A comparative analysis of stably expressed genes across diverse angiosperms exposes flexibility in underlying promoter architecture

Eric J.Y. Yang, Cassandra J. Maranas, Jennifer L. Nemhauser*
University of Washington, Department of Biology, Seattle, WA 98105-1800, USA

*email: jn7@uw.edu

Abstract

Promoters regulate both the amplitude and pattern of gene expression—key factors needed for optimization of many synthetic biology applications. Previous work in *Arabidopsis* found that promoters that contain a TATA-box element tend to be expressed only under specific conditions or in particular tissues, while promoters which lack any known promoter elements, thus designated as Coreless, tend to be expressed more ubiquitously. To test whether this trend represents a conserved promoter design rule, we identified stably expressed genes across multiple angiosperm species using publicly available RNA-seq data. Comparisons between core promoter architectures and gene expression stability revealed differences in core promoter usage in monocots and eudicots. Furthermore, when tracing the evolution of a given promoter across species, we found that core promoter type was not a strong predictor of expression stability. Our analysis suggests that core promoter types are correlative rather than causative in promoter expression patterns and highlights the challenges in finding or building constitutive promoters that will work across diverse plant species.

Introduction

Precise control over gene expression is essential for development and survival. One of the first regulatory steps in expression regulation is transcription initiation, which is controlled by DNA regions designated as promoters. Current understanding of eukaryotic promoters is still remarkably limited, and we have difficulty even identifying a precise promoter region given an arbitrary sequence (Donczew & Hahn, 2017). A core promoter region is functionally defined as the minimal region required for transcription initiation, associated with binding of RNA Polymerase II (RNAPII) and General Transcription Factors (GTFs). Proximal and distal cis-regulatory elements contribute to the modulation of the core promoter's activity and give it its characteristic expression profile. A sequence containing the proximal cis-regulatory elements as well as the core promoters is often referred to as the "promoter" region (Andersson & Sandelin, 2020; Biłas et al., 2016; Haberle & Stark, 2018; Schmitz et al., 2022). In practice, cloning and analysis projects often pick an arbitrary length (e.g., up to 2000 base pairs or until the next coding sequence) upstream of the transcription start site to define as the promoter region (Andersson & Sandelin, 2020; Schmitz et al., 2022).

Many core promoter elements have been identified within the core promoter region that are important in directing RNAPII and determining the transcription start site (TSS). The TATA-box motif is the most well-understood of the core promoter elements, yet TATA-box-containing promoters only account for about 20% of eukaryotic promoters and about 30% of *Arabidopsis* promoters (Donczew & Hahn, 2017; Molina & Grotewold, 2005). In plants, additional core promoter types were proposed by Yamamoto and colleagues based on their identification of over-represented motifs around a fixed distance from the transcription start site (Yamamoto et al., 2007, 2009). Y patch, or pyrimidine patch, motifs are C and T rich motifs whose presence had been recently shown experimentally to associate with stronger expression (Jores et al., 2021). CA and GA are additional core promoter elements, represented in approximately 20% and 1% of genic promoters, respectively (Yamamoto et al., 2009). Unlike the TATA-box which has a known GTF-binding protein associated with it, the molecular mechanism of the Y patch, CA and GA elements remain largely unknown. Core promoters that do not contain any of the identified core promoter types have been termed Coreless (Yamamoto et al., 2009, 2011). In *Arabidopsis*, Coreless promoters tend to be expressed more weakly but more broadly than those that contain TATA-boxes (Das & Bansal, 2019; Yamamoto et al., 2011).

Constitutive promoters, defined here as promoters that are on in all tissues at all times, are versatile tools in synthetic biology due to their desirable expression pattern (Yang & Nemhauser, 2022; Zhou et al., 2023). They are often used to drive expression of components used in synthetic circuits or metabolic engineering (Brophy et al., 2022; Patron, 2020; South et al., 2019; Wu et al., 2014). Core promoter regions of constitutive promoters (such as the Cauliflower Mosaic Virus 35S promoter) have often been used as the starting point to build synthetic promoters by introducing natural cis-elements or synthetic TF-binding sites upstream of these core promoter regions to artificially tune expression strength or confer new expression patterns (Ali & Kim, 2019; Belcher et al., 2020; Brophy et al., 2022; Brückner et al., 2015; Cai et al., 2020; Moreno-Giménez et al., 2022). However, a lack of understanding of the design constraints around promoters had made engineering synthetic promoters challenging. Current approaches often require trial and error or high throughput screening to identify functional synthetic promoters (Belcher et al., 2020; Brophy et al., 2022; Brückner et al., 2015; Cai et al., 2020; Moreno-Giménez et al., 2022). A better understanding of the contributions and limitations of core promoters in controlling expression patterns can therefore be essential in engineering better synthetic promoters.

Here, by leveraging publicly available RNA-seq atlases of fifteen angiosperms, we were able to map gene expression pattern onto core promoter type in multiple genomic contexts. While TATA-box-containing promoters are over-represented in conditionally-expressed genes in all of the species we examined, the pattern for Coreless promoters was less clear. In most eudicots, Coreless promoters were over-represented in stably expressed genes, but the opposite trend was observed in monocots. Additionally, by identifying orthologous gene groups within these species, we were able to track changes in core promoter type and expression pattern for groups of evolutionarily related promoters. We found that stably expressed genes are also more likely to have orthologs in other species compared to unstably expressed genes, and the orthologs tend to retain similar expression patterns. Lastly, we show that changes in core promoter types do not explain changes in expression pattern. This evolution-guided approach reveals design rules surrounding core promoter architecture and expression patterns.

