leaf RNA of soybean plants transformed with the Zm.GST7 cassette was analyzed by high-throughput sequencing of 3′ RACE libraries. The resulting sequence reads were mapped to the cassette sequence to identify polyadenylation sites. A total of 479,738 trimmed reads were mapped to the expression cassette. Of those, 3,219 reads aligned with the 3′ sequencing adaptor to define polyadenylation sites. As expected, two dominant and concentrated polyadenylation sites were found, centered around nucleotide positions 3,126 and 3,186 (**Figure 8A**). Nucleotide 3,126 is an A in a YA dinucleotide within a T-rich region that is downstream of 2 AATAAA consensus sites, the closer being 17 bp away, which is one of the expected configurations of polyadenylation sites (Li and Hunt, 1995). Nucleotide 3,186 is positioned at the end of a poly Tract, which has also been found to be enriched near cleavage sites (Wu et al., 2011). Overall, the polyadenylation sites were found to be consistent with plant native 3′ UTRs and other 3′ UTRs commonly used in plant biotechnology.

TABLE 2 | GrassRootsIntrons (GSIs) demonstrate predictable splicing that is dependent on functional splicing motifs.

Intron description	5′ splice site	3′ splice site	Splicing category	Trimmed read count	% reads obs at expected 5′ splice site	% reads obs at expected 3′ splice site
At.Cyco	511	853	Expected splicing	5,432	98.8	98.5
At.Cyco	511	853	Unspliced	65 (5′), 80 (3′)	1.2	1.5
At.GSI17	511	805	Expected splicing	52,618	93.1	96.7
At.GSI17	511	805	Unspliced	1,599 (5′), 1,809 (3′)	2.8	3.3
At.GSI17	*511*	*818*	*Unexpected splicing*	*2,304*	*4.1*	-
At.GSI17_IME	511	805	Expected splicing	37,254	93.1	96.8
At.GSI17_IME	511	805	Unspliced	1,070 (5′), 1,241 (3′)	2.7	3.2
At.GSI17_IME	*511*	*818*	*Unexpected splicing*	*1,674*	*4.2*	-
At.GSI17_splicesite	511	805	Expected splicing	0	0	0
At.GSI17_splicesite	511	805	Unspliced	28,436 (5′), 20,394 (3′)	100	100
At.GSI21	511	814	Expected splicing	17,324	82.6	83.5
At.GSI21	511	814	Unspliced	2,269 (5′), 2,212 (3′)	10.8	10.7
At.GSI21	*511*	*673*	*Unexpected splicing*	*1,385*	*6.6*	-
At.GSI21	*688*	*814*	*Unexpected splicing*	*187*	-	*0.9*
At.GSI21	*723*	*814*	*Unexpected splicing*	*1,036*	-	*5.7*
At.GSI21_IME	511	814	Expected splicing	14,207	79.7	80.9
At.GSI21_IME	511	814	Unspliced	2,570 (5′), 2,626 (3′)	14.4	15.0
At.GSI21_IME	*511*	*673*	*Unexpected splicing*	*1,047*	*5.9*	-
At.GSI21_IME	*688*	*814*	*Unexpected splicing*	*2*	-	*0.0*
At.GSI21_IME	*723*	*814*	*Unexpected splicing*	*724*	-	*4.1*
At.GSI21_splicesite	511	814	Expected splicing	0	0	0
At.GSI21_splicesite	511	814	Unspliced	6,746 (5′), 5,768 (3′)	100	100
At.GSI21_splicesite	*580*	*673*	*Unexpected splicing*	*704*	-	-
At.GSI21_splicesite	*580*	*827*	*Unexpected splicing*	*165*	-	-
At.GSI21_splicesite	*633*	*827*	*Unexpected splicing*	*146*	-	-
At.GSI21_splicesite	*723*	*827*	*Unexpected splicing*	*716*	-	-

Rows with italic values have unexpected splicing.

To determine if Zm.GST7 enabled effective transcript processing, qRT-PCR assays were designed to quantify transcript read through. This assay is comprised of two DNA primer-probe sets that enabled quantification of the relative ratio between transcripts detected within the GUS-coding region and transcripts detected downstream of the 3′ UTR. *Gb.Fbl2*_3′ UTR has typically shown a read through of 17–20% when used in a soy transgene (data not shown). For Zm.GST7, read through was detected at <5% of the GUS-coding sequence (**Figures 8B,C**). In maize, Zm.GST43 read through was very low, near the limit of quantitation (**Supplementary Material S5**). These results indicate that the transcripts processed on GSTs are predictable and have minimal read through to downstream sequences. Therefore, computationally derived 3′ UTRs present minimal risk with regard to having an adverse impact on neighboring transgenes in a trait stack.

To further assess the role of known motifs for polyadenylation such as the NUE or the T-rich tracts, variants of Zm.GST7 were generated with these motifs disrupted. The resulting Zm.GST7 variants, Zm.GST7_NUE, and Zm.GST7_T-rich_tracts were introduced into the same expression cassette in soybean to substitute for the original Zm.GST7. The two Zm.GST7 variants were then compared with the original in terms of both expression and molecular function. Both Zm.GST7_NUE and Zm.GST7_T_rich tracts cassettes were found to have significantly reduced expression of the GUS reporter as compared with Zm.GST7 in both leaf and root ($p < 0.05$) by >50% (**Figure 7**). GUS transcript analysis from leaf tissue corroborated the GUS enzymatic assay, with transcript reductions in GUS-coding sequence of the variants as compared with the original Zm.GST7 ($p < 0.05$; **Figures 8B,C**). As expected, increased levels of read-through transcripts were detected in both Zm.GST7 variants, as compared with the original Zm.GST7 ($p < 0.05$), resulting in an overall increase in percentage read through, with a 6-fold increase in the read through observed in Zm.GST7_NUE and a >10-fold increase in Zm.GST7_T-rich_tracts. Interestingly, while the mutations in Zm.GST7_NUE reduced expression and increased read through, the overall distribution of detected polyadenylation transcripts that mapped within the cassette was still similar to Zm.GST7, whereas, for Zm.GST7_T-rich_tracts, no polyadenylation transcripts were mapped within the 3′ UTR (**Supplementary Material S6**). These results indicate that, while the NUE contributes to the efficiency of 3′ UTR processing and expression, the T-rich tracts are required for defining the cleavage site for polyadenylation.

