Learning from nature: understanding and engineering transcription regulation in plants

Eric Yang

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2023

Reading Committee:
Jennifer Nemhauser, Chair
Jesse Zalatan
Eric Klavins
Ning Zheng

Program Authorized to Offer Degree
Molecular Engineering

University of Washington

Abstract

Learning from nature: understanding and engineering transcription regulation in plants

Eric Yang

Chair of the Supervisory Committee:

Jennifer Nemhauser

Department of Biology

Synthetic biology offers tools to modify plants in an environment that is changing at a faster pace than can be matched by evolution alone. The tools available come from parts either borrowed from nature or designs built upon our understanding of foundational molecular mechanisms. Some of the indispensable tools in the synthetic biology toolkit involve ways to regulate transcription strength and transcription pattern. The ability to control when and where genes are turned on is essential to survival. By leveraging publicly available RNA-seq atlases, we were able to identify a set of some of the most stably expressed genes in the genome of the reference plant *Arabidopsis thaliana*. We evaluated these promoter parts in transient assays in *Nicotiana benthamiana* and in stable transgenic lines of *Arabidopsis.* To provide additional functionality to these promoter parts borrowed from nature, we introduced gRNA-target sites recognized by a dCas9-repressor construct to turn these constitutive promoters into repressible NOR logic gates in *N. benthamiana*. To explore the fundamental design rules behind constitutive promoters, we did an *in silico* experiment that expanded the RNA-seq atlas pipeline to identify stably expressed genes across multiple angiosperm species. Comparisons between core promoter architectures and gene expression stability revealed potential differences in core promoter usage in monocots and eudicots. Furthermore, evaluating groups of evolutionarily related promoters across species found a lack of strong evolutionary preference for core promoter types for expression stability. To improve upon the repression aspect of transcriptional regulation, we used machine learning models to predict and optimize a short alpha helical repression domain, and identified potential key residues that contribute to repression. Taken together, this work contributes to our ability to engineer transcriptional regulation in plants by providing a new set of tools, as well as revealing design rules behind both gene expression pattern and repression.

Table of Contents

List of Figures

Acknowledgements

I am tremendously fortunate to have the support, mentorship, and camaraderie of many wonderful people throughout my PhD career. I would like to thank my advisor, Dr. Jennifer Nemhauser, for her patience and kindness. She is truly invested in my success, and I am extremely grateful for her guidance both regarding my research projects and my professional development. I would also like to thank my committee: Dr. Eric Klavins, Dr. Jesse Zalatan, and Dr. Ning Zheng. All of you have been nothing but kind and supportive, and our conversations always provided me with new insights.

I would also like to thank my wonderful lab mates Roman Baez, Hardik Gala, Amy Lanctot, Alexander Leydon, and Sarah Guiziou for their mentorship, as well as Benjamin Downing, Wesley George, Cassandra Maranas, and Janet Sanchez for their camaraderie and putting up with me when I'm stressed out. I also had the fortune of working with many talented undergraduates: Dante Fisher, Ani Gallman, Khushi Tawde, and Cayden Weiszmann. Thank you for working and growing with me.

Thank you to my high school friends Josh and Wayne for keeping in touch after all these years, and I am still amazed by the brief couple of years where we all managed to live in Seattle. I would also like to thank the MolES 2017 cohort for always forcing me to hang out, as you have become some of my closest friends here in Seattle.

I want to thank my family for always being there for me even though we are physically separated by about 10,000 km. Thank you for always checking up on me and keeping me updated on everything that's happening back home. I would like to thank Po and Suki, for being wonderful kitty cats. Finally, thank you to Camille, for being the kindest, sweetest partner anyone could ask for.

GEN Biotechnology
Volume 1, Number 5, 2022

DOI: 10.1089/genbio.2022.29060.jne

GEN Biotechnology

VIEWS & NEWS

En ROOT to Plant Logic

Eric J.Y. Yang and Jennifer L. Nemhauser*

A report in* Science *from Jennifer Brophy and colleagues uses engineering principles to design synthetic logic gates, which are then used to produce novel gene expression patterns and alter root architecture.

By definition, logic gates use some form of computation to turn one or more binary input signals (e.g., present or absent) into a single predictable binary output. Not surprisingly, all living things perform complex logic operations that are essential for survival. Plants are no exception. The famous ABC model of floral development captures the logic of whether sepal, petal, stamen, or carpel will develop based on the combination of transcription factors present in a cell.[1] Expression of both gene A *and* gene B gives an outcome completely different from expression of gene A *not* gene B. Many environmental responses also can be understood through the lens of logic gates. For example, plants integrate sensory information about what is attacking them (e.g., chewing caterpillars vs. sap-sucking aphids) to mount an optimal defense response through production and transport of different hormones.[2]

From a synthetic biology perspective, the ability to perform artificial logic operations can mean better spatial-temporal control over gene expression. Such control is an important step toward engineering interventions that can work effectively within the complex natural computational environment. Logic gates have already been implemented in bacteria and mammalian cells to build biosensors, event recorders, and state detection.[3–5] In plants, logic gates could be a powerful tool to engineer new plant morphologies and stress responses. These endeavors are especially important with a rapidly changing climate that is outpacing evolution.

But how can we introduce synthetic logic pathways into plants that can produce the desired outputs given arbitrary inputs? Some of the main obstacles are device orthogonality from the host system, robust characterization of parts that can respond within an appropriate dynamic range, and composability, which means making sure the individual parts work properly when assembled into more complex circuits.[6,7]

A recent groundbreaking report from Brophy et al in *Science* has successfully surmounted all of these obstacles.[8] The authors report on a suite of new open-source plant synthetic biology tools and, perhaps even more importantly, share some critical lessons about plant engineering.

The researchers used two basic gate designs: those with low basal expression that could be activated in the presence of the right combination of signals, and those with strong constitutive expression that could be repressed in the specified conditions. For both applications, they used the Cauliflower Mosaic Virus 35S promoter—a workhorse of transgenesis in plants. To either the minimal 35S promoter (for activatable gates) or the full-length 35S promoter (for repressible gates), the authors added operator sequences that could recruit synthetic transcription factors carrying cognate DNA-binding domains. Both operators and DNA-binding domains were from bacteria that allowed these circuits to be well insulated from the native plant genome. By employing a range of strategies to modulate the activity of the synthetic promoters (e.g., location and number of operator sites), they were able to tune the degree of repression or activation on a given gate.

With the individual components characterized, Brophy et al then moved on to assembling a series of logic gates, some of which were quite complex. For example, an OR gate required two separate activators, a NOR gate required two separate repressors, and a NIMPLY gate required both an activator and a repressor. Some gates also required layering, where the inputs were modulating the expression of another transcription factor instead of directly controlling the output promoter. Similar to most synthetic biology endeavors, the individual parts sometimes interacted in unforeseen ways, and caused complications when assembled. But through a series of Design–Build–Test–Learn cycles, Brophy et al were ultimately successful (Fig. 1). Using *Nicotiana benthamiana* (tobacco) as a rapid prototyping platform, they were able to demonstrate all of the single-input logic gates and four two-input logic gates (OR, NOR, AND, and NAND), along with IMPLY and NIMPLY gates.

Conventional wisdom says that gene expression is combinatorial, that the final activity of a promoter is determined by the interaction of various cell-type or environment-specific regulators that are recruited to a given promoter.[9,10] One of the applications of logic circuits is to derive any pattern of interest by combining native promoters in appropriate relation to one another, which could be invaluable for targeted gene perturbation experiments. Using well-characterized tissue-specific promoters with overlapping yet distinct expression patterns as inputs, the

Department of Biology, University of Washington, Seattle, Washington, USA.

*Address correspondence to: *Jennifer L. Nemhauser, Department of Biology, University of Washington, Seattle, WA 98195-1800, USA,* E-mail: jn7@uw.edu

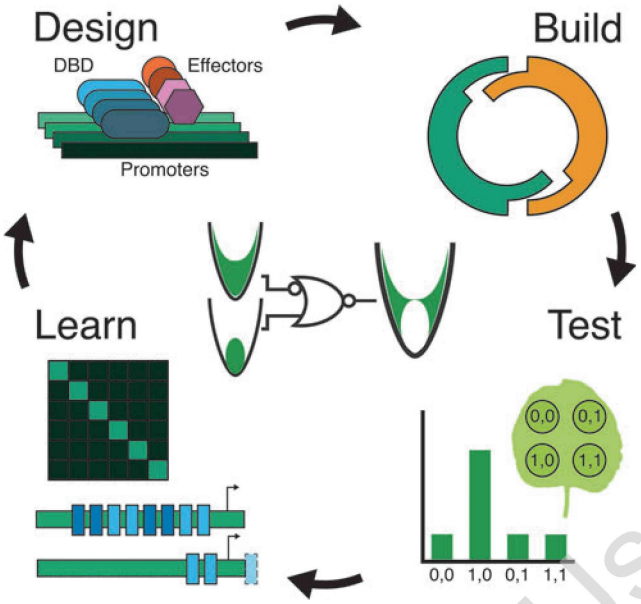


FIG. 1. Using a Design–Build–Test–Learn cycle, Brophy et al[8] produced novel gene expression patterns and altered root architecture.
(Credit: Eric Yang)

authors created brand new expression patterns in *Arabidopsis thaliana* roots. Porting the logic gates from tobacco to *Arabidopsis* revealed that there were significant differences in the behavior of circuit components between plants and/or between transient versus stable transgenic expression. This provided the authors with a further chance to apply their optimization cycle before ultimately achieving the expected reporter expression patterns.