Results:

We began this project by identifying species with RNA-seq Atlases, which we defined as datasets containing at least ten different tissue samples and with samples that represented at least two distinct developmental stages. Details regarding the dataset and their references can be found in Supplemental Table S1. Figure1A shows a phylogenetic tree of the fifteen species that fit our criteria, which spans a range of angiosperms including multiple monocots and eudicots. The datasets were processed through a custom pipeline (Figure1B-D). In brief, Kallisto was used for RNA-seq quantification and MultiQC was used to summarize all the outputs up till DESeq2 (Supplemental Data S7) (Bray et al., 2016; Ewels et al., 2016). For each species, normalized counts from each tissue were then converted to stability information using the coefficient of variation (CV) as a metric. In this analysis, lower CV corresponds to more stable expression, meaning comparable expression in all tissues. Higher CV, on the other hand, means less stable and more tissue-specific expression. To facilitate comparison between species, we used percentile rank of CV as the primary metric, which represents the percentage of CVs that are less than or equal to a given value.

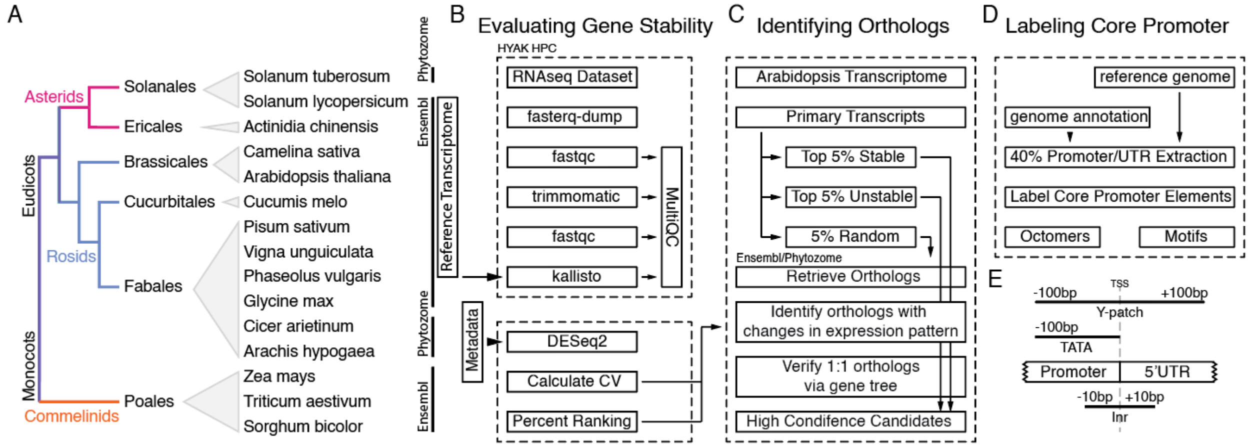

Figure 1. An outline of the bioinformatics pipelines. A) The fifteen angiosperms included in this study and their phylogenetic relationship. B-D) The three major data processing steps performed in the study. Detailed parameters are included in the Methods section. Reference genomes, transcriptomes and gene orthologs were retrieved via either Ensembl (Cunningham et al., 2021) or Phytozome (Goodstein et al., 2012) databases depending on the species. E) Regions searched for each core promoter motif.

To determine whether the characteristic differences in expression patterns between different core promoter types seen in *Arabidopsis* holds across all the species in our dataset, we extracted the -100bp to +100bp region around the TSS as the "core promoter region" for 40% of all promoters in each species (Figure1D). TATA box, Y patch, and Inr motifs were screened according to methods detailed in Jores et al. 2021. The regions scanned for each motif are more relaxed than their known regions in *Arabidopsis*, as we applied the scan to multiple species and wanted to avoid falsely labeling promoters as Coreless. Illustration of the regions scanned for each core promoter type are illustrated in Figure1E.

Forty percent of all promoters for each species were labeled as either TATA or Y patch. If a promoter did not contain either element, we labeled them as "Coreless". It is important to note that the definition of Coreless promoters introduced by Yamamoto and colleagues is somewhat more strict than the definition used here, as they also screened for the relatively rare CA and GA core promoter elements (Yamamoto et al., 2009). We then plotted the distribution of CV for each species, broken down by core promoter types (Fig. 2). Similar results for Y patch, Inr and a random set of promoters that serve as a control are in Supplemental Figure S2.