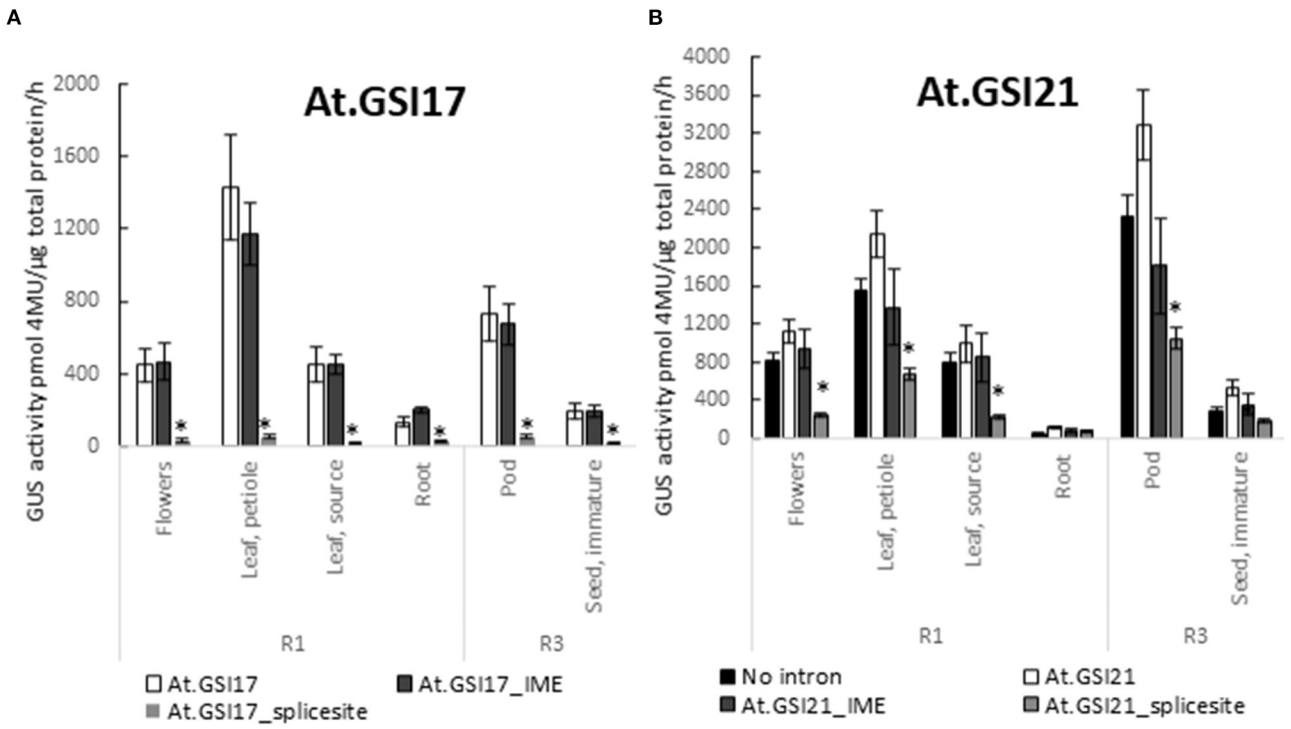

FIGURE 6 | GSIs require intron motifs for function. Expression of GUS reporter genes in the IME motif and splice site variants of GSIs was compared with the original At.GSI17 **(A)** and At.GSI21 **(B)** introns. At.GSI17_IME and At.GSI21_IME intron variants were generated by introducing transversions in the IME motifs and splicing motifs in the respective intron sequences (**Supplementary Material S1**). Intron variants were tested in the same context as the original introns as in **Figure 5**. GUS reporter gene activity was performed on various tissues from transgenic soybean plants at R1 and R3 stages. GUS enzyme activity on MUG substrate was normalized to total protein and reported as pmol 4-MU/µg total protein/h. At least six independent transgenic events were analyzed, and the data are reported as the mean with standard error. *Indicates a statistically significant difference between original and mutant variants when compared in the same tissue type with *p*-value threshold corrections to control for false discovery < 0.05 in multiple testing.

Finally, to demonstrate that these computationally derived expression elements can function together in an expression cassette, the plant sourced promoter and leader, intron, and 3′ UTR were fully substituted for At.GSP571, At.GSI21, and Zm.GST7, respectively (**Supplementary Material S7**). The expression was compared to the well-known CaMV.35S promoter as a reference. The data showed that the computationally derived expression elements were compatible and together, were able to drive leaf expression in the vegetative stage, which are comparable to 35S with intron enhancement. The results indicate that these new expression elements can be used not only to replace individual elements; they can also be used to generate new and unique expression cassette combinations that can further expand the opportunities for optimizing transgenic traits.

DISCUSSION

The availability of diversified expression elements for efficacious and predictable gene expression regulation is one of the key challenges in developing new plant biotechnology traits to meet the growing demands of farmers and consumers. In this study, we have generated a set of new and functional gene expression elements by using computational methods to learn from native sequences sourced from co-expressed plant genes. Our results demonstrate that these new expression elements can drive effective expression of a transgene and perform with the molecular characteristics of native expression elements. In all cases, we have found that overall expression levels from cassettes with these computationally derived elements were at least comparable to the reference cassette with plant native expression elements across multiple tissues and developmental stages. The new expression elements all contributed to the unique expression profile and levels of the transgene. The specific expression levels and profiles conferred by the promoters could be further modified by introns and 3′ UTRs. By developing new promoters, introns, and 3′ UTRs, we have generated a modular and diversified expression tool kit for optimizing plant biotechnology traits. Moreover, these computationally derived expression elements have no significant sequence identity to the source genome of the training set, thus offering the additional potential for sequence diversification, while recapitulating the intended molecular functions.