Finally, the authors wanted to showcase how the logic gates could help engineer development. In these experiments, they took advantage of the dominant *solitary root* (*slr-1*) mutant that lacks secondary (or lateral) roots. The *slr-1* allele was expressed as the output of a BUFFER gate. Using BUFFER gates of varying strengths, the authors were able to create a series of plants with lateral root density ranging from zero (strong *slr-1* expression) to approaching wild-type architecture (weakest *slr-1* expression). Moreover, by using a lateral root-specific promoter to drive the synthetic transcription factors acting as inputs, they were able to minimize pleiotropic effects such as reduced gravity sensing. The tunability of an agriculturally relevant phenotype is a model for future engineering efforts and provides an experimental tool for quantifying dosage effects of any gene, or variant, of interest.

This report, along with several other recent publications exploring synthetic signaling,[10–13] ushers in a new era for engineering plant biology. Beyond new tools, this study demonstrates the fragility of synthetic systems and how many challenges still exist in building the large synthetic circuits of our dreams. It is also another great proof of the power of the Design–Build–Test–Learn cycle applied to biology. Not only did this approach ultimately lead to functional circuits, but it also yielded generalizable strategies for tuning parts and helped generate hypotheses for why some designs failed. These hard-earned lessons will no doubt help accelerate progress toward the smart plants of the future.

References

1. Irish V. The ABC model of floral development. Curr Biol 2017;27(17):R887–R890; doi: 10.1016/j.cub.2017.03.045
2. Broekgaarden C, Caarls L, Vos IA, et al. Ethylene: Traffic controller on hormonal crossroads to defense. Plant Physiol 2015;169(4):2371–2379; doi: 10.1104/pp.15.01020
3. Wang B, Barahona M, Buck M. A modular cell-based biosensor using engineered genetic logic circuits to detect and integrate multiple environmental signals. Biosens Bioelectron 2013;40(1):368–376; doi: 10.1016/j.bios.2012.08.011
4. Hsiao V, Hori Y, Rothemund PW, et al. A population-based temporal logic gate for timing and recording chemical events. Mol Syst Biol 2016;12(5):869; doi: 10.15252/msb.20156663
5. Liu Y, Zeng Y, Liu L, et al. Synthesizing AND gate genetic circuits based on CRISPR-Cas9 for identification of bladder cancer cells. Nat Commun 2014;5(1):5393; doi: 10.1038/ncomms6393
6. Brophy JAN, Voigt CA. Principles of genetic circuit design. Nat Methods 2014;11(5):508–520; doi: 10.1038/nmeth.2926
7. Xiang Y, Dalchau N, Wang B. Scaling up genetic circuit design for cellular computing: Advances and prospects. Nat Comput 2018;17(4):833–853; doi: 10.1007/s11047-018-9715-9
8. Brophy JAN, Magallon KJ, Duan L, et al. Synthetic genetic circuits as a means of reprogramming plant roots. Science 2022;377(6607):747–751; doi: 10.1126/science.abo4326
9. Yaschenko AE, Fenech M, Mazzoni-Putman S, et al. Deciphering the molecular basis of tissue-specific gene expression in plants: Can synthetic biology help? Curr Opin Plant Biol 2022;68:102241; doi: 10.1016/j.pbi.2022.102241
10. Cai Y-M, Kallam K, Tidd H, et al. Rational design of minimal synthetic promoters for plants. Nucleic Acids Res 2020;48(21):11845–11856; doi: 10.1093/nar/gkaa682
11. Andreou AI, Nirkko J, Ochoa-Villarreal M, et al. Mobius assembly for plant systems highlights promoter-terminator interaction in gene regulation. bioRxiv 2021;2021.03.31.437819; doi: 10.1101/2021.03.31.437819
12. Belcher MS, Vuu KM, Zhou A, et al. Design of orthogonal regulatory systems for modulating gene expression in plants. Nat Chem Biol 2020;16(8):857–865; doi: 10.1038/s41589-020-0547-4
13. Moreno-Giménez E, Selma S, Calvache C, et al. GB_SynP: A modular dCas9-regulated synthetic promoter collection for fine-tuned recombinant gene expression in plants. ACS Synth Biol 2022;11(9):3037–3048; doi: 10.1021/acssynbio.2c00238

Chapter1: Building A pipeline to identify and engineer constitutive and repressible promoters

Eric J.Y. Yang and Jennifer L. Nemhauser*

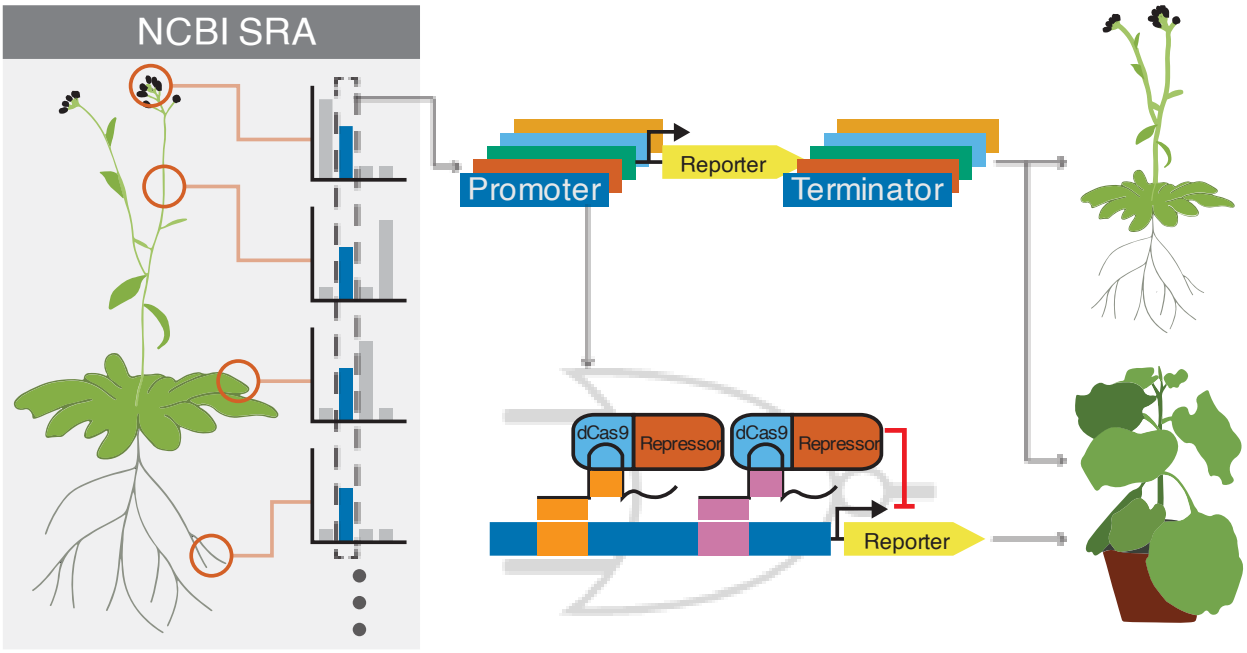
University of Washington, Department of Biology, Seattle, WA 98105-1800, USA

*email: jn7@uw.edu

Word Count: 5175

Keywords: Synthetic Biology, Boolean Logic, *Arabidopsis thaliana*, *Nicotiana benthamian*a

Abstract and Graphical Abstract



To support the increasingly complex circuits needed for plant synthetic biology applications, additional constitutive promoters are essential. Reusing promoter parts can lead to difficulty in cloning, increased heterogeneity between transformants, transgene silencing and trait instability. We have developed a pipeline to identify genes that have stable expression across a wide range of *Arabidopsis* tissues at different developmental stages, and have identified a number of promoters that are well expressed in both transient (*Nicotiana benthamiana*) and stable (*Arabidopsis*) transformation assays. We have also introduced two genome-orthogonal gRNA target-sites in a subset of the screened promoters, converting them into NOR logic gates. The work here establishes a pipeline to screen for additional constitutive promoters, and can form the basis of constructing more complex information processing circuits in the future.

Introduction

Plant synthetic biology aims to provide greater control over plant form and function, a goal that is beginning to be realized. Several projects have produced measurable gains in photosynthetic efficiency (Batista-Silva et al., 2020; Orr et al., 2017), and others have intervened in hormone response pathways to change plant architecture (Khakhar et al., 2018) or environmental response (Lim et al., 2020; Park et al., 2015). These advances rely on well-characterized promoters to ensure expression of transgene in desired tissues.

Promoters can be broadly broken down into three categories based on expression pattern: constitutive, spatiotemporally-restricted, and inducible (Peremarti et al., 2010). Constitutive promoters are defined here as promoters expressed in all tissues at all times. These promoters regulate the transcription of what are commonly referred to as "housekeeping genes". While each category of promoter is useful in plant engineering, constitutive promoters are often used to confer novel traits such as herbicide tolerance, to drive synthetic circuits, and used in metabolic engineering projects due to their broad tissue coverage (Bak & Emerson, 2020; Brophy et al., 2022; South et al., 2019). Some of the most widely used plant constitutive promoters include variants from the *Cauliflower Mosaic Virus* 35S (35S) promoter, and promoters from members of the ubiquitin and actin families (Jiang et al., 2018; Peremarti et al., 2010). However, the list of available plant constitutive promoters is short, and this lack of parts poses many challenges (Peremarti et al., 2010). Having to reuse the limited number of promoters in increasingly complex plant gene circuits or metabolic engineering projects can quickly lead to instability of the transformed construct due to repeated elements rearranging and homology-dependent gene silencing, which is heritable (De Wilde et al., 2000; Peremarti et al., 2010; Rajeev Kumar et al., 2015).