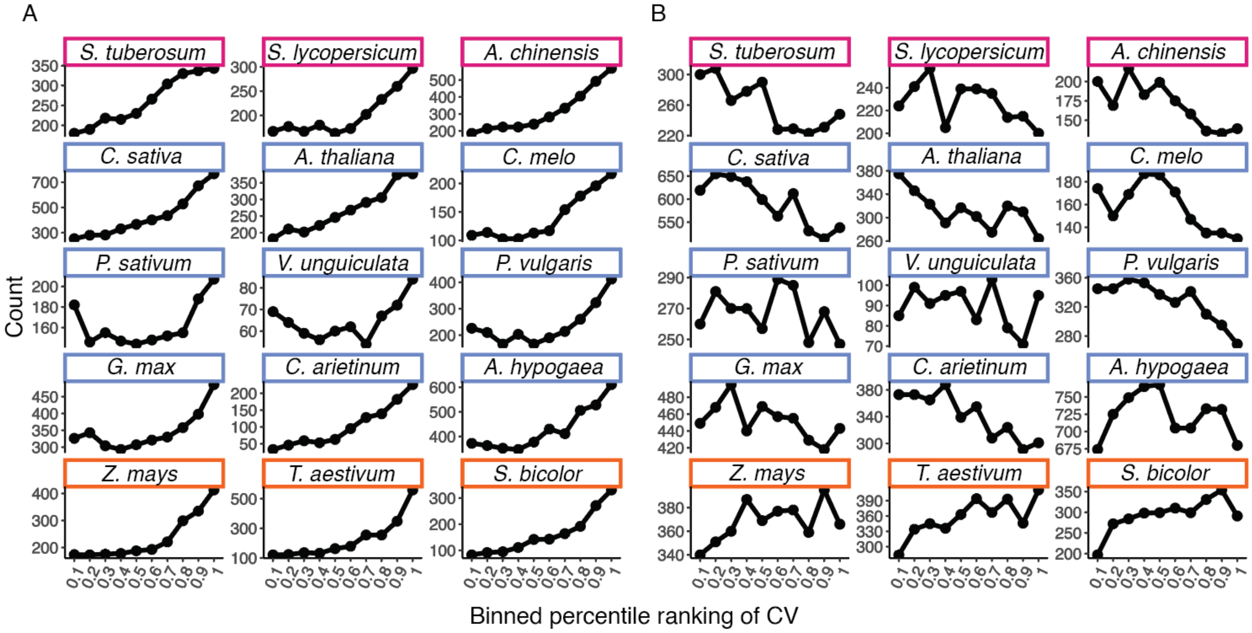

Figure 2. Distribution of relative specificity or uniformity of TATA-box-containing and Coreless promoters. Higher Coefficient of Variation (CV) rankings indicate more specificity, while lower CV rankings indicate more uniformity. A random subsampling of forty percent of promoters from each species are shown here. A) TATA-box containing promoters, and B) Promoters termed Coreless as they lacked both TATA-box and Y-path motifs. Colors correspond to phylogeny shown in Figure 1A.

Using microarray data, Yamamoto and colleagues had found that Coreless promoters are under-represented in genes that responds to stimulus (i.e. more constitutively expressed) (Yamamoto et al., 2011). However, we did not see the same trend until we removed the lowest expressing transcripts from the analysis (transcripts with an average of less than 1 read). These extremely low read counts are likely to be unreliable and an analysis of the weak-expressing genes that we removed revealed that they bias towards higher CV when compared to the rest of the genes in the dataset (Supplemental Figure S3). This same minimum read number requirement was then applied to the rest of the species.

Overall, the expected trend of TATA box-containing promoters being over-represented in unstable genes is observed across all the species analyzed (Fig. 2). In contrast, the trend of Coreless promoters being associated with more stably expressed genes was weaker and only observed in a subset of the eudicots. The monocots (*Zea mays*, *Triticum aestivum*, and *Sorghum bicolor*) all exhibited a strong trend of Coreless promoters associating with unstable genes (e.g., those with higher CV values), along with an enrichment of Y patch-containing promoters being associated with stable expression (Fig. 2 and Supplemental Figure S2). This inverted pattern could be explained in two ways given that a promoter not labeled as containing a TATA box or Y patch is labeled as Coreless. Under this classification scheme, an apparent enrichment by one category of promoters could reflect a surplus of that type of promoter in a particular CV ranking bin or a depletion of the other two promoter categories in that same bin. The latter explanation seems more likely for the Y patch promoters in monocots, but further experimental tests are required to fully resolve this question. The surprising pattern of Coreless genes "flipping" their behavior in monocots might also reflect an as yet undefined promoter element that is lumped into the Coreless category here. For example, there may be slight differences in TATA motif, as has been described for maize (Mejía-Guerra et al., 2015). Accounting for this known source of variation, we did not see any significant decrease in the Coreless trend towards conditionally-expressed genes (Supplemental Figure S2).

To determine whether core promoter type is tightly linked to expression stability for a given gene, we identified a set of orthologous genes (Figure1C). *Arabidopsis thaliana* is the most well-annotated genome, and it has 47,684 transcripts with a non-zero transcript count in at least one of the sampled tissues. Of this total, we retained only the primary transcripts of each non-mitochondrial and non-chloroplast gene, resulting in a final total of 26,842 genes. The top 5% most stable and top 5% least stable genes were selected based on CV, along with a randomly selected control set of equal size (n=1343 genes in each category). The sets of genes were used to query the Ensembl or Phytozome database for orthologs in the rest of the 14 species in our dataset (Cunningham et al., 2021; Goodstein et al., 2012). The orthologs were searched for in the database where their reference transcriptome was downloaded to ensure matching of the target transcript name with the transcript counts. Orthologs of *Arachis hypogaea, Cicer arietinum, and Solanum tuberosum* were found using Phytozome, and the remaining species were found in Ensembl.

Orthologous genes tended to retain their expression pattern across species (Fig. 3A). While orthologs corresponding to the random set of *Arabidopsis* genes were spread quite uniformly across distribution of CV rankings, the orthologs of the top 5% stable set of *Arabidopsis* genes were skewed heavily towards the more stable, lower percentage CV rankings. The orthologs of the 5% least stable set of *Arabidopsis* genes showed a more subtle skew towards higher CV ranking. This trend was more visible in some species than others, partially due to the overall lower gene counts. One notable trend was that the least stable gene set retrieved significantly fewer orthologs compared to the random or most stable gene sets (Fig. 3B). This is possibly because stable genes are associated with more fundamental cellular functions, and therefore more likely to be conserved across species (Klepikova et al., 2016). Following a similar logic, unstable genes tend to be more tissue-specific, and therefore are more easily lost during species divergence.