The computationally derived expression elements demonstrated molecular functions that are consistent with

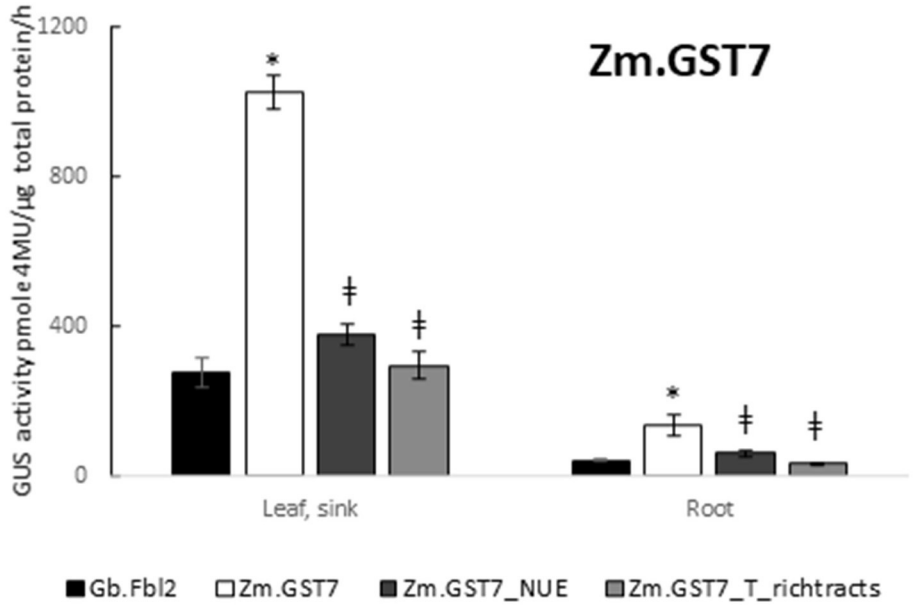

FIGURE 7 | GST7 enables effective expression that is dependent on functional polyadenylation and cleavage signals. Expression of GUS reporter genes with Zm.GST7 and its NUE polyadenylation signal and T-rich tract cleavage site variants were compared with Gb.Fbl_3′ UTR in the same expression cassette context with At.GSP571 and At.Cyco_intron. Zm.GST7_NUE and Zm.GST7_T-rich_tracts 3′ UTR variants were generated by introducing transversions in the relevant motifs. GUS reporter gene activity was performed on leaf and root tissues from transgenic soybean plants at the V3 stage. GUS enzyme activity on MUG substrate was normalized to total protein and reported as pmol 4-MU/µg total protein/h. At least 12 independent transgenic events were analyzed, and the data are reported as the mean with standard error. *Indicates a statistically significant difference between Zm.GST7 and Gb.Fbl2, whereas ≠ indicates a statistically significant difference between original Zm.GST7 and mutant variant when compared in the same tissue type with p-value threshold corrections to control for false discovery < 0.05 in multiple testing.

well-understood mechanisms in native plant genes. For example, transcription start sites from GSPs were detected ∼30 bp downstream of the AT-rich TATA box and occurred near Inr-like motifs, as observed in genome-wide analysis of plant transcription (Hetzel et al., 2016). The majority of transcripts from expression cassettes with GSIs spliced as expected, with a minor amount of intron retention and alternative splicing observed. This is consistent with reports that up to 70% of genes in plants, including Arabidopsis and soybean, are alternatively spliced (Lorkovic et al., 2000; Filichkin et al., 2010; Marquez et al., 2012; Iñiguez et al., 2017; Chaudhary et al., 2019). GSTs enabled transcript processing and termination with low- to no-read-through downstream of the 3′ UTR. For each of the GSTs, two polyadenylation sites were observed near known motifs in native plant 3′ UTRs. These motifs include the AU-rich Near Upstream Element (NUE) upstream of the polyadenylation site and the T-rich tract near the cleavage site (Li and Hunt, 1995; Wu et al., 2011). The presence of two polyadenylation sites was also consistent with the use of multiple polyadenylation sites in ∼70% of genes in plants (Shen et al., 2008; Wu et al., 2011). Mutations of key motifs associated with these molecular functions in native sequences, including the promoter TATA box, intron splice sites, and 3′ UTR NUE, and T-rich tracts resulted in reduced expression levels, indicating that these key motifs are necessary for driving expression through interactions with the cellular machinery. These data, together, support the idea that the computationally derived expression elements leverage the same cellular and molecular mechanisms as native expression elements to recapitulate these molecular properties.

In addition, the new expression elements have demonstrated desirable properties for optimizing plant biotechnology transgenes. For example, the GSPs demonstrated a dominant transcriptional start site (TSS) peak that is predictably positioned ∼30 bp downstream of the intended TATA box. Some native plant and animal promoters have been observed with a dominant TSS peak, but others have a broad transcription start region spread across the promoter and 5′ UTR (Morton et al., 2014; Mejia-Guerra et al., 2015). While both of these TSS profiles can drive transgene expression, as we have observed with *At. Cyco*_promoter and GSPs, a broad start region has some drawbacks for biotechnology. First, in angiosperms, small open-reading frames upstream of the main coding sequence are common but can decrease protein expression (von Arnim et al., 2014). Second, the scanning model for translation favors initiation at the first ATG codon encountered by the ribosome (Kozak, 1989a,b). Therefore, optimized promoters with a dominant transcription start site and 5′UTRs that are devoid of ATGs can enable a more predictable translation of the transgene-coding sequence.