To expand the number of promoters available, several groups have recently used distinct strategies to engineer both constitutively and conditionally expressed promoters. One approach builds synthetic promoters by adding cis-elements to a "minimal promoter region", which is often 35S-derived. By varying the number and type of cis-elements, researchers were able to generate promoters with a wide range of expression levels and expression patterns (Ali & Kim, 2019; Belcher et al., 2020; Brophy et al., 2022; Liu & Stewart, 2016). Another approach uses sequences upstream of the minimal promoter region as a landing dock for synthetic activators guided by zinc-finger, TALE, or dCas9 to promote expression (Liu & Stewart, 2016). The expression strength of these promoters can be tuned by varying the number of target sites for the synthetic activators (Cai et al., 2020; Moreno-Giménez et al., 2022). These approaches, while quite powerful, are limited by the small number of characterized minimal promoters available to build upon and may still lead to repeated units in large constructs if the same minimal promoters were used.

Here, we employed an alternative approach for finding constitutive promoters. Instead of building and testing synthetic promoters, we looked to natural promoters found in the *Arabidopsis* genome. This approach has a few advantages. Synthetic promoters require extensive characterization to determine their expression pattern and, because of practical constraints, are often only tested in a few selected tissues. In contrast, the wealth of RNA-seq data available for *Arabidopsis* provides highly detailed information about a given promoter's likely expression potential, including the expression level of the gene throughout many developmental stages, tissue types, and even various growth and/or stress conditions. The expression of a native promoter has already been subject to selective pressures, and so is potentially more

likely to remain stable across generations. By introducing a set of unique sequences, natural promoters also have the potential to minimize the likelihood of gene silencing or unwanted recombination through repeated units in multigenic constructs. Lastly, by employing some of the techniques in generating synthetic promoters described above, these native promoters could potentially form the foundation for suites of derived promoters with even more refined expression levels. A similar approach successfully expanded the range of promoter expressions available in *B. subtilis* (Guiziou et al., 2016). Since the argument for the need of additional promoter parts can be directly extended to the need for additional terminators, and terminators are known regulators of gene expression (Andreou et al., 2021; P.-H. Wang et al., 2020), we screened promoter-terminator pairs together. To further extend the utility of the new promoter/terminator pairs, we also introduced dCas9 target-sites with sequences not found elsewhere in the *Arabidopsis* genome, thereby enabling specific repression by synthetic transcription factors without interfering with the cognate native genes.

Results

To identify the most stably expressing promoters available in the *Arabidopsis* genome, we analyzed publicly available RNAseq datasets. The majority of the RNAseq dataset came from the Klepikova transcriptome profile which included multiple tissues from different development stages (Klepikova et al., 2016). We supplemented this dataset with an RNAseq dataset for pollen (Loraine et al., 2013), as this cell-type was not represented in the Klepikova dataset. After processing the RNAseq datasets, there were 10,096 genes that were expressed in all the datasets (i.e. have none zero read counts) (Figure1A). Coefficient of variation (CV) of expression across different tissues is often used as a metric for identifying stably expressed genes (Czechowski et al., 2005; Huang et al., 2019; Z. Wang et al., 2019). Within the lowest 3% CV, there were 303 genes, which corresponds to a CV cutoff of 0.26 (Figure1A,D). To facilitate dissemination of the parts quantified in this study, we adopted the Golden Gate MoClo system and cloned the promoter + 5'UTR together as a standard MoClo part and similarly with 3'UTR + terminator (Figure1C) (Engler et al., 2014; Weber et al., 2011). Since MoClo uses BsaI and BbsI type-II restriction enzymes for cloning, we removed any candidates with these restriction sites within the cloned regions. These cloning constraints left us with 61 candidate genes.

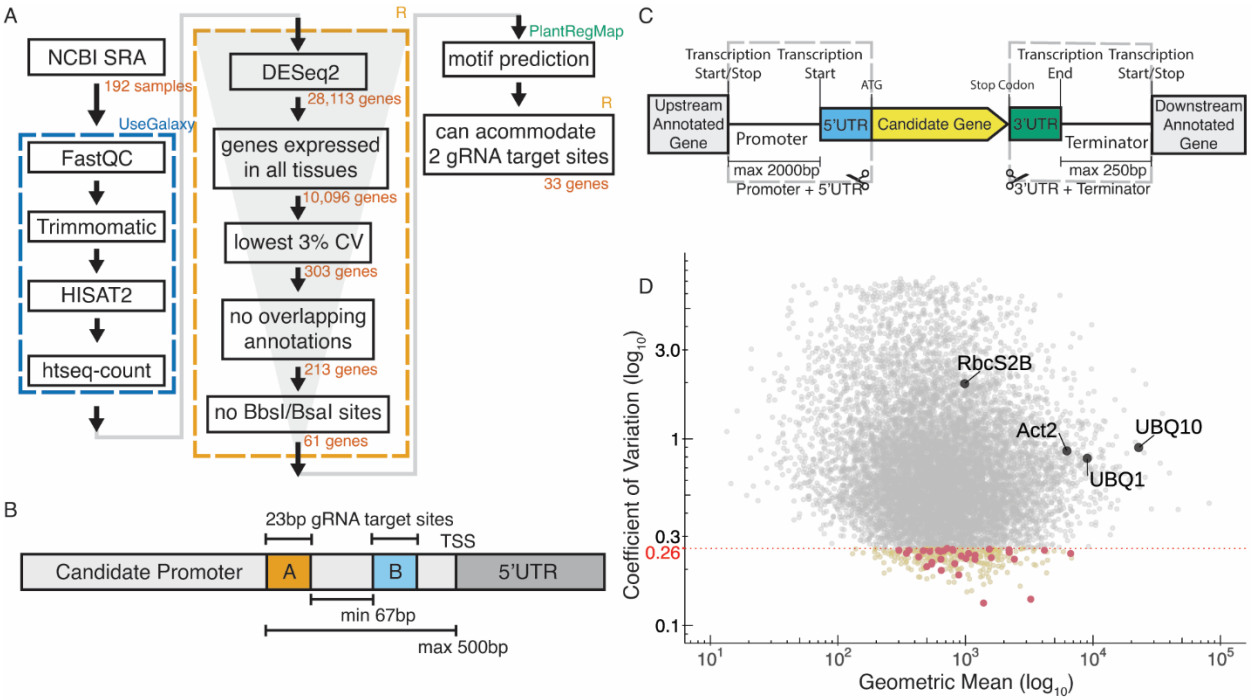


Figure1. A) Pipeline to identify constitutive promoters. The number of genes that pass each filter are indicated, along with the software used to implement the analysis. SRA is the "Sequence Read Archive". Detailed methods, including parameters for each filter, are described in the Methods section. B) Schematic of filters used to select candidate promoters to engineer with synthetic gRNA target-sites. C) Schematic describing how we defined "promoter" and "terminator". The "promoter" was defined here as starting from the transcription start site and going upstream to a maximum of 2000bp or to the next annotated neighboring gene, whichever is shorter. Similarly, a "terminator" was defined as starting from the transcription end site and going downstream a maximum of 250bp or to the next annotated neighboring gene, whichever is shorter. Promoters and terminators were cloned, along with their respective UTRs, following the Golden Gate MoClo system. D) Plot showing values for the 10,096 genes expressed in all tissues. The geometric mean of expression across samples is plotted on the x-axis with the coefficient of variation (CV) on the y-axis. Both axes are on a base-10 log scale. Lowest 3% CV corresponds to a 0.26 CV cutoff, and the 303 genes with CV lower than 0.26 are highlighted yellow. The final 33 candidates that fulfilled all criteria are highlighted in red. Several common promoters used in plant synthetic biology are annotated for reference.

To selectively activate or repress promoters in the context of a synthetic circuit, we wanted to modify segments of the promoter region to allow genome-orthogonal dCas9 targeting. While there are no specific guidelines on optimal placement for gRNA target-sites in plants (Pan et al., 2021), studies in other eukaryotes have pointed to -50 to +300bp from TSS in mammalian cells for CRISPRi, and within -200bp from TSS in yeast (Jensen, 2018). Using the "Binding Site Prediction" function from PlantRegMap we screened for predicted motifs within 500 bp of the promoter region from the TSS (Tian et al., 2020). We retained promoters that could accommodate two 23bp gRNA target-sites (20bp target sequence and 3bp PAM site) without interrupting any predicted motifs and were at least 67bp apart, following the spacing used in Gander et al. (Figure1B). We were left with 33 candidate genes. Compared to the commonly used native *Arabidopsis* constitutive promoters, the candidates identified here were more stably expressed but

have mostly weaker mean expression (Figure1D). Detailed information of the 33 candidates can be found in Supplementary Table S2.

While one of the main goals of this study is to identify the best available natural stable genes through analysis of RNAseq data, the "stability" of the candidate genes we screened for in this paper is constrained by the choice of RNAseq dataset used. The Klepikova dataset included stress-treated leaf samples with heat, cold, and wounding treatments, but they were not included in the CV calculation since the samples were only collected from mature third leaves and no other tissue types. Instead, we normalized the stress data with untreated "mature whole third leaf" and calculated their CV and included the result in Supplementary Table S2 for reference. Similarly, while the datasets capture coarse temporal resolutions throughout development, they cannot identify the fluctuation of circadian genes and therefore we supplemented the final table with identified circadian genes from CGDB for reference (Li et al., 2017).

Of the 33 stably expressed genes identified from the bioinformatics pipeline, we successfully cloned twenty-two promoter-terminator pairs. We tested the promoters in *Nicotiana benthamiana* (tobacco) transient agroinfiltration assays and identified sixteen promoters that had expressions that are significantly different from the negative control (Figure2).