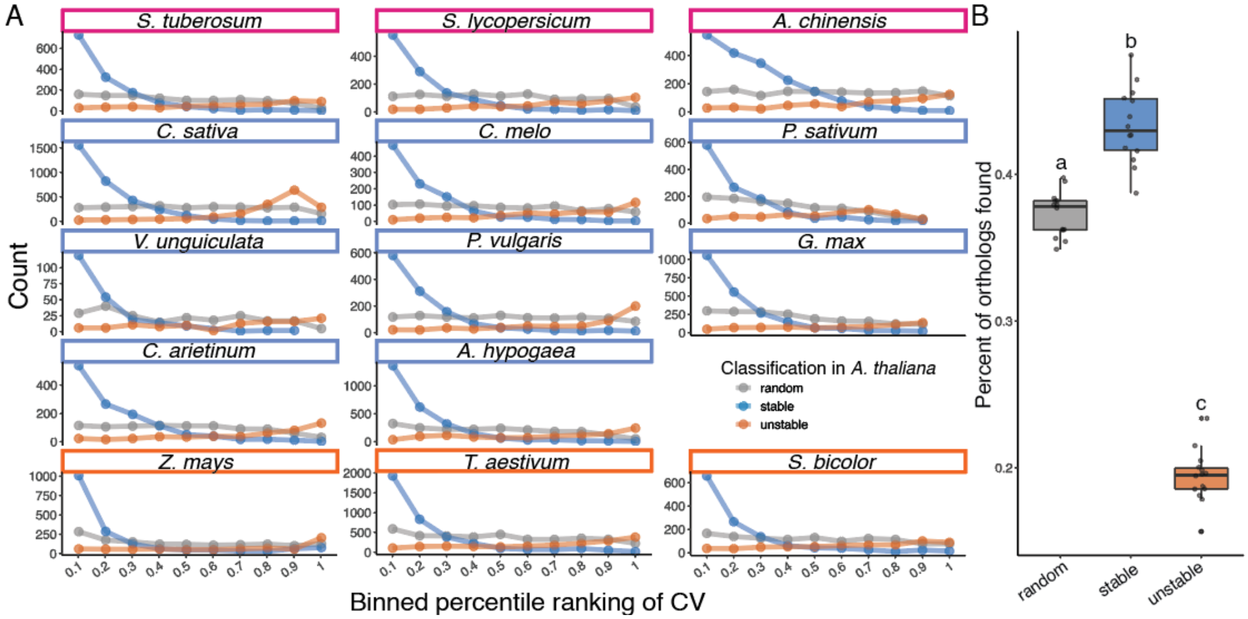

Figure 3. Genes that show uniform expression in *A. thaliana* tend to behave similarly in other species. A) Distribution of CVs for orthologs of stable (blue), unstable (orange) or random (grey) *A. thaliana* genes. The color of boxes around species names corresponds to Figure1A. B) Percent of orthologs found for each set of *A. thaliana* genes for each species. Each dot corresponds to a single species. Statistical tests were performed by one-way ANOVA followed by Tukey HSD. All three groups are significantly different from one another.

Even when looking at genes that fell at the tail ends of the expression stability distribution from *Arabidopsis*, we could find orthologs positioned across the full range of CV rankings (Fig. 3A). In other words, expression stability of a given gene can vary dramatically across species. To investigate this further, we curated a set of evolutionarily-related genes that showed this type of switching behavior. Starting with the set of all the orthologs retrieved through Ensembl and

Phytozome, we first filtered the target orthologs to count only the highest expressing transcript for each gene, thereby limiting each gene to a single representative transcript. We filtered the list of orthologs to include *Arabidopsis* transcripts that had only a single ortholog found in the transcriptome of each other species. We considered any target transcripts that crossed the 50th percentile in CV as “changing expression pattern”, and we limited the *Arabidopsis* transcripts to those where transcripts changed expression pattern in at least two different species. These changes were mapped onto the phylogenetic tree to identify clusters where changes could be associated with a specific node.

Gene trees were built for the most promising candidates, and when more than one ortholog was found in the target species, those genes were removed from further analysis (Fig. 1C). These stringent parameters maximize the likelihood that the remaining candidates are true orthologs, and that any changes in expression pattern could be biologically significant. Seven high-confidence orthologous gene groups were found with three *Arabidopsis* transcripts (AT3G17020.1, AT3G18215.1, AT4G40045.1) that are from the top 5% stable genes list and four *Arabidopsis* transcripts (AT1G04700.1, AT5G17400.1, AT5G18910.1, AT5G20410.1) from the top 5% unstable genes list. A summary of the filters and numbers of target orthologs as well as *Arabidopsis* query transcripts left after each step can be found in Supplemental Table S4.

The promoters for these seven sets of orthologs were extracted and TATA, Y patch, Inr motifs were screened for as described above (for clarity, this analysis will be referred to as Motif Scan) (Figure1D). In parallel, these promoters were also screened for TATA, Y patch, Inr, CA, GA octamers as defined in Yamamoto et al. 2009 (Octamer Scan), and an illustration of the regions scanned for each octamers can be found in Supplemental Figure S5. Comparing the two methods, the Motif Scan resulted in more identified core promoters due to its more relaxed parameters. Only two promoters were labeled as Y patch by the Octamer Scan but not the Motif Scan. A core promoter element was considered present if either method returned a positive result (Supplemental Table S6). Within each orthologous gene group, changes in the presence of TATA or Y patch elements did not appear to correlate with changes in expression patterns (Fig. 4). In each group, there are examples of promoters having the same core promoter type but different expression patterns, as well as cases of promoters having the same expression pattern but different core promoter types. Since there were only seven TATA-box-containing promoters (~15.5% of the promoters), we were not able to observe instances where two related TATA-box containing promoters having different expression patterns, but there are multiple instances where changes in presence of TATA motif did not change expression pattern. This result suggests that the presence or absence of a TATA or Y patch is not sufficient to change expression pattern.