Similarly, the GSIs and GSTs present opportunities for enhancing the predictability of transgene expression. For example, alternative splicing has been widely reported in native

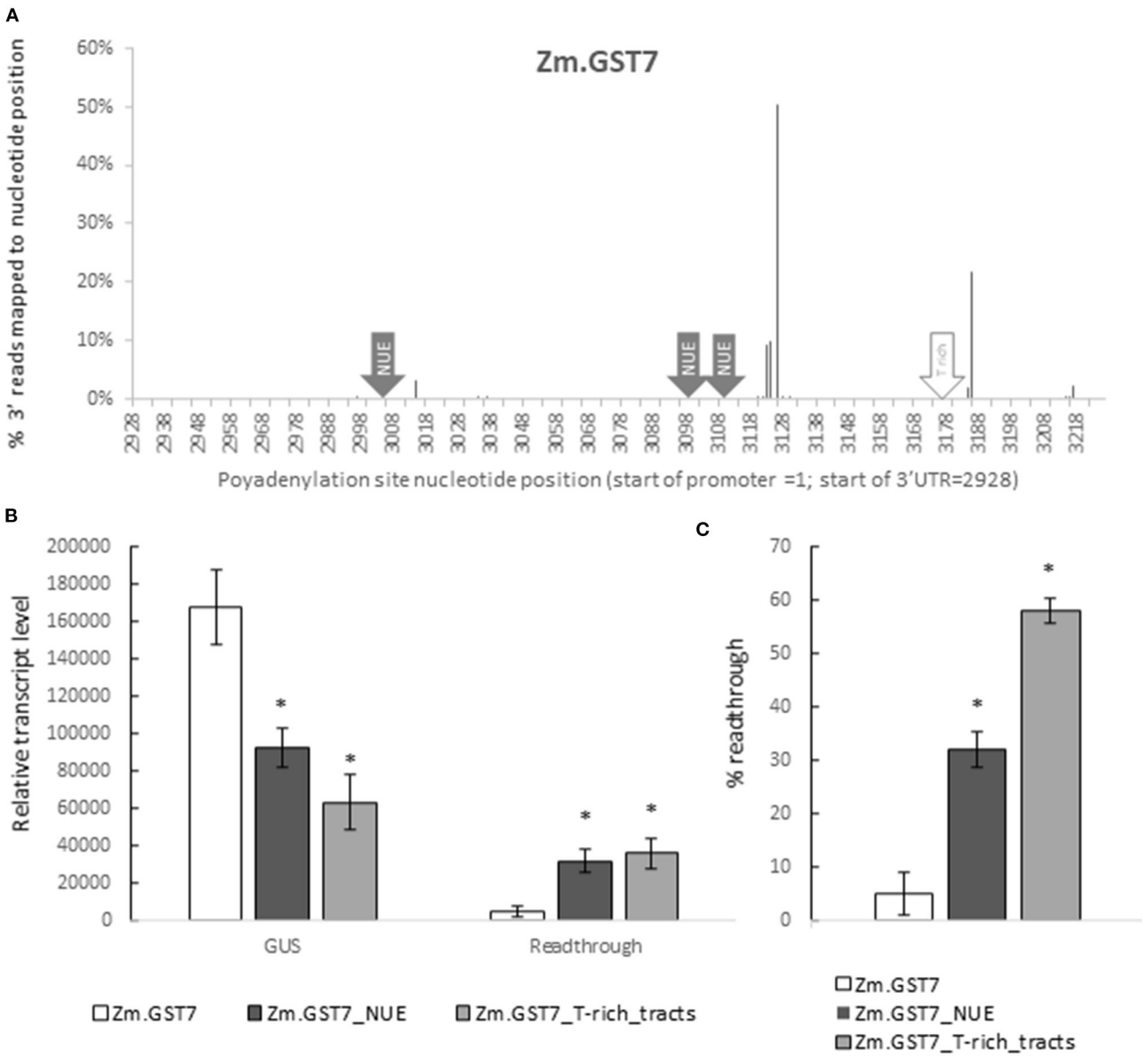

FIGURE 8 | GST7 enables effective transcript termination as predicted. This function is dependent on functional polyadenylation and cleavage signals. 3′ polyadenylation sites for GUS reporter gene with Zm.GST7 were characterized by sequencing **(A)**. 3′ ends of GUS transcript fragments from leaf RNA were captured and sequenced by high throughput sequencing. About 12 independent transgenic events were characterized to produce 479,738 reads mapped to the expression cassette after trimming. About 3,219 of these mapped reads with the 3′ sequencing adapter were identified as polyadenylation sites. The resulting sequencing reads were pooled to report the polyA site distribution as % reads mapped to each nucleotide position. Nucleotide positions are reported with the 5′ most position of the promoter = 1.GUS transcripts and read through from reporter genes with Zm.GST7 and variants in the NUE polyadenylation signal and T-rich tract cleavage site signals were analyzed by qRT-PCR **(B,C)**. GUS and read-through transcripts were detected by primer/probe sets in the GUS-coding sequence and downstream of the 3′ UTR, respectively. Transcript levels within each tissue are normalized by two [-DeltaDeltaC(T)] methods and reported as a relative expression value. At least 12 independent transgenic events were analyzed, and the data are reported as the mean with standard error. Percent read through is calculated by normalizing read-through transcript to GUS transcript **(C)**. *Indicates a statistically significant difference between original and mutant variants with corrections to control for false discovery in multiple testing.

plant genes, including intron retention, exon skipping, and the use of alternative splice sites, and has been proposed to be a mechanism for tissue or development-specific gene regulation, or response to environmental conditions (Lorkovic et al., 2000; Filichkin et al., 2010; Marquez et al., 2012; Iñiguez et al., 2017; Chaudhary et al., 2019; Song et al., 2020; Martín et al., 2021). The GSIs characterized in this study demonstrated mostly predictable splicing and did not disrupt the GUS-coding sequence, indicating that this is a potential way to reduce expression variability. While our current data demonstrate predictable splicing in

one tissue, our computational approach has the potential for learning from new genomic datasets to further optimize introns for predictable splicing across tissues and conditions. Recent improvements in molecular characterization and sequencing of transcripts (Steijger et al., 2013; Wang et al., 2016) will further improve the resolution of training sets to enable optimization of expression predictability and specificity of these elements.