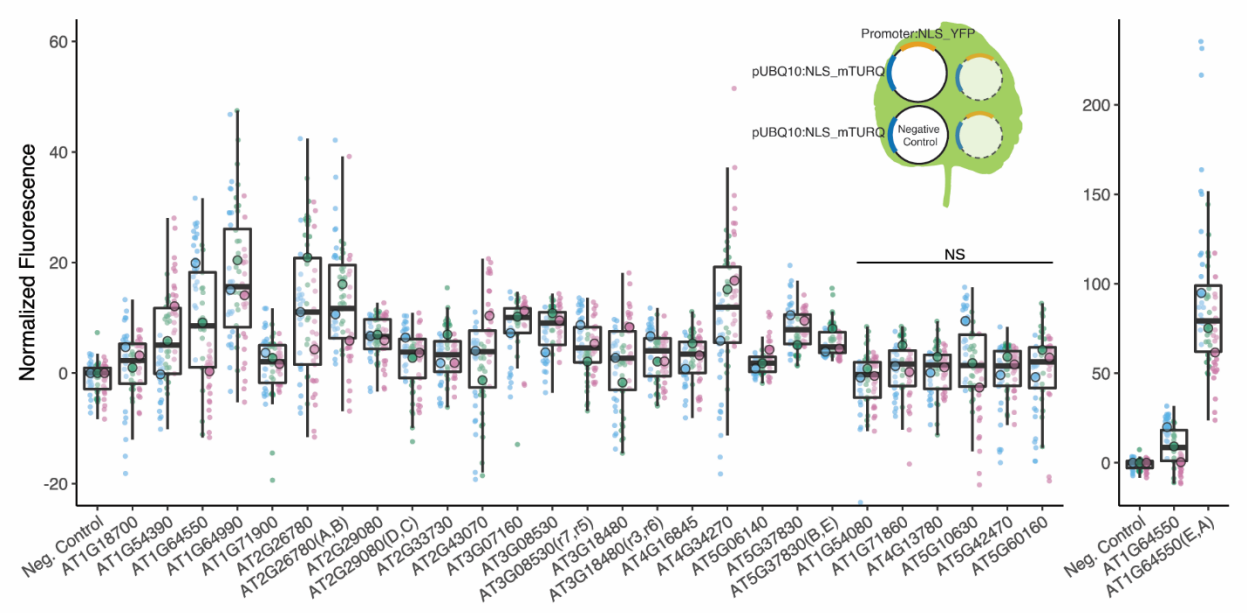


Figure2. We identified sixteen promoters that expressed in *N. benthamiana*. Six promoters were modified to introduce gRNA target sites. These sites are designated by brackets following the gene name. Three different constructs were injected per leaf, each containing a promoter to be tested driving NLS_YFP and an internal control of pUBQ10:NLS_mTURQ. Each leaf also has a negative control injection that only contains

pUBQ10:NLS_mTURQ. Normalization is performed using the formula: $\frac{YFP_{promoter} - median(YFP_{Neg.Control})}{mTURQ_{promoter}}$. For each construct, the three replicates with median fluorescence levels closest to the median of the group were selected for visualization and statistical analysis. Each biological replicate is represented by a beeswarm plot of 24 datapoints (12 per leaf disc, 2 disc per injection) collected from the plate reader as well as a single summarizing datapoint representing the median. The boxplots represent all biological replicates. Significance test was performed using Dunnett's test for comparing multiple treatments with control at 95% family-wise confidence level. Non-significant constructs are marked as NS. For a given construct, the colors signify datapoints derived from the same biological replicate.

To determine whether the promoters showed constitutive expression in *Arabidopsis*, twelve of the promoters were selected to drive expression of the RUBY reporter (He et al., 2020) in stable transformants. Since RUBY is a pigment that allows for simple visual readout, we were hoping it would be an effective way of evaluating the expression of the promoters in all the tissues throughout development. Three representative T1 lines were selected for each construct and six T2s per T1 line were observed at the seedling stage (12 days) and as mature plants (day 34). Eleven of the twelve promoters transformed showed expression in *N. benthamiana*, yet we only identified three promoters that displayed RUBY expression in *Arabidopsis.* Representative individuals are shown in Figure 3A (Supplementary Figure S3) with the intensity of RUBY coloring quantified in Fiji (Supplementary Figure S4)

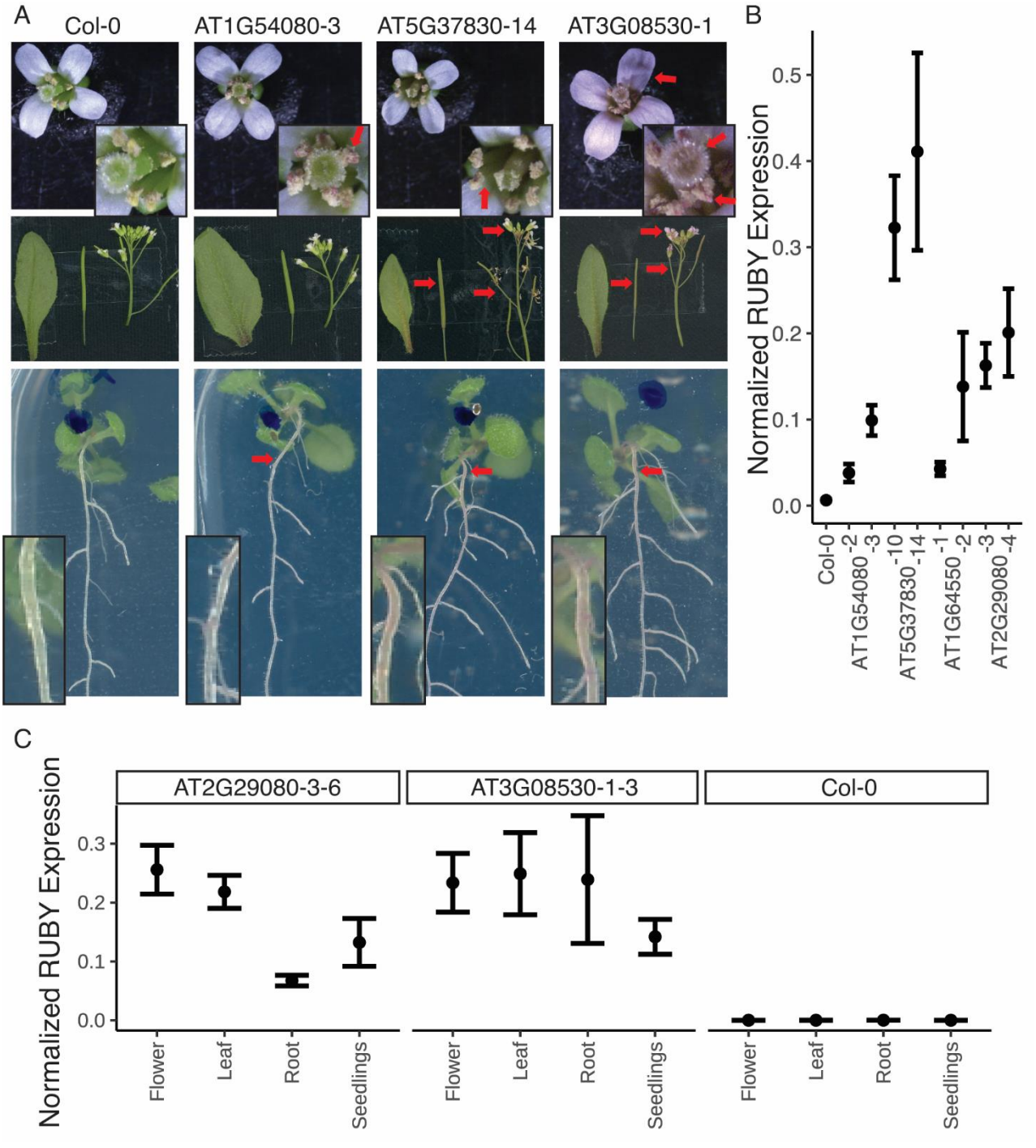


Figure3. A) Three promoters showed expression of RUBY in *Arabidopsis* T2 plants. The flowers, siliques, and leaves were imaged on day 34, while the seedling images were imaged on day 12. The inset boxes are the same images at higher magnification. Red arrows indicate areas where RUBY expression is visible by eye. B) qPCR data on T2 whole seedlings in three biological replicates for each line and C) qPCR data on tissues collected from T3 plants, each with three biological replicates and two technical replicates. RUBY expression was normalized against the reference gene PP2AA3, and the bars represent the mean expression of the RUBY reporter and the standard error of the mean (SEM).

Given that expression in *N. benthamiana* doesn't perfectly predict expression in *Arabidopsis*, we included two promoters (AT1G54080, AT1G71860) that did not show expression in tobacco infiltration in our *Arabidopsis* stable transformation experiment. Interestingly, AT1G54080 displayed RUBY expression in roots and pollen. AT5G37830 had visible expression in pollen, siliques, stems, and roots. AT3G08530 had the most ubiquitous expression and had visible expression in the flowers, pollen, siliques, stems, and roots. A visual summary of the *Arabidopsis* and *N. benthamiana* experiments can be found in Supplementary Figure S5.

Given that the majority of the promoters had no visible expression of RUBY by eye, we performed qPCR on whole seedlings from four promoter lines: two with visible RUBY expression in roots (AT1G54080 and AT5G37830) and two without (AT1G64550 and AT2G29080). The two lines without visible RUBY expression both had qPCR expression level between the two lines with visible RUBY expression in the roots. This result suggests that RUBY was not a reliable reporter for these low expressing promoters, and that the promoters were indeed functional in *Arabidopsis* (Figure3B). To further confirm whether the promoters were truly constitutively expressed, we chose our strongest expressing line (Figure3A) and one of the lines that did not appear red but had detectable expression by qPCR (Figure3B) for more careful qPCR analysis on seedlings, adult roots, flowers, and leaves (Figure3C). The result confirms that RUBY expression was detected in all the tissues analyzed, even when the tissues might not appear red by visual inspection. An interesting observation from the qPCR experiments was that the expression level of RUBY mRNA detected through qPCR is weaker than expected from the RNAseq dataset. While the predicted expression level of all four of the genes in the qPCR experiment are higher than the reference gene PP2AA3, the measured result showed the opposite (Supplementary Table S7). This discrepancy could be attributed to the RUBY reporter or potential limitations in identifying additional transcriptional regulators (see discussion).