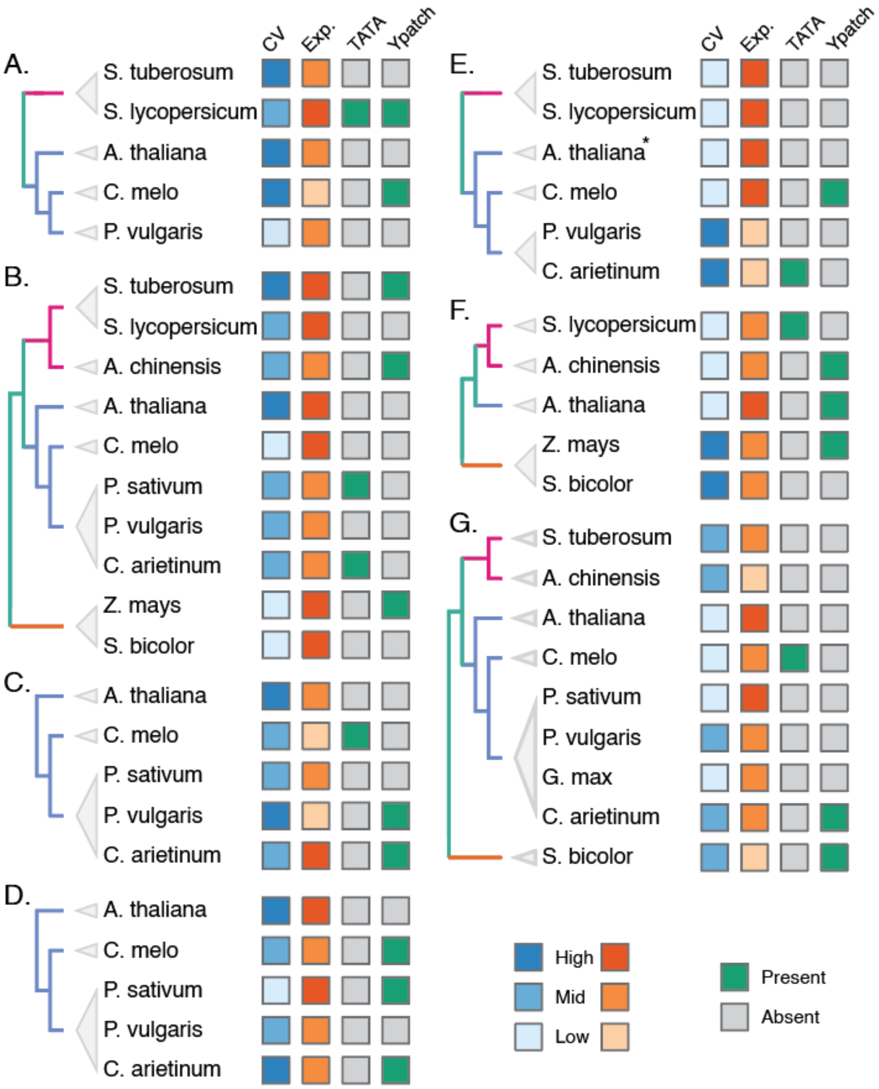

Figure4. Individual gene trees where expression stability changes can be observed. A-D) The gene is unstably expressed in *A. thaliana* but stably expressed in another species. E-G) The gene is stably expressed in *A. thaliana* but unstably expressed in another species. CV and expression strength (Exp.) is grouped by percentile ranking of 0.66~1.00 (High), 0.33~0.66 (Mid), or 0.00~0.33 (Low) and color coded accordingly. Presence (green) or absence (grey) of TATA and Y patch motifs are indicated. **A. thaliana* has no identifiable core promoter identified as the intergenic region is only 8 bp.

Discussion:

Understanding the rules that govern the performance of natural promoters could inspire the construction of synthetic promoters that are able to retain their behavior over multiple generations in transgenic plants. Here, we mined RNA-seq atlases from fifteen different angiosperms to extract patterns connected to the relative specificity or uniformity of gene expression across developmental stages and tissue types. We found that the previously observed trend that TATA-box-containing promoters are over-represented in conditionally expressed genes is highly conserved. In contrast, the relative uniformity versus specificity of expression from Coreless promoters is not as well conserved. Coreless promoters from eudicots analyzed in this study were, in general, more highly associated with stable expression patterns. Coreless promoters from monocot species, however, exhibited the opposite trend. In addition, we found that promoters tend to maintain their expression pattern across species, with the caveat that stably expressed genes are more likely to have identifiable orthologs when compared to unstably expressed genes. Lastly, by tracking expression pattern and promoter type within the evolutionary trajectory of individual genes, we could test the hypothesis that promoter architecture is responsible for the level and pattern of gene expression. We found that none of the core promoter types screened for in this work are consistently associated with changes in expression pattern or strength. This suggests that while there may be a correlation between promoter architecture and transcription parameters, the underlying molecular mechanism that determines whether a gene is conditionally or specifically expressed remains unknown.

While the general trend that TATA-box-containing promoters are found in genes that are only expressed in specific times and/or locations was highly conserved, close study of single gene phylogenies reveals that the TATA-box is not the determinant of this expression pattern. The overall lack of pattern for TATA and Y patch motifs on the phylogenetic tree also suggest that the gain and loss of these promoter elements, at least in the genes studied here, are sporadic events that do not experience strong positive selection for maintenance. In the future, it would be interesting to add the additional dimension of tracking the relative conservation versus divergence of the coding regions of the genes associated with each promoter type; however, the small number of promoters in each category would likely limit the potential to detect a clear pattern.

From a synthetic biology perspective, there are two major implications from the analysis described here. First, the hope of finding strong, constitutive natural promoters that work across diverse species may be even more challenging than we originally thought. For example, it is unlikely that there are natural promoter architectures that will work equally well as constitutive promoters in monocot and eudicot crops. Second, and more hopefully, our analysis suggests that the approach currently being taken by multiple labs for engineering synthetic promoters is likely to find solutions that work well across species (Belcher et al., 2020; Brophy et al., 2022; Cai et al., 2020; Moreno-Giménez et al., 2022). The overall scheme of many of these groups is to take a core promoter region containing a TATA-box, and then add natural cis-elements or synthetic transcription factor target sequences. We found that the same core promoter could support widely varied expression patterns. This is consistent with the emerging hypothesis that cis-elements contribute more to expression pattern than the core promoter itself (Cai et al., 2020), and that any desired expression pattern can be achieved regardless of core promoter type. Why Coreless promoters are enriched in constitutively expressed genes in eudicots, and whether this mode of regulation leads to greater robustness of expression pattern over time, will require a more detailed understanding of transcription initiation events at a range of promoters in multiple species.

Methods

Phylogenetic tree

A phylogenetic tree was constructed referencing NCBI's Taxonomy Browser and Li et al. 2021.