GrassRoots Terminators provide another example of how computationally derived elements can be used to increase the predictability of transgene expression. Native plant genes utilize a variety of alternative polyadenylation sites, which can sometimes produce alternative coding sequences (Wu et al., 2011). The GST characterized in this study demonstrated just two polyadenylation sites as intended, and with little to no detectable read through into downstream sequences. Hence, a combination of optimized 3′ UTRs, in addition to codon optimization of the transgene to avoid unintended polyadenylation, can enable the production of transcripts with predictable coding sequences.

While our results have demonstrated predictability of native expression element motifs in these new expression elements, our results also show that additional mechanisms are providing function outside of these identified motifs. For example, disrupting the TATA box significantly reduced but did not fully abolish expression. IME motifs that were tested could only partially explain the intron-enhancing effect on expression. Specific mutations of the NUE reduced expression, and disruptions of T-rich regions increased read through, but neither set of mutations completely abolished expression. These results suggest that the computationally derived expression elements contain additional motifs within the respective expression elements that provide function in the variants. Alternative motifs for transcription initiation (Morton et al., 2014; Mejia-Guerra et al., 2015; Hetzel et al., 2016), IME (Parra et al., 2011), polyadenylation, and cleavage (Wu et al., 2011) have been reported. Recent reports of expression elements generated by combinatorial approaches with discreet modular sequence fragments or motifs have demonstrated enhanced function in plants (Liu et al., 2014; Sahoo et al., 2014; Rushton, 2016; Belcher et al., 2020). With a computational approach, our expression elements are the result of learning from the analysis of a larger set of native sequences, in which the generated sequences are predicted to contain combinations of motifs that can potentially enhance the robustness of expression. Further advances in genomic datasets and learning algorithms can enhance motif discovery and lead to further improvement in robustness of expression element performance.

Expression elements derived from larger sequence training sets from co-expressed genes may drive gene expression profiles with increased robustness and predictability. For example, At.Cyco has previously been identified in the Arabidopsis expression Atlas (Schmid et al., 2005; Klepikova et al., 2016) as a broadly expressed gene with an annotated TSS (Swarbreck et al., 2008). Interestingly, when the native promoter from this plant gene was utilized to drive the expression of a GUS reporter transgene, we found that the resulting expression profile was developmentally regulated. This indicates that transgene expression profiles do not always recapitulate those observed in nature when using native expression elements and could change depending on the gene expression cassette composition and/or genomic context, which underscores the importance of testing and characterization in the transgenic context. Interestingly, the characterized GSPs in this study generally recapitulated the overall expression pattern preferences of the training set. While our current data only showcase three promoter examples from two different training sets, this highlights the opportunity for using this computational approach to optimize tissue and developmental-stage specific expression of plant biotechnology traits.

The focus of this research is to optimize expression elements to recapitulate properties of native elements and drive efficacious expression of transgenes. Although our work features computationally derived expression elements, it does not constitute the synthetic biology of plants. In microbes, synthetic biology has assembled entire synthetic metabolic pathways of sequentially acting enzymes, or genetic circuits of interacting regulatory factors (Patron et al., 2015; Shih et al., 2016; McCarthy and Medford, 2020; Sorg et al., 2020). The expression elements described in this study are DNA sequences that can be used to drive the expression of plant biotechnology traits, but in and of themselves are not intended to be expressed as enzymes or regulatory components. While these new expression elements can be useful for metabolic engineering applications, they only serve the same roles that native elements have in previously developed, risk assessed, and commercialized plant biotechnology traits to deliver functional expression of transgenes.

Overall, we have generated functional expression elements by learning from native plant sequences and expression data. We have generated new elements by using sequence training sets from multiple plant species (Arabidopsis and maize) to be applied to different target crops (soybean and maize), and we have shown that these design principles can be applied to different types of expression elements (promoters, 5′ UTRs, introns, and 3′ UTRs). Furthermore, these different expression elements can be used together or with native elements to generate unique and optimized combinations for the expression of agronomic trait genes. Applications of this technology can be expanded to other crops and element functions to further optimize and diversify expression elements for developing future plant biotechnology traits.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ID, JT, MM, and AS conceived and designed the experiments. JT and TE supervised the execution of research. CB, KD, RG, ZG, OH, JJ, HL, MM, BO'B, AS, and JT generated the data and contributed to the analysis. JT assembled the figures and

wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

Funding for this research was provided by NSF small business grants to GrassRoots Biotechnology on constitutive promoters for crop improvement: NSF 0810649 (STTR Phase I), NSF 0957836 (STTR Phase II), and NSF 0957836 (STTR Phase IIB), as well as resources from GrassRoots Biotechnology, Monsanto Company, and Bayer Crop Science.