To make the promoters screened in this experiment more versatile, we next introduced two gRNA target-sites into six of the promoters screened with target-site sequences not found in the *Arabidopsis* genome. A constitutive promoter with two unique gRNA target-sites can function as a NOR gate (a two-input logic gate where the output is only ON when neither of the inputs are present) in the presence of a dCas9-guided repressor. The inputs for such a gate are the gRNAs. When either or both of the gRNAs are present, the dCas9-guided repressor should be able to keep the promoter OFF (Figure4B). Only when neither of the guides are present can the promoter be turned ON. Nine different functional gRNA sequences (A-I) were selected from literature (Supplementary Table S8).

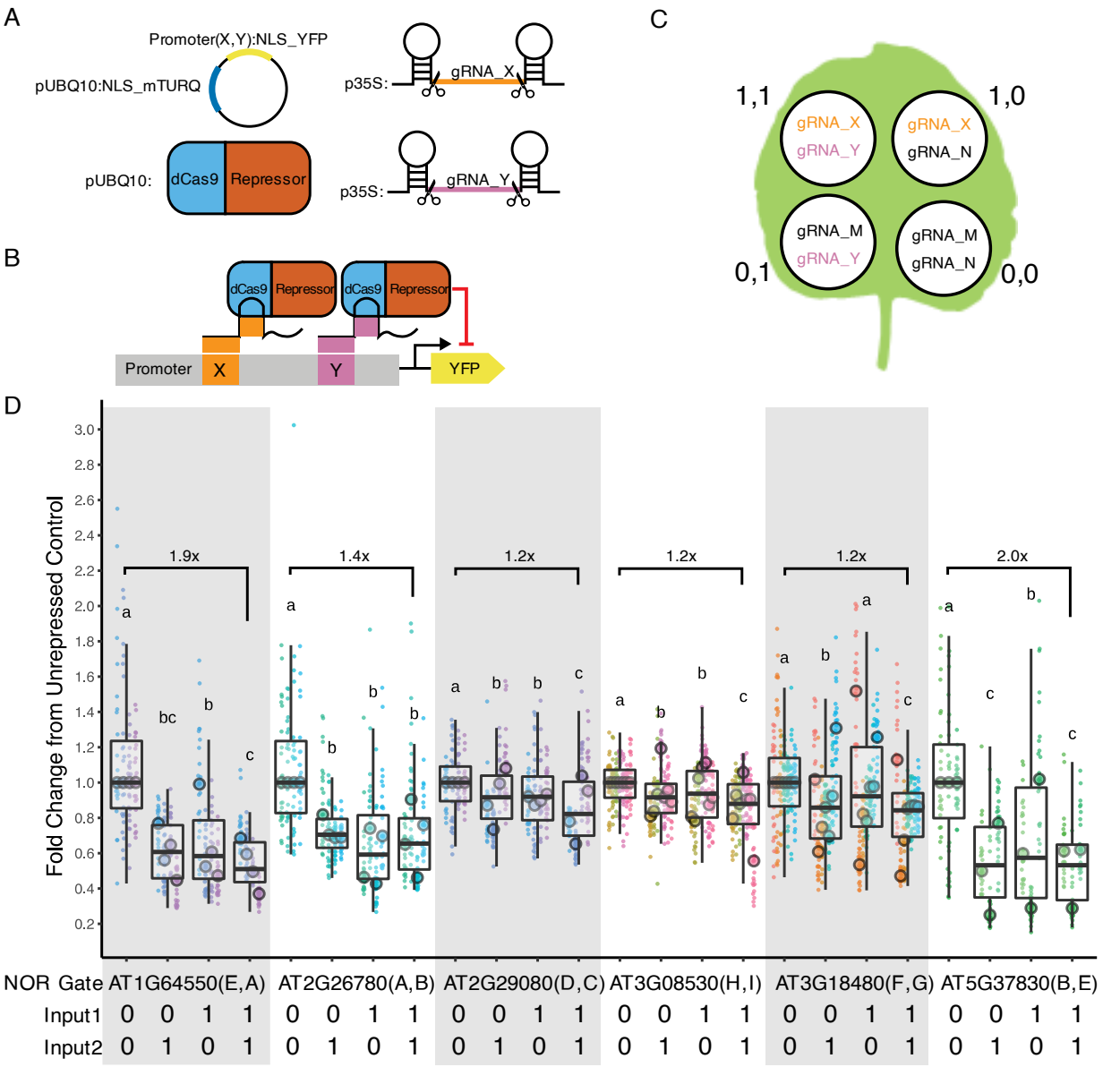


Figure4. A) The four constructs co-injected for each injection. The injection always contains the mPromoter, dCas9-guided repressor, and the two self-cleaving input gRNAs. The gRNAs are denoted with X and Y representing a variable input. B) Schematic of the NOR gate when both input gRNAs are present. C) Pattern of injection for the four possible input combinations and the gRNAs used for each injection. (1,1) represents both guides are present while (0,0) represents neither are present. When a guide is not present, a non-matching gRNA is injected in its place, denoted here as gRNA_M or N. D) Five of the six mPromoters functioned as NOR gates. All guides apart from gRNA_F are independently repressible. Each biological replicate is represented by a beeswarm plot of 24 datapoints (12 per leaf disc, 2 disc per injection) collected from the plate reader as well as a single summarizing datapoint representing the median. The boxplots represent all biological replicates. The signal is measured as the YFP fluorescence (driven by the promoters being tested) divided by the mTURQ fluorescence (driven by pUBQ10). In each set of NOR gate injections, the (0,0) injection serves as the unrepressed control, and the dataset is normalized by dividing all values by the median of the unrepressed control on a per-leaf basis. The y-axis represents fold changes from the unrepressed control and each biological replicate of the control is centered on 1. Each color

represents a unique leaf. Letters above each boxplot are Compact Letter Display (CLD) for all pairwise comparisons within each set of injections using ANOVA followed by Tukey's Honest Significant Difference Test. Numbers above the boxplot represent fold repression between the (0,0) and (1,1) injections.

We first confirmed that the introduction of the target-sites did not abolish promoter expression (Figure2). While in most cases there was little difference in expression between modified and native promoters, in one case [AT1G64550(E,A)], the expression level increased dramatically, possibly due to the introduction of new TF binding sites at the junction of the introduced gRNA target-site (Supplementary Figure S6). The repressibility of the modified promoters were tested in *N. benthamiana* with each infiltration containing all constructs required for repression (Figure2A). Each set of experiment contains four possible input combinations for each repressible promoter and the extent of repression was evaluated against the non-repressed control using two non-matching gRNAs (Figure2C,D). Five of the modified promoters (mPromoters) functioned as NOR gates while AT3G18480(F,G) acted as a NOT gate with input2 (gRNA_G) (Figure4D). Of the NOR gates, AT2G26780(A,B) repressed to similar extents with either or both inputs. AT1G64550(E,A), AT2G29080(D,C) and AT3G08530(H,I) all displayed additive effects where having both inputs gave stronger repression than just having one alone. AT5G37830(B,E) had a well repressed target-site with input2 (gRNA_E) while input1 (gRNA_B) alone resulted in a weaker repressed state. The strongest repression was observed for AT5G37830(B,E) and AT1G64550(E,A) with about a two-fold repression between input(1,1) and input(0,0), while the rest of the promoters had around a 1.2-fold repression. The repression strength observed in the assay is quite modest, and it is likely because the promoters are quite weak to begin with, making strong repression more difficult. The result displayed as normalized fluorescence and not as fold repression can be found in Supplementary Figure S9.

Discussion

Constitutive promoters are essential staples in stocking the synthetic biology toolbox. They are versatile due to their wide expression coverage, and form the foundation from which many synthetic promoters are built. Here, we report on the establishment of a pipeline to find the most stably expressing promoters in *Arabidopsis*. We successfully used this approach to identify sixteen promoters that are predicted to be more stably expressed than some of the most widely used native plant constitutive promoters, and showed they can drive expression in transient transformations of *N. benthamiana*. We attempted to capture the expression pattern of these promoters in stably transformed *Arabidopsis* using the visual RUBY reporter, and uncovered limitations in its utility, but we identified at least two promoters that showed expression in all the tissues tested throughout development via qPCR. Lastly, we engineered repressible versions of six promoters and showed that five of these can function as NOR logic gates.

One of the biggest challenges in having a small selection of promoters to choose from is the need to reuse promoters in larger constructs, which could pose challenges to long term stability. The promoters identified in this paper were selected from some of the most stably expressed genes available in the *Arabidopsis* genome and all have distinct sequences. A lack of promoter parts also means a lack of flexibility when it comes to the range of expression strength. Most of the promoters used in plant synthetic biology are quite strong and that is not ideal for every application. The availability of weaker, broadly expressed promoters like those characterized here allows more flexibility in promoter choices when excess production of target proteins can be a problem. For example, they can be beneficial in avoiding toxic intermediates or optimizing flux in metabolic engineering projects (Brückner et al., 2015; Patron, 2020). If a minimal promoter sequence can be identified from these native promoters, they can also serve as the foundation of additional synthetic promoters where the expression pattern and strength can be freely modified by adding cis-elements or synthetic transcription factor binding sites. The pipeline employed in this paper to arrive at new native constitutive promoters should be readily adaptable to other organisms if there is sufficiently broad sampling of transcriptomes and a reference genome. The pipeline could also be modified to identify native promoters with particular expression patterns. One caveat is that the promoters that can be extracted in this way are, by definition, limited by what is naturally available in the organism. On the other hand, they have the advantage of already being assayed in a whole range of tissue types and developmental stages—a breadth of information that can be logistically challenging to collect for synthetic promoters. It will be interesting to see if synthetic devices made with these modified native promoters prove more resilient to mutation than those using fully engineered promoters, as these sequences have presumably maintained stable expression in the face of mutation and selection.