RNA-seq dataset processing

RNA-seq atlases were located in the NCBI Sequence Read Archive (SRA) database. The references for the datasets can be found in Supplemental Table S1. The individual datasets were retrieved using sratoolkit-3.0.1 prefetch followed by fasterq-dump functions. Fastqc-0.11.9 were used to generate a QC report for each dataset. Trimmomatic-0.39 were used for adaptor and low quality ends trimming using the following settings: 'SLIDINGWINDOW:4:20 MINLEN:36'. ILLUMINACLIP files TruSEq3-PE-2.fa was supplied for paired end data and TruSEq3-SE.fa were supplied for single end data. Reference transcriptome were downloaded from the Ensembl Plants (http://plants.ensembl.org/index.html) for *Arabidopsis thaliana, Camelina sativa, Cucumis melo, Glycine max, Phaseolus vulgaris, Pisum sativum, Vigna unguiculata, Sorghum bicolor, Zea mays, Solanum lycopersicum, Actinidia chinensis, Triticum aestivum*. and Phytozome (https://phytozome-next.jgi.doe.gov) for *Arachis hypogaea, Cicer arietinum, and Solanum tuberosum* (Cunningham et al., 2021; Goodstein et al., 2012). An index file was generated and the reads aligned and counted using Kallisto-0.44.0 with '-o counts -b 500'. For single end data, Fragment Length and Standard Deviation were required, but the information is difficult to locate, and so a default value of '-l 200 -s 20' were used across the board.

Another Fastqc was performed on the trimmed files, and a final MultiQC-1.13 were run on the entire folder encompassing all the log files that Fastqc, Trimmomatic, and Kallisto generated. The MultiQC report was inspected to ensure the trimming step improved read quality and there were no major warnings.

Normalizing count, Calculating CV and Percent Ranking

(Relevant files: 1_Metadata_from_RUNselector.Rmd, 2_MOR_Normalization.Rmd)

Using an R script, the raw counts for each species were normalized using the DESeq2 package using a metadata file curated from the original study for the RNA-seq datasets. The coefficient of variation across all samples for a given atlas was used as a metric for stability for each gene, and the percentile ranking for each gene was calculated. The geometric mean for each gene was also calculated across all samples.

Extracting intergenic region and 5'UTR

(Relevant files: 3_ExtractPromUTR(ALL_Transcripts).ipynb, 8_ExtractPromUTR(Orthologs).ipynb)

Gff3 annotation files and reference genomes were downloaded from Ensembl or Phytozome depending on where the reference transcriptomes were retrieved from. 40% of transcripts were selected from the total transcriptome and their intergenic region and 5'UTR were extracted from the Gff3 annotation. Intergenic region and 5'UTRs of identified orthologs were extracted in a similar manner.

Labeling core promoter types

(Relevant files: 4_Label_Promoters.Rmd, 9_Motif_Scan.Rmd, 10_Octamer_Scan.ipynb)

Motif Scan: Intergenic regions and 5'UTR sequences are trimmed to only regions to be scanned for each core promoter types: TATA box (-100 to TSS), Y patch (-100 to +100), and Inr (-10 to +10). Intergenic regions shorter than 100bps were excluded from analysis. Each regions were scanned for their respective motifs according using motif files as well as methods outlined in (Jores et al., 2021). A motif is considered to be present when the relative motif scores are above 0.85.

Octamer Scan: Intergenic regions and 5'UTR sequences were trimmed based on the positions relative to the TSS outlined in Yamamoto et al. 2009 (TATA, −45 to −18; Y Patch, −50 to +50; CA, −35 to −1; GA, −35 to +75). Each region was scanned for the presence of octamer motifs from the TATA, Y patch, GA, and CA lists outlined in Yamamoto et al. 2009. If the specified region contained at least one motif for a given promoter type, it was labeled as positive.

Ortholog Analysis

(Relevant files: 5_At_gene_ranking.Rmd, 6_Identifying_orthologs.Rmd, 7_Processing_orthologs.Rmd)

The *Arabidopsis* transcriptome was filtered to only include primary transcripts, and mitochondria as well as chloroplast transcripts were removed. Top 5% stable genes by CV, bottom 5% stable genes by CV and a random set of 1343 genes (5%) were randomly selected.

Using biomaRt in R, the Ensembl and Phytozome databases were queried for orthologs for the selected set of *Arabdiopsis* genes for each species (Durinck et al., 2009). Orthologs from *Arachis hypogaea, Cicer arietinum,* and *Solanum tuberosum* were retrieved from Phytozome, and the rest of the species from Ensembl. For analysis in Figure3B, significance test of done by ANOVA followed by Tukey's HSD. For each target gene that matched to an *Arabidopsis* transcript, only the highest expressing transcript was kept. If an *Arabidopsis* transcript retrieved more than one orthologs from a target species, these pairs of orthologs were removed from analysis. We only kept orthologous gene groups that had a "change" in expression pattern, defined as crossing the 50th percentile CV, in two target species, and the remaining candidates were manually mapped onto the phylogenetic tree to identify gene groups that had changes in expression pattern that are consistent with the tree. This means having changes in expression pattern that are mostly found in the same clade. Gene trees were built for these candidates using blast-align-tree (https://github.com/steinbrennerlab/blast-align-tree) and the candidate lists were further trimmed based on the gene trees to ensure a 1:1 relationship between all members in the gene group.

Data availability

All scripts and datasets necessary to perform the analysis in the article are available at https://doi.org/10.5061/dryad.9w0vt4bmk

Acknowledgements

We thank Dr. Alexander Leydon, and Janet Solano Sanchez for careful reading of the manuscript, and Dr. Adam Steinbrenner for advice on identifying orthologs. We also thank other members of the Di Stilio, Imaizumi, Steinbrenner, and Nemhauser lab for their feedback on this project. This work was supported by the National Science Foundation (IOS-1546873), the National Institute of Health (R01-GM107084) and the Howard Hughes Medical Institute Faculty Scholar Award.

Author Contributions

Experimental design and analysis by EJYY, CJM and JLN. Research performed by EJYY and CJM. Manuscript written by EJYY, CJM and JLN.