ACKNOWLEDGMENTS

We express our sincere thanks to PNB for invaluable input to this research. We truly appreciate the contributions of all the pipeline and functional teams who have enabled the production and analysis of plant materials for this study including Vector Production, Transformation, Controlled Environment, Molecular QC, Nucleic Acid Technologies, and Genomics and Data Sciences. We thank ML, BG, MV-S, SB, CS, QT, MM, TLR, CMD, and AC-R for many helpful discussions. We also thank CS, SS, QT, WU, TLR, TY, HL, LB, RFC, AC-R, JV, and DB for critically reviewing and providing constructive feedback for this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2021.712179/full#supplementary-material

REFERENCES

Aldemita, R. R., Reaño, I. M. E., Solis, R. O., and Hautea, R. A. (2015). Trends in global approvals of biotech crops (1992-2014). *GM Crops Food* 6, 150–166. doi: 10.1080/21645698.2015.1056972

Alexandrov, N. N., Brover, V. V., Freidin, S., Troukhan, M. E., Tatarinova, T. V., Zhang, H., et al. (2008). Insights into corn genes derived from large-scale cDNA sequencing. *Plant Mol. Biol.* 69, 179. doi: 10.1007/s11103-008-9415-4

Belcher, M. S., Vuu, K. M., Zhou, A., Mansoori, N., Agosto Ramos, A., Thompson, M. G., et al. (2020). Design of orthogonal regulatory systems for modulating gene expression in plants. *Nat. Chem. Biol.* 16, 857–865. doi: 10.1038/s41589-020-0547-4

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Brady, S. M., Orlando, D. A., Lee, J. Y., Wang, J. Y., Koch, J., Dinneny, J. R., et al. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* 318, 801–806. doi: 10.1126/science.1146265

Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., and Collins, J. J. (2018). Next-Generation Machine Learning for Biological Networks. *Cell1* 73, 1581–1592. doi: 10.1016/j.cell.2018.05.015

Chaudhary, S., Khokhar, W., Jabre, I., Reddy, A. S. N., Byrne, L. J., Wilson, C. M., et al. (2019). Alternative splicing and protein diversity: plants versus animals. *Front. Plant Sci.* 10:708. doi: 10.3389/fpls.2019.00708

Codex_Alimentarius (2009). *Foods Derived From Modern Biotechnology*. Rome: World Health Organization (WHO and Food and Agriculture Organization of the United Nations (FAO).

Datta, A. (2013). Genetic engineering for improving quality and productivity of crops. *Agric. Food Secur.* 2, 15. doi: 10.1186/2048-7010-2-15

Davis, I. W., Benninger, C., Benfey, P. N., and Elich, T. (2012). POWRS: position-sensitive motif discovery. *PLoS ONE* 7:e40373. doi: 10.1371/journal.pone.0040373

Depicker, A., Stachel, S., Dhaese, P., Zambryski, P., and Goodman, H. M. (1982). Nopaline synthase: transcript mapping and DNA sequence. *J. Mol. Appl. Genet.* 1, 561–573.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi: 10.1093/bioinformatics/bts635

Fagard, M., and Vaucheret, H. (2000). (TRANS)GENE SILENCING IN PLANTS: How Many Mechanisms? *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 51, 167–194. doi: 10.1146/annurev.arplant.51.1.167

FAO (2017). *The Future of Food and Agriculture- Trends and Challenges*. Rome: Food and Agriculture Organization of the United Nations.

Filichkin, S. A., Priest, H. D., Givan, S. A., Shen, R., Bryant, D. W., Fox, S. E., et al. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome Res.* 20, 45–58. doi: 10.1101/gr.093302.109

Gilman, J., Singleton, C., Tennant, R. K., James, P., Howard, T. P., Lux, T., et al. (2019). Rapid, Heuristic Discovery and Design of Promoter Collections in Non-Model Microbes for Industrial Applications. *ACS Synth Bio l*8, 1175–1186. doi: 10.1021/acssynbio.9b00061

Grant, T. N., De La Torre, C. M., Zhang, N., and Finer, J. J. (2017). Synthetic introns help identify sequences in the 5' UTR intron of the Glycine max polyubiquitin (Gmubi) promoter that give increased promoter activity. *Planta* 245, 849–860. doi: 10.1007/s00425-016-2646-8

Hetzel, J., Duttke, S. H., Benner, C., and Chory, J. (2016). Nascent RNA sequencing reveals distinct features in plant transcription. *Proc. Natl. Acad. Sci. U.S.A.* 113, 12316–12321. doi: 10.1073/pnas.1603217113

Hollerer, S., Papaxanthos, L., Gumpinger, A. C., Fischer, K., Beisel, C., Borgwardt, K., et al. (2020). Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping. *Nat. Commun.* 11, 3551. doi: 10.1101/2020.01.23.915405

Huang, J., Ellis, C., Hauge, B., Qi, Y., and Varagona, M. J. (2015). "Herbicide tolerance," in *Recent Advancements in Gene Expression and Enabling Technologies in Crop Plants,* eds K. Azhakanandam, A. Silverstone, H. Daniell, and M. R. Davey (New York, NY: Springer New York), 213–237.

Iñiguez, L. P., Ramírez, M., Barbazuk, W. B., and Hernández, G. (2017). Identification and analysis of alternative splicing events in *Phaseolus vulgaris* and Glycine max. *BMC Genomics* 18:650. doi: 10.1186/s12864-017-4054-2

Jafar, Z., Tariq, S., Sadiq, I., Nawaz, T., and Akhtar, M. N. (2019). Genome-wide profiling of polyadenylation events in maize using high-throughput transcriptomic sequences. *G3* 9, 2749–2760. doi: 10.1534/g3.119.400196

Jores, T., Tonnies, J., Wrightsman, T., Buckler, E. S., T., Cuperus, J., et al. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *bioRxiv* 2021.2001.2007.425784. doi: 10.1101/2021.01.07.425784

Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., and Penin, A. A. (2016). A high resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J.* 88, 1058–1070. doi: 10.1111/tpj.13312

Kooter, J. M., Matzke, M. A., and Meyer, P. (1999). Listening to the silent genes: transgene silencing, gene regulation and pathogen control. *Trends Plant Sci.* 4, 340–347. doi: 10.1016/S1360-1385(99)01467-3

Korkuc, P., Schippers, J. H., and Walther, D. (2014). Characterization and identification of cis-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. *Plant Physiol.* 164, 181–200. doi: 10.1104/pp.113.229716

Kotopka, B. J., and Smolke, C. D. (2020). Model-driven generation of artificial yeast promoters. *Nat. Commun.* 11, 2113. doi: 10.1038/s41467-020-15977-4