Evaluating the promoters using RUBY revealed that the novel reporter had limited sensitivity when driven by weaker promoters. We were able to detect RUBY expression in seedlings and adult tissues without visible coloration using qPCR, a more sensitive assay. However, it is important to note that detecting transcripts doesn't always imply comparable levels of protein production due to post-transcriptional and post-translational regulation. In our design, we attempted to capture the effects of any post-transcriptional regulation by including the UTRs, but other potential transcriptional regulators could be missed. The lower-than-expected RUBY mRNA levels detected could be due to such regulators. Promoter-proximal introns after the translation start codon, for example, would not be captured in the cloning pipeline though it is known to contribute to gene expression (Rose, 2019; Rose et al., 2008). Distally located regulatory regions would also not be captured, but they should be rare in the compact genome of *Arabidopsis* (Galli et al., 2020; Lu et al., 2019).

Working with native promoters also provided an opportunity to learn more about the biology of promoters themselves. Yamamoto and colleagues had suggested that plant promoters can be grouped into

a few core promoter categories based on the presence or absence of certain location-sensitive motifs (Yamamoto et al., 2011). Interestingly, they reported that TATA-box containing promoters tend to be regulated promoters while Coreless promoters (promoters that don't have any characteristic location-sensitive motifs) tend to be constitutively expressed. The vast majority of the constitutive promoters used today in plant synthetic biology are from the TATA promoter class, and we also have a much better understanding of how their expression is regulated (Cai et al., 2020). If the goal is to find constitutive promoters, however, the analysis by Yamamoto and colleagues would suggest that we should look to Coreless promoters instead. Indeed, only 9% (3/33) of the candidate genes identified in this study contain TATA boxes, while 45% (15/33) are Coreless (Supplementary Table S2).

The ability to selectively activate and repress genes provides the tools necessary to perform Boolean logic, which would allow more complex computations (Kassaw et al., 2018). Plants naturally perform complex computations to determine when and where a gene should be expressed by integrating internal and external signals, and genetic logic gates provide a modular way to synthetically construct these input-output relationships by using simple genetic parts. There are many ways to achieve the different logic operations using molecular biology (Patron, 2020). A NOR gate is powerful in that it can be used to construct any logic gate by just stringing together multiple NOR gates, and its efficacy had been demonstrated in yeast (Gander et al., 2017). To date, the feasibility of building more complex logic circuits in plants has been hindered by the lack of unique and strongly repressible promoter parts. With just our design constraints and no additional refinement, five of the six NOR gates built showed the correct behavior, suggesting the pipeline used holds promise in identifying additional promoter candidates for engineering. The repressible promoters evaluated in this work can be further improved through additional design-build-test cycles to optimize the individual gRNA target sites by varying their position and sequence. The repressor design can also be potentially improved upon to lower the overall OFF state. While *N. benthamiana* serves as a great prototyping platform, the performance of the gates would also need to be evaluated in stable *Arabidopsis* lines to validate their viability. The pipeline and the repressible promoter screened in this work contributes to the construction of more complex, synthetic plant logic operations in the future.

Methods

Downloading and processing RNA-seq datasets

We used a custom UseGalaxy pipeline to process the RNA-seq datasets (Afgan et al., 2016). SRR accession codes from BioProject IDs PRJNA314076 (138 samples; Klepikova et al., 2016), PRJNA268115 (20 samples; Klepikova et al. 2015), PRJNA324514 (32 samples), PRJNA194429 (2 samples; Loraine et al., 2013) were input into "Faster Download and Extract Reads in FASTQ (Galaxy Version 2.10.8+galaxy0)" with default settings. The FASTQ files were pipped into "FastQC (Galaxy Version 0.73+galaxy0)" and "Trimmomatic (Galaxy Version 0.38.0)" with sliding window trimming averaging across 4 bases with required average quality 20, and a minimum read length of 36. The trimmed files were input into "HISAT2 (Galaxy Version 2.1.0+galaxy5)" with reference genome assembly TAIR10 and Araport11 genome annotation from The *Arabidopsis* Information Resource (TAIR). Minimum intron length was set to 60, and maximum intron length was set to 6000 (Marquez et al., 2012). Features from the Araport11 annotation were counted with "htseq-count (Galaxy Version 0.9.1)" set to Union Mode and counting only reads within regions defined as "exons" in the Araport11 annotation while not counting non-unique/ambiguous reads (Klepikova et al., 2016). The counted features were downloaded, and subsequent analysis was done in R (R Core Team, 2022).

Identifying stable promoters

All samples excluding the stress dataset (PRJNA324514) were normalized using the Median Ratios method from the DESeq2 package in R (Love et al., 2014). Coefficient of variation (CV) for each gene was calculated from the normalized data. Genes with the lowest 3% CV were kept for further analysis. Stress dataset from PRJNA324514 was normalized with "mature whole third leaf" from PRJNA314076 for CV calculation, separate from the rest of the data.

Extracting promoter and terminator sequences

Promoter+5'UTR region (from before the start codon and extending upstream till the first annotated neighboring gene or to a maximum of 2kb from the transcription start site, whichever is shorter) and 3'UTR+terminators (from after the stop codon and extending downstream till the first annotated neighboring gene or to a maximum of 250bp past the transcription end site, whichever is shorter) of the remaining genes were extracted using the Araport11 genome annotation and the "3000bp upstream and downstream" sequence files from the TAIR website. The extracted sequences were screened for BbsI and BsaI restriction enzyme cut sites and only those without were kept. Any genes with their promoter+5'UTR and 3'UTR+terminator overlapping annotations from neighboring genes in the Araport11 annotation were also removed.

Transcription factor binding site prediction

The promoter sequences of the remaining genes were uploaded onto PlantRegMap using the "Binding Site Prediction" function (Tian et al., 2020) and the predicted motifs for each promoter sequence were downloaded. Only genes that can fit two 23bp gRNA target sites at least 67bp apart without interrupting any of the predicted motifs while being within 500bp of the TSS were kept.

Annotating candidate genes

For the final 33 candidate genes, CV for the Stress Dataset (StressCV) and promoter and terminator sequences were extracted as described above. The promoter core type was annotated from Tokizawa et al. 2017. A list of experimentally determined circadian genes in *Arabidopsis* was downloaded from CGDB (Li et al., 2017), and any UniprotKB identifiers were converted to ATG identifiers with the Uniprot Retrieve/ID mapping tool. Gene Descriptions (Representative Gene Model Name, Gene Description, Gene Model Type, Primary Gene Symbol, and All Gene Symbols) were retrieved from TAIR.

Construction of plasmids

Promoter+5'UTR and 3'UTR+terminator for each candidate genes as defined above were cloned with PCR from extracted genomic Col-0 DNA into their respective MoClo level zero acceptors (pICH41295 and pICH41276 respectively) (Engler et al., 2014). The promoter and terminator pair of the candidate genes were paired with nuclear localized Venus to make level one constructs in "position one". Venus level one constructs were paired with pUBQ10 promoter driving nuclear localized mTURQ with an Act2 terminator from the MoClo Plant Parts Kit (pICH44300) in "position two" to form ratio-metric lvl2s with a binary Ti vector backbone (pAGM4673 or pAGM4723). RUBY from (He et al., 2020) was cloned into level zero constructs and then cloned directly into level-2 Ti vector backbone with the promoter and terminator pairs (pICH86966). List of primers and plasmid maps can be found in Supplementary Table S10 and S11 and Genbank files can be found in Supplementary Data S13.

gRNA target-site introduction

gRNA target-sites were cloned into regions that do not disrupt any predicted TF binding sites (as described above) through Gibson assembly by replacing the original sequence (Gibson et al., 2009). Primers can be found in Supplementary Table S10.

Agrobacterium infiltration

5mL cultures of *Agrobacterium* containing constructs to be injected along with a separate 25mL culture of P19 (Win and Kamoun 2004) were grown overnight at 30C with the appropriate antibiotics. On the following day, the overnights were centrifuged at 3000xg for 10 minutes. The pellets were resuspended

with 1mL MMA (10 mM MgCl2, 10mM MES (pH 5.6), 100uM acetosyringone). The OD of the cultures were measured and about 1~2mL volume mixture with 5.0 OD for construct to be tested and 5.0 OD for P19 were prepared. The infiltration mix were rotated to mix at room temperature for 3hr before injecting into fully emerged *N. benthamiana* leaves with a 1mL syringe. The injections were always injected as triplicates on three separate leaves on three separate tobacco plants. Each leaf is also always injected with a pUBQ10:mTURQ control.

Fluorescence quantification in N. benthamiana

At 3 days post infiltration, the leaves were clipped off and visualized in the Azure C600 Western Blot imaging system with exposure times Cy5=0sec, Cy3=15sec, Cy2=5sec. Two hole-punches were taken out of representative regions of each injection, and damaged regions with high background fluorescence were avoided. The leaf discs were placed in a 96-well plate on top of 200uL of water, and the plates were read with a TECAN SPARK plate reader with YFP: excitation 506(15) and emission 541(15), Gain 100. mTurq: excitation 430(15) and emission 480(15), Gain 50. mScarlet: excitation 565(15) and emission 600(15), Gain 100. Settings: Multiple Reads Per Well; Circle (Filled) 4x4 with border 800uM. Each leaf disc was read 12 times giving a total of 24 datapoints per injection per leaf. The output data was read into a custom R file for annotation, clean-up, and visualization. Multiple biological replicates were assayed for each promoter being tested, and the three replicates closest to the median of all replicates were kept for visualization and statistical analysis. Each injection's YFP value is subtracted by the median YFP value of the pUBQ10:mTURQ negative control on the same leaf. The YFP value is then divided by the mTURQ value for each injection to normalize across results. A Dunnett Test, a post hoc pairwise multiple comparison test from the DescTools package, was used to determine whether the injections were significantly different from the negative control (Signorell et al., 2022).