References

Ali, S., & Kim, W.-C. (2019). A Fruitful Decade Using Synthetic Promoters in the Improvement of Transgenic Plants. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01433

Andersson, R., & Sandelin, A. (2020). Determinants of enhancer and promoter activities of regulatory elements. *Nature Reviews Genetics*, *21*(2), Article 2. https://doi.org/10.1038/s41576-019-0173-8

Belcher, M. S., Vuu, K. M., Zhou, A., Mansoori, N., Agosto Ramos, A., Thompson, M. G., Scheller, H. V., Loqué, D., & Shih, P. M. (2020). Design of orthogonal regulatory systems for modulating gene expression in plants. *Nature Chemical Biology*, *16*(8), 857–865. https://doi.org/10.1038/s41589-020-0547-4

Biłas, R., Szafran, K., Hnatuszko-Konka, K., & Kononowicz, A. K. (2016). Cis-regulatory elements used to control gene expression in plants. *Plant Cell, Tissue and Organ Culture (PCTOC)*, *127*(2), 269–287. https://doi.org/10.1007/s11240-016-1057-7

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), Article 5. https://doi.org/10.1038/nbt.3519

Brian, L., Warren, B., McAtee, P., Rodrigues, J., Nieuwenhuizen, N., Pasha, A., David, K. M., Richardson, A., Provart, N. J., Allan, A. C., Varkonyi-Gasic, E., & Schaffer, R. J. (2021). A gene expression atlas for kiwifruit (Actinidia chinensis) and network analysis of transcription factors. *BMC Plant Biology*, *21*(1), 121. https://doi.org/10.1186/s12870-021-02894-x

Brophy, J. A. N., Magallon, K. J., Duan, L., Zhong, V., Ramachandran, P., Kniazev, K., & Dinneny, J. R. (2022). Synthetic genetic circuits as a means of reprogramming plant roots. *Science*, *377*(6607), 747–751. https://doi.org/10.1126/science.abo4326

Brückner, K., Schäfer, P., Weber, E., Grützner, R., Marillonnet, S., & Tissier, A. (2015). A library of synthetic transcription activator-like effector-activated promoters for coordinated orthogonal gene expression in plants. *The Plant Journal*, *82*(4), 707–716. https://doi.org/10.1111/tpj.12843

Cai, Y.-M., Kallam, K., Tidd, H., Gendarini, G., Salzman, A., & Patron, N. J. (2020). Rational design of minimal synthetic promoters for plants. *Nucleic Acids Research*, *48*(21), 11845–11856. https://doi.org/10.1093/nar/gkaa682

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., … Flicek, P. (2021). Ensembl 2022. *Nucleic Acids Research*, *50*(D1), D988–D995. https://doi.org/10.1093/nar/gkab1049

Das, S., & Bansal, M. (2019). Variation of gene expression in plants is influenced by gene architecture and structural properties of promoters. *PLOS ONE*, *14*(3), e0212678. https://doi.org/10.1371/journal.pone.0212678

Donczew, R., & Hahn, S. (2017). Mechanistic Differences in Transcription Initiation at TATA-Less and TATA-Containing Promoters. *Molecular and Cellular Biology*, *38*(1), e00448-17. https://doi.org/10.1128/MCB.00448-17

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols*, *4*(8), Article 8. https://doi.org/10.1038/nprot.2009.97

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., & Rokhsar, D. S. (2012). Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(D1), D1178–D1186. https://doi.org/10.1093/nar/gkr944

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, *19*(10), Article 10. https://doi.org/10.1038/s41580-018-0028-8

Jores, T., Tonnies, J., Wrightsman, T., Buckler, E. S., Cuperus, J. T., Fields, S., & Queitsch, C. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nature Plants*, *7*(6), 842–855. https://doi.org/10.1038/s41477-021-00932-y

Kagale, S., Koh, C., Nixon, J., Bollina, V., Clarke, W. E., Tuteja, R., Spillane, C., Robinson, S. J., Links, M. G., Clarke, C., Higgins, E. E., Huebert, T., Sharpe, A. G., & Parkin, I. A. P. (2014). The emerging biofuel crop Camelina sativa retains a highly undifferentiated hexaploid genome structure. *Nature Communications*, *5*, 3706. https://doi.org/10.1038/ncomms4706

Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., & Penin, A. A. (2016). A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *The Plant Journal*, *88*(6), 1058–1070. https://doi.org/10.1111/tpj.13312

Kudapa, H., Garg, V., Chitikineni, A., & Varshney, R. K. (2018). The RNA-Seq-based high resolution gene expression atlas of chickpea (Cicer arietinum L.) reveals dynamic spatio-temporal changes associated with growth and development. *Plant, Cell & Environment*, *41*(9), 2209–2225. https://doi.org/10.1111/pce.13210

Li, H.-T., Luo, Y., Gan, L., Ma, P.-F., Gao, L.-M., Yang, J.-B., Cai, J., Gitzendanner, M. A., Fritsch, P. W., Zhang, T., Jin, J.-J., Zeng, C.-X., Wang, H., Yu, W.-B., Zhang, R., van der Bank, M., Olmstead, R. G., Hollingsworth, P. M., Chase, M. W., … Li, D.-Z. (2021). Plastid phylogenomic insights into relationships of all flowering plant families. *BMC Biology*, *19*(1), 232. https://doi.org/10.1186/s12915-021-01166-2

Libault, M., Farmer, A., Joshi, T., Takahashi, K., Langley, R. J., Franklin, L. D., He, J., Xu, D., May, G., & Stacey, G. (2010). An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. *The Plant Journal*, *63*(1), 86–99. https://doi.org/10.1111/j.1365-313X.2010.04222.x

Loraine, A. E., McCormick, S., Estrada, A., Patel, K., & Qin, P. (2013). RNA-Seq of Arabidopsis Pollen Uncovers Novel Transcription and Alternative Splicing1[C][W][OA]. *Plant Physiology*, *162*(2), 1092–1109. https://doi.org/10.1104/pp.112.211441

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B. D., McKinley, B., Mattison, A., Morishige, D. T., Grimwood, J., Schmutz, J., & Mullet, J. E. (2018). The Sorghum bicolor reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, *93*(2), 338–354. https://doi.org/10.1111/tpj.13781

Mejía-Guerra, M. K., Li, W., Galeano, N. F., Vidal, M., Gray, J., Doseff, A. I., & Grotewold, E. (2015). Core Promoter Plasticity Between Maize Tissues and Genotypes Contrasts with

Predominance of Sharp Transcription Initiation Sites[OPEN]. *The Plant Cell*, *27*(12), 3309–3320. https://doi.org/10.1105/tpc.15.00630