Kozak, M. (1989a). Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs. *Mol. Cell. Biol.* 9, 5134–5142. doi: 10.1128/mcb.9.11.5134-5142.1989

Kozak, M. (1989b). The scanning model for translation: an update. *J. Cell. Biol.* 108, 229–241. doi: 10.1083/jcb.108.2.229

Krueger, F. (2012). *Trim Galore*. Available online at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

Le Hir, H., Nott, A., and Moore, M. J. (2003). How introns influence and enhance eukaryotic gene expression. *Trends Biochem. Sci.* 28, 215–220. doi: 10.1016/S0968-0004(03)00052-5

Li, G.-X., Xu, B.-C., Yin, L.-N., Wang, S.-W., Zhang, S.-Q., Shan, L., et al. (2020). Dryland agricultural environment and sustainable productivity. *Plant Biotechnol. Rep.* 14, 169–176. doi: 10.1007/s11816-020-00613-w

Li, Q., and Hunt, A. G. (1995). A near-upstream element in a plant polyadenylation signal consists of more than six nucleotides. *Plant Mol. Biol.* 28, 927–934. doi: 10.1007/BF00042076

Licht, M. (2014). *Soybean Growth and Development*. Ames, IA: Iowa State University Extension and Outreach.

Liu, W., Mazarei, M., Peng, Y., Fethe, M. H., Rudis, M. R., Lin, J., et al. (2014). Computational discovery of soybean promoter cis-regulatory elements for the construction of soybean cyst nematode-inducible synthetic promoters. *Plant Biotechnol. J.* 12, 1015–1026. doi: 10.1111/pbi.12206

Lorkovic, Z. J., Wieczorek Kirk, D. A., Lambermon, M. H., and Filipowicz, W. (2000). Pre-mRNA splicing in higher plants. *Trends Plant Sci.*5, 160–167. doi: 10.1016/S1360-1385(00)01595-8

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., and Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.* 22, 1184–1195. doi: 10.1101/gr.134106.111

Martín, G., Márquez, Y., Mantica, F., Duque, P., and Irimia, M. (2021). Alternative splicing landscapes in *Arabidopsis thaliana* across tissues and stress conditions highlight major functional differences with animals. *Genome Biol.* 22, 35–35. doi: 10.1186/s13059-020-02258-y

Maruyama, K., Ogata, T., Kanamori, N., Yoshiwara, K., Goto, S., Yamamoto, Y. Y., et al. (2017). Design of an optimal promoter involved in the heat-induced transcriptional pathway in Arabidopsis, soybean, rice and maize. *Plant J.* 89, 671–680. doi: 10.1111/tpj.13420

McCarthy, D. M., and Medford, J. I. (2020). Quantitative and predictive genetic parts for plant synthetic biology. *Front. Plant Sci.* 11:512526. doi: 10.3389/fpls.2020.512526

Mejia-Guerra, M. K., Li, W., Galeano, N. F., Vidal, M., Gray, J., Doseff, A. I., et al. (2015). Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. *Plant Cell* 27, 3309–3320. doi: 10.1105/tpc.15.00630

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628. doi: 10.1038/nmeth.1226

Morton, T., Petricka, J., Corcoran, D. L., Li, S., Winter, C. M., Carda, A., et al. (2014). Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell* 26, 2746–2760. doi: 10.1105/tpc.114.125617

Nakamura, M., Tsunoda, T., and Obokata, J. (2002). Photosynthesis nuclear genes generally lack TATA-boxes: a tobacco photosystem I gene responds to light through an initiator. *Plant J.* 29, 1–10. doi: 10.1046/j.0960-7412.2001.01188.x

Nuccio, M. L. (2018). A brief history of promoter development for use in transgenic maize applications. *Methods Mol. Biol.* 1676, 61–93. doi: 10.1007/978-1-4939-7315-6_4

Parra, G., Bradnam, K., Rose, A. B., and Korf, I. (2011). Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. *Nucleic Acids Res.* 39, 5328–5337. doi: 10.1093/nar/gkr043

Patron, N. J., Orzaez, D., Marillonnet, S., Warzecha, H., Matthewman, C., Youles, M., et al. (2015). Standards for plant synthetic biology: a common syntax for exchange of DNA parts. *New Phytol.* 208, 13–19. doi: 10.1111/nph.13532

Peremarti, A., Twyman, R. M., Gomez-Galera, S., Naqvi, S., Farre, G., Sabalza, M., et al. (2010). Promoter diversity in multigene transformation. *Plant Mol. Biol.* 73, 363–378. doi: 10.1007/s11103-010-9628-1

Proudfoot, N. (2004). New perspectives on connecting messenger RNA 3′ end formation to transcription. *Curr. Opin. Cell Biol.* 16, 272–278. doi: 10.1016/j.ceb.2004.03.007

Que, Q., Chilton, M. D., de Fontes, C. M., He, C., Nuccio, M., Zhu, T., et al. (2010). Trait stacking in transgenic crops: challenges and opportunities. *GM Crops* 1, 220–229. doi: 10.4161/gmcr.1.4.13439

Ricroch, A. E., and Hénard-Damave, M.-C. (2016). Next biotech plants: new traits, crops, developers and technologies for addressing global challenges. *Crit. Rev. Biotechnol.* 36, 675–690. doi: 10.3109/07388551.2015.1004521

Rose, A. B. (2008). Intron-mediated regulation of gene expression. *Curr. Top. Microbiol. Immunol.* 326, 277–290. doi: 10.1007/978-3-540-76776-3_15

Rose, A. B. (2018). Introns as gene regulators: a brick on the accelerator. *Front. Genet.* 9:672. doi: 10.3389/fgene.2018.00672

Rose, A. B., and Beliakoff, J. A. (2000). Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. *Plant Physiol.* 122, 535–542. doi: 10.1104/pp.122.2.535