Repression assays

Repression assays were performed using the modified promoters (mPromoters) with gRNA target-sites driving NLS-YFP and pUBQ10:NLS-mTURQ internal control as the reporter. The mPromoters (5OD) were co-injected with P19 (1OD), TPL repressor (1OD), self-cleaving gRNA_1 (1OD), and self-cleaving gRNA_2 (1OD). To test the mPromoters' NOR gate functionality, two gRNA inputs were required. In cases where only one input is present, the other self-cleaving gRNA will be a non-matching guide to the mPromoter. When neither inputs are present, sometimes two non-matching gRNAs (1OD each) were co-injected and sometimes only one (2OD), but the final total OD were always consistent. The exact injection combinations can be found in the R script. The TPL repressor construct contains pUBQ1:tdTomato-pUBQ10:dCas9_TPL(N188) and is modified from Khakhar et al. 2018. Self-cleaving gRNAs were designed in accordance to Zhang et al. 2017, and the modifications (gRNA and

complementary sequences) were introduced in one step using Q5 mutagenesis (NEB) and was placed in the MoClo pICH86988 acceptor with a 35S promoter. The four possible input combinations for the NOR gate for each promoter were always injected on the same leaf, and the result was read with a plate reader as described above. The YFP value of each injection was divided by the mTURQ value to normalize the data, and the value of each injection was divided by the median of the no-input control. An ANOVA followed by a Tukey's Honest Significant Difference Test was used to determine significant differences in expression between samples. List of plasmid maps used can be found in Supplementary Table S11 and Genbank files in Supplementary Data S13.

RUBY expression in Arabidopsis

Constructs with the candidate promoters driving RUBY were transformed into Col-0 through floral dipping method (Clough & Bent, 1998). T1 seeds were selected on 0.5x LS+50ug/mL Kanamycin+0.8% bactoagar. Plates were stratified for 2 days, light pulsed for 6 hours then kept in the dark for 3 days. Resistant seedlings were transplanted to soil to collect T2 seeds. For each promoter lines, three representative T1 lines were chosen to have their T2 seedlings phenotyped, and for each line, 19 T2 seeds were plated on 120x120x17 square petri dishes with 0.5x LS+0.8% bactoagar without selection. The plates were imaged on day 4, 8, and 12 post germination, and six representative seedlings were transplanted to soil. The plants were imaged with a digital camera on day 34. The flowers were imaged under a Leica S8AP0 dissecting scope. A representative leaf, a segment of the inflorescence, and silique were placed between two clear projector sheets and scanned with a flatbed scanner.

RUBY redness quantification

Images of were loaded into Fiji (Schindelin et al., 2012), and then converted to Lab stack to isolate the a* stack . The default color of the extracted stack was green, so the image was further converted to an RGB stack so that the Green-channel could be used for region of interest (ROI) quantification.

qPCR

T2 seedlings were grown vertically on 0.5x LS+0.8% Phytoagar and without selection. The plates were stratified at 4ºC for 2 days. On day 12, approximately five seedlings per replicate were frozen in liquid nitrogen. Three biological replicates were prepared for each genotype. Col-0 seedlings were also collected on day 12 as a single biological replicate. For T3 tissues, seedlings were grown on vertical 0.5xLS+0.8% Phytoagar plates without selection. On day 10, three sets of four seedlings were collected for each line and frozen in liquid nitrogen. Four seedlings per line were transplanted to fresh LS+Phytoagar plates to collect older roots from, and the rest of the seedlings were transplanted to soil. On day 22, three whole

roots were collected from plants on plates and the tissues were frozen. The entire inflorescence and one leaf from three plants on soil were collected for each line between days 27 and 31. The tissues were collected in 2mL tubes and powdered with a metal bead using a Retsch MM400 shaker after freezing the samples in liquid nitrogen. RNA was purified using an Illustra RNAspin Mini Kit (GE Healthcare). 1µg of extracted RNA was then used with the iScript cDNA synthesis kit (BIO-RAD). qPCR was performed using the iQ SYBR Green Supermix (BIO-RAD). PP2AA3 were used as a reference gene and the primers for PP2AA3 and RUBY can be found in Supplementary Table S1. The standard curves were established using a pool of all the cDNAs. T2 RUBY seedlings were run with three biological replicates per line while Col-0 seedlings were run as four technical replicates. T3 tissues were run with three biological replicates per tissue and two technical replicates per line. The qPCR were performed on a C1000 Thermal Cycler (BIO-RAD) and the result were read using the Bio-Rad CFX Maestro software and analyzed using standard methods (Pfaffl, 2001).

Acknowledgements

We thank Wesley George, Cassandra Maranas, Dr. Román Ramos Báez, Dr. Sarah Guiziou and Dr. Alexander Leydon for careful reading of the manuscript, as well as other members of the Nemhauser, Imaizumi, and Steinbrenner labs for their feedback on this project. We thank Dr. Nicholas J. Provart for his help with the RNA-seq datasets.

Author Contributions

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

Financial Support

This work was supported by the National Institute of Health (R01- GM107084), the National Science Foundation (IOS-1546873), and a Faculty Scholar Award from the Howard Hughes Medical Institute.

Conflicts of Interest

Authors declare no competing interests.

Data and Coding Availability Statement

The codes and datasets used in this study can be found on Github at https://github.com/Nemhauserlab/StablePromoters, and on Zenodo with DOI: 10.5281/zenodo.8170303. The repositories contain all the raw data as well as scripts to annotate, normalize, and generate the figures used in the article. Datasets before annotation and annotated data before normalization are both available. To minimize the supplemental file size, scripts without the datasets can be found in Supplementary Data S12.

References

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., & Goecks, J. (2016). The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, *44*(Web Server issue), W3–W10. https://doi.org/10.1093/nar/gkw343

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

Andreou, A. I., Nirkko, J., Ochoa-Villarreal, M., & Nakayama, N. (2021). *Mobius Assembly for Plant Systems highlights promoter-terminator interaction in gene regulation* (p. 2021.03.31.437819). bioRxiv. https://doi.org/10.1101/2021.03.31.437819

Bak, A., & Emerson, J. B. (2020). Cauliflower mosaic virus (CaMV) Biology, Management, and Relevance to GM Plant Detection for Sustainable Organic Agriculture. *Frontiers in Sustainable Food Systems*, *4*. https://www.frontiersin.org/articles/10.3389/fsufs.2020.00021

Batista-Silva, W., da Fonseca-Pereira, P., Martins, A. O., Zsögön, A., Nunes-Nesi, A., & Araújo, W. L. (2020). Engineering Improved Photosynthesis in the Era of Synthetic Biology. *Plant Communications*, *1*(2), 100032. https://doi.org/10.1016/j.xplc.2020.100032

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

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

Clough, S. J., & Bent, A. F. (1998). Floral dip: A simplified method for Agrobacterium -mediated transformation of Arabidopsis thaliana. *The Plant Journal*, *16*(6), 735–743. https://doi.org/10.1046/j.1365-313x.1998.00343.x

Czechowski, T., Stitt, M., Altmann, T., Udvardi, M. K., & Scheible, W.-R. (2005). Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis. *Plant Physiology*, *139*(1), 5–17. https://doi.org/10.1104/pp.105.063743

De Wilde, C., Van Houdt, H., De Buck, S., Angenon, G., De Jaeger, G., & Depicker, A. (2000). Plants as bioreactors for protein production: Avoiding the problem of transgene silencing. *Plant Molecular Biology*, *43*(2), 347–359. https://doi.org/10.1023/A:1006464304199

Engler, C., Youles, M., Gruetzner, R., Ehnert, T.-M., Werner, S., Jones, J. D. G., Patron, N. J., & Marillonnet, S. (2014). A Golden Gate Modular Cloning Toolbox for Plants. *ACS Synthetic Biology*, *3*(11), 839–843. https://doi.org/10.1021/sb4001504

Galli, M., Feng, F., & Gallavotti, A. (2020). Mapping Regulatory Determinants in Plants. *Frontiers in Genetics*, *11*, 591194. https://doi.org/10.3389/fgene.2020.591194

Gander, M. W., Vrana, J. D., Voje, W. E., Carothers, J. M., & Klavins, E. (2017). Digital logic circuits in yeast with CRISPR-dCas9 NOR gates. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15459

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, *6*(5), Article 5. https://doi.org/10.1038/nmeth.1318

Guiziou, S., Sauveplane, V., Chang, H.-J., Clerté, C., Declerck, N., Jules, M., & Bonnet, J. (2016). A part toolbox to tune genetic expression in Bacillus subtilis. *Nucleic Acids Research*, *44*(15), 7495–7508. https://doi.org/10.1093/nar/gkw624

He, Y., Zhang, T., Sun, H., Zhan, H., & Zhao, Y. (2020). A reporter for noninvasively monitoring gene expression and plant transformation. *Horticulture Research*, *7*(1), Article 1. https://doi.org/10.1038/s41438-020-00390-1

Huang, X., Li, S., & Zhan, A. (2019). Genome-Wide Identification and Evaluation of New Reference Genes for Gene Expression Analysis Under Temperature and Salinity Stresses in Ciona savignyi. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00071

Jensen, M. K. (2018). Design principles for nuclease-deficient CRISPR-based transcriptional regulators. *FEMS Yeast Research*, *18*(4), foy039. https://doi.org/10.1093/femsyr/foy039