Molina, C., & Grotewold, E. (2005). Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, *6*, 25. https://doi.org/10.1186/1471-2164-6-25

Moreno-Giménez, E., Selma, S., Calvache, C., & Orzáez, D. (2022). *GB_SynP: A modular dCas9-regulated synthetic promoter collection for fine-tuned recombinant gene expression in plants* (p. 2022.04.28.489949). bioRxiv. https://doi.org/10.1101/2022.04.28.489949

Patron, N. J. (2020). Beyond natural: Synthetic expansions of botanical form and function. *New Phytologist*, *227*(2), 295–310. https://doi.org/10.1111/nph.16562

Penin, A. A., Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., & Logacheva, M. D. (2019). Comparative Analysis of Developmental Transcriptome Maps of Arabidopsis thaliana and Solanum lycopersicum. *Genes*, *10*(1), 50. https://doi.org/10.3390/genes10010050

Potato Genome Sequencing Consortium, Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De la Cruz, G., Chakrabarti, S. K., … Visser, R. G. F. (2011). Genome sequence and analysis of the tuber crop potato. *Nature*, *475*(7355), 189–195. https://doi.org/10.1038/nature10158

Ramírez-González, R. H., Borrill, P., Lang, D., Harrington, S. A., Brinton, J., Venturini, L., Davey, M., Jacobs, J., van Ex, F., Pasha, A., Khedikar, Y., Robinson, S. J., Cory, A. T., Florio, T., Concia, L., Juery, C., Schoonbeek, H., Steuernagel, B., Xiang, D., … Uauy, C. (2018). The transcriptional landscape of polyploid wheat. *Science (New York, N.Y.)*, *361*(6403), eaar6089. https://doi.org/10.1126/science.aar6089

Schmitz, R. J., Grotewold, E., & Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. *The Plant Cell*, *34*(2), 718–741. https://doi.org/10.1093/plcell/koab281

South, P. F., Cavanagh, A. P., Liu, H. W., & Ort, D. R. (2019). Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field. *Science*, *363*(6422). https://doi.org/10.1126/science.aat9077

Stelpflug, S. C., Sekhon, R. S., Vaillancourt, B., Hirsch, C. N., Buell, C. R., de Leon, N., & Kaeppler, S. M. (2016). An Expanded Maize Gene Expression Atlas based on RNA Sequencing and its Use to Explore Root Development. *The Plant Genome*, *9*(1), plantgenome2015.04.0025. https://doi.org/10.3835/plantgenome2015.04.0025

Sudheesh, S., Sawbridge, T. I., Cogan, N. O., Kennedy, P., Forster, J. W., & Kaur, S. (2015). De novo assembly and characterisation of the field pea transcriptome using RNA-Seq. *BMC Genomics*, *16*(1), 611. https://doi.org/10.1186/s12864-015-1815-7

Vlasova, A., Capella-Gutiérrez, S., Rendón-Anaya, M., Hernández-Oñate, M., Minoche, A. E., Erb, I., Câmara, F., Prieto-Barja, P., Corvelo, A., Sanseverino, W., Westergaard, G., Dohm, J. C., Pappas, G. J., Saburido-Alvarez, S., Kedra, D., Gonzalez, I., Cozzuto, L., Gómez-Garrido, J., Aguilar-Morón, M. A., … Guigó, R. (2016). Genome and transcriptome analysis of the Mesoamerican common bean and the role of gene duplications in establishing tissue and temporal specialization of genes. *Genome Biology*, *17*(1), 32. https://doi.org/10.1186/s13059-016-0883-6

Wu, Y., Wang, Y., Li, J., Li, W., Zhang, L., Li, Y., Li, X., Li, J., Zhu, L., & Wu, G. (2014). Development of a general method for detection and quantification of the P35S promoter based on assessment of existing methods. *Scientific Reports*, *4*(1), Article 1. https://doi.org/10.1038/srep07358

Yamamoto, Y. Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., & Abe, T. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics*, *8*(1), 67. https://doi.org/10.1186/1471-2164-8-67

Yamamoto, Y. Y., Yoshioka, Y., Hyakumachi, M., & Obokata, J. (2011). Characteristics of Core Promoter Types with respect to Gene Structure and Expression in Arabidopsis thaliana. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *18*(5), 333–342. https://doi.org/10.1093/dnares/dsr020

Yamamoto, Y. Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., & Obokata, J. (2009). Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *The Plant Journal: For Cell and Molecular Biology*, *60*(2), 350–362. https://doi.org/10.1111/j.1365-313X.2009.03958.x

Yang, E. J. Y., & Nemhauser, J. L. (2022). *Expanding the synthetic biology toolbox with a library of constitutive and repressible promoters* (p. 2022.10.10.511673). bioRxiv. https://doi.org/10.1101/2022.10.10.511673

Yano, R., Nonaka, S., & Ezura, H. (2018). Melonet-DB, a Grand RNA-Seq Gene Expression Atlas in Melon (Cucumis melo L.). *Plant and Cell Physiology*, *59*(1), e4. https://doi.org/10.1093/pcp/pcx193

Yao, S., Jiang, C., Huang, Z., Torres-Jerez, I., Chang, J., Zhang, H., Udvardi, M., Liu, R., & Verdier, J. (2016). The Vigna unguiculata Gene Expression Atlas (VuGEA) from de novo assembly and quantification of RNA-seq data provides insights into seed maturation mechanisms. *The Plant Journal: For Cell and Molecular Biology*, *88*(2), 318–327. https://doi.org/10.1111/tpj.13279

Zhou, A., Kirkpatrick, L. D., Ornelas, I. J., Washington, L. J., Hummel, N. F. C., Gee, C. W., Tang, S. N., Barnum, C. R., Scheller, H. V., & Shih, P. M. (2023). A Suite of Constitutive Promoters for Tuning Gene Expression in Plants. *ACS Synthetic Biology*, *12*(5), 1533–1545. https://doi.org/10.1021/acssynbio.3c00075