Rushton, P. J. (2016). What have we learned about synthetic promoter construction? *Methods Mol. Biol.* 1482, 1–13. doi: 10.1007/978-1-4939-6396-6_1

Sahoo, D. K., Sarkar, S., Raha, S., Maiti, I. B., and Dey, N. (2014). Comparative analysis of synthetic DNA promoters for high-level gene expression in plants. *Planta* 240, 855–875. doi: 10.1007/s00425-014-2135-x

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., et al. (2005). A gene expression map of *Arabidopsis thaliana* development. *Nat. Genet.* 37, 501–506. doi: 10.1038/ng1543

Schmittgen, T. D., and Livak, K. J. (2008). Analyzing real-time PCR data by the comparative C(T) method. *Nat. Protoc.* 3, 1101–1108. doi: 10.1038/nprot.2008.73

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science* 326, 1112–1115. doi: 10.1126/science.1178534

Shen, Y., Ji, G., Haas, B. J., Wu, X., Zheng, J., Reese, G. J., et al. (2008). Genome level analysis of rice mRNA 3'-end processing signals and alternative polyadenylation. *Nucleic Acids Res.* 36, 3150–3161. doi: 10.1093/nar/gkn158

Shih, P. M., Liang, Y., and Loque, D. (2016). Biotechnology and synthetic biology approaches for metabolic engineering of bioenergy crops. *Plant J.*87, 103–117. doi: 10.1111/tpj.13176

Smale, S. T., and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annu. Rev. Biochem.* 72, 449–479. doi: 10.1146/annurev.biochem.72.121801.161520

Soderlund, C., Descour, A., Kudrna, D., Bomhoff, M., Boyd, L., Currie, J., et al. (2009). Sequencing, Mapping, and Analysis of 27,455 Maize Full-Length cDNAs. *PLoS Genet.* 5:e1000740. doi: 10.1371/journal.pgen.1000740

Sonenberg, N., and Hinnebusch, A. G. (2009). Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell* 136, 731–745. doi: 10.1016/j.cell.2009.01.042

Song, L., Pan, Z., Chen, L., Dai, Y., Wan, J., Ye, H., et al. (2020). Analysis of whole transcriptome RNA-seq data reveals many alternative splicing events in soybean roots under drought stress conditions. *Genes (Basel)* 11, 1520. doi: 10.3390/genes11121520

Sorg, R. A., Gallay, C., Van Maele, L., Sirard, J. C., and Veening, J. W. (2020). Synthetic gene-regulatory networks in the opportunistic human pathogen Streptococcus pneumoniae. *Proc. Natl. Acad. Sci. U.S.A.* 117, 27608–27619. doi: 10.1073/pnas.1920015117

Steijger, T., Abril, J. F., Engström, P. G., Kokocinski, F., Hubbard, T. J., Guig,ó, R., et al. (2013). Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* 10, 1177–1184. doi, doi: 10.1038/nmeth.2714

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., et al. (2008). The arabidopsis information resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.* 36, D1009–1014. doi: 10.1093/nar/gkm965

Tilman, D., Balzer, C., Hill, J., and Befort, B. L. (2011). Global food demand and the sustainable intensification of agriculture. *Proc. Nat. Acad. Sci U.S.A.* 108, 20260–20264. doi: 10.1073/pnas.1116437108

Troukhan, M., Tatarinova, T., Bouck, J., Flavell, R. B., and Alexandrov, N. N. (2009). Genome-wide discovery of cis-elements in promoter sequences using gene expression. *OMICS* 13, 139–151. doi: 10.1089/omi.2008.0034

Vaucheret, H., Beclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J. B., et al. (1998). Transgene-induced gene silencing in plants. *Plant J.* 16, 651–659. doi: 10.1046/j.1365-313x.1998.00337.x

Venter, M. (2007). Synthetic promoters: genetic control through cis engineering. *Trends Plant Sci.* 12, 118–124. doi: 10.1016/j.tplants.2007.01.002

von Arnim, A. G., Jia, Q., and Vaughn, J. N. (2014). Regulation of plant translation by upstream open reading frames. *Plant Sci.* 214, 1–12. doi: 10.1016/j.plantsci.2013.09.006

Wang, B., Tseng, E., Regulski, M., Clark, T. A., Hon, T., Jiao, Y., et al. (2016). Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat. Commun.* 7, 1–13. doi: 10.1038/ncomms11708

Wang, Y., Wang, H., Wei, L., Li, S., Liu, L., and Wang, X. (2020). Synthetic promoter design in Escherichia coli based on a deep generative network. *Nucleic Acids Res.* 48, 6403–6412. doi: 10.1093/nar/gkaa325

Wu, X., Liu, M., Downie, B., Liang, C., Ji, G., Li, Q. Q., et al. (2011). Genome-wide landscape of polyadenylation in *Arabidopsis* provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. U.S.A.* 108, 12533–12538. doi: 10.1073/pnas.1019732108

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., and Spencer, F. (2004). A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.* 99, 909–917. doi: 10.1198/016214504000000683

Zhu, Q., Dabi, T., and Lamb, C. (1995). TATA box and initiator functions in the accurate transcription of a plant minimal promoter *in vitro*. *Plant Cell* 7, 1681–1689. doi: 10.1105/tpc.7.10.1681

Zuo, Y. C., and Li, Q. Z. (2011). Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility. *Genomics* 97, 112–120. doi: 10.1016/j.ygeno.2010.11.002

Conflict of Interest: The research activities in this report were conducted by teams at GrassRoots Biotechnology, Monsanto Company, and Bayer Crop Science. The authors contributed to this research as employees of one or more of the above entities. ID and AS are inventors of the new plant regulatory elements on a granted patent assigned to Monsanto Company.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 To, Davis, Marengo, Shariff, Baublite, Decker, Galvão, Gao, Haragutchi, Jung, Li, O'Brien, Sant and Elich. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.