Jiang, P., Zhang, K., Ding, Z., He, Q., Li, W., Zhu, S., Cheng, W., Zhang, K., & Li, K. (2018). Characterization of a strong and constitutive promoter from the Arabidopsis serine carboxypeptidase-like gene AtSCPL30 as a potential tool for crop transgenic breeding. *BMC Biotechnology*, *18*(1), 59. https://doi.org/10.1186/s12896-018-0470-x

Kassaw, T. K., Donayre-Torres, A. J., Antunes, M. S., Morey, K. J., & Medford, J. I. (2018). Engineering synthetic regulatory circuits in plants. *Plant Science*, *273*, 13–22. https://doi.org/10.1016/j.plantsci.2018.04.005

554 Khakhar, A., Leydon, A. R., Lemmex, A. C., Klavins, E., & Nemhauser, J. L. (2018). Synthetic
555 hormone-responsive transcription factors can monitor and re-program plant
556 development. *ELife*, *7*, e34702. https://doi.org/10.7554/eLife.34702
557 Klepikova, A. V., Kasianov, A. S., Gerasimov, E. S., Logacheva, M. D., & Penin, A. A. (2016). A
558 high resolution map of the Arabidopsis thaliana developmental transcriptome based on
559 RNA-seq profiling. *The Plant Journal*, *88*(6), 1058–1070.
560 https://doi.org/10.1111/tpj.13312
561 Li, S., Shui, K., Zhang, Y., Lv, Y., Deng, W., Ullah, S., Zhang, L., & Xue, Y. (2017). CGDB: A
562 database of circadian genes in eukaryotes. *Nucleic Acids Research*, *45*(Database
563 issue), D397–D403. https://doi.org/10.1093/nar/gkw1028
564 Lim, S. D., Mayer, J. A., Yim, W. C., & Cushman, J. C. (2020). Plant tissue succulence
565 engineering improves water-use efficiency, water-deficit stress attenuation and salinity
566 tolerance in Arabidopsis. *The Plant Journal*, *103*(3), 1049–1072.
567 https://doi.org/10.1111/tpj.14783
568 Liu, W., & Stewart, C. N. (2016). Plant synthetic promoters and transcription factors. *Current*
569 *Opinion in Biotechnology*, *37*, 36–44. https://doi.org/10.1016/j.copbio.2015.10.001
570 Loraine, A. E., McCormick, S., Estrada, A., Patel, K., & Qin, P. (2013). RNA-Seq of Arabidopsis
571 Pollen Uncovers Novel Transcription and Alternative Splicing1[C][W][OA]. *Plant*
572 *Physiology*, *162*(2), 1092–1109. https://doi.org/10.1104/pp.112.211441
573 Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and
574 dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.
575 https://doi.org/10.1186/s13059-014-0550-8
576 Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019). The
577 prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature*
578 *Plants*, *5*(12), Article 12. https://doi.org/10.1038/s41477-019-0548-z

Marquez, Y., Brown, J. W. S., Simpson, C., Barta, A., & Kalyna, M. (2012). Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Research*, *22*(6), 1184–1195. https://doi.org/10.1101/gr.134106.111

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

Orr, D. J., Pereira, A. M., da Fonseca Pereira, P., Pereira-Lima, Í. A., Zsögön, A., & Araújo, W. L. (2017). Engineering photosynthesis: Progress and perspectives. *F1000Research*, *6*, 1891. https://doi.org/10.12688/f1000research.12181.1

Pan, C., Sretenovic, S., & Qi, Y. (2021). CRISPR/dCas-mediated transcriptional and epigenetic regulation in plants. *Current Opinion in Plant Biology*, *60*, 101980. https://doi.org/10.1016/j.pbi.2020.101980

Park, S.-Y., Peterson, F. C., Mosquna, A., Yao, J., Volkman, B. F., & Cutler, S. R. (2015). Agrochemical control of plant water use using engineered abscisic acid receptors. *Nature*, *520*(7548), Article 7548. https://doi.org/10.1038/nature14123

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

Peremarti, A., Twyman, R. M., Gómez-Galera, S., Naqvi, S., Farré, G., Sabalza, M., Miralpeix, B., Dashevskaya, S., Yuan, D., Ramessar, K., Christou, P., Zhu, C., Bassie, L., & Capell, T. (2010). Promoter diversity in multigene transformation. *Plant Molecular Biology*, *73*(4), 363–378. https://doi.org/10.1007/s11103-010-9628-1

Pfaffl, M. W. (2001). A new mathematical model for relative quantification in real-time RT–PCR. *Nucleic Acids Research*, *29*(9), e45.

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajeev Kumar, S., Anunanthini, P., & Ramalingam, S. (2015). Epigenetic silencing in transgenic plants. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00693

Rose, A. B. (2019). Introns as Gene Regulators: A Brick on the Accelerator. *Frontiers in Genetics*, *9*. https://www.frontiersin.org/articles/10.3389/fgene.2018.00672

Rose, A. B., Elfersi, T., Parra, G., & Korf, I. (2008). Promoter-Proximal Introns in Arabidopsis thaliana Are Enriched in Dispersed Signals that Elevate Gene Expression. *The Plant Cell*, *20*(3), 543–551. https://doi.org/10.1105/tpc.107.057190

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., Tinevez, J.-Y., White, D. J., Hartenstein, V., Eliceiri, K., Tomancak, P., & Cardona, A. (2012). Fiji: An open-source platform for biological-image analysis. *Nature Methods*, *9*(7), Article 7. https://doi.org/10.1038/nmeth.2019

Signorell, A., Aho, K., Alfons, A., Anderegg, N., & Aragon, T. (2022). *DescTools: Tools for Descriptive Statistics*. https://cran.r-project.org/package=DescTools

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

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., & Gao, G. (2020). PlantRegMap: Charting functional regulatory maps in plants. *Nucleic Acids Research*, *48*(D1), D1104–D1113. https://doi.org/10.1093/nar/gkz1020

Tokizawa, M., Kusunoki, K., Koyama, H., Kurotani, A., Sakurai, T., Suzuki, Y., Sakamoto, T., Kurata, T., & Yamamoto, Y. Y. (2017). Identification of Arabidopsis genic and non-genic promoters by paired-end sequencing of TSS tags. *The Plant Journal: For Cell and Molecular Biology*, *90*(3), 587–605. https://doi.org/10.1111/tpj.13511

Wang, P.-H., Kumar, S., Zeng, J., McEwan, R., Wright, T. R., & Gupta, M. (2020). Transcription Terminator-Mediated Enhancement in Transgene Expression in Maize: Preponderance

of the AUGAAU Motif Overlapping With Poly(A) Signals. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.570778

Wang, Z., Lyu, Z., Pan, L., Zeng, G., & Randhawa, P. (2019). Defining housekeeping genes suitable for RNA-seq analysis of the human allograft kidney biopsy tissue. *BMC Medical Genomics*, *12*(1), 86. https://doi.org/10.1186/s12920-019-0538-z

Weber, E., Engler, C., Gruetzner, R., Werner, S., & Marillonnet, S. (2011). A Modular Cloning System for Standardized Assembly of Multigene Constructs. *PLOS ONE*, *6*(2), e16765. https://doi.org/10.1371/journal.pone.0016765

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

Zhang, T., Gao, Y., Wang, R., & Zhao, Y. (2017). Production of Guide RNAs in vitro and in vivo for CRISPR Using Ribozymes and RNA Polymerase II Promoters. *Bio-Protocol*, *7*(4), e2148. https://doi.org/10.21769/BioProtoc.2148

Chapter1: Supplementary materials

Supplementary Table S1. qPCR primers for RUBY and PP2AA3

Gene	Forward Primer	Reverse Primer
RUBY	AAACAGGGCAAGCTCGTGTA	ATCCGCAGTGGGTGAGAAAG
PP2AA3	AACGTGGCCAAAATGATGC	AACCGCTTGGTCGACTATCG

Supplementary Table S2. Top 3% Candidate Genes

AGI is the *Arabidopsis* Genome Initiative locus code. Coefficient of variation (CV) and geometric mean (geom_mean) are calculated without the stress dataset while coefficient of variation for the stress dataset (StressCV) is given separately. The core promoter types (TATA, Ypatch, CA, GA, Coreless) are taken from Tokizawa et al. 2017, and "1" signifies that core type is predicted in the promoter. CGDB circadian genes are whether the promoter is found to be circadian regulated.

Supplementary Table S2

AGI	CV	geom_mean	StressCV	TATA	Ypatch	CA	GA	Coreless	CGDB Circadian	Representative Gene Model Name
AT1G11060	0.206645	497.6316	0.202739	0	0	0	0	1	N	AT1G11060.1
AT1G18700	0.18655	886.1216	0.476444	0	0	0	0	1	Y	AT1G18700.5
AT1G54080	0.138365	3252.131	0.196132	1	1	0	1	0	N	AT1G54080.2
AT1G54390	0.217939	545.5878	0.352891	0	0	0	0	1	N	AT1G54390.2
AT1G60670	0.243686	345.225	0.291584	0	1	0	0	0	N	AT1G60670.2
AT1G64550	0.23089	1595.881	0.29768	0	1	0	0	0	Y	AT1G64550.1
AT1G64990	0.214651	824.9956	0.538037	0	0	0	0	1	N	AT1G64990.1
AT1G71860	0.243063	1059.69	0.391842	0	1	0	0	0	N	AT1G71860.1
AT1G71900	0.252404	612.8375	0.369024	0	1	0	0	0	N	AT1G71900.2
AT2G16880	0.251567	466.0495	0.346558	0	1	0	0	0	N	AT2G16880.1
AT2G26780	0.226062	1188.741	0.212312	0	0	0	1	0	Y	AT2G26780.1
AT2G29080	0.226666	2415.667	0.388664	0	1	0	0	0	N	AT2G29080.1
AT2G33730	0.246879	2201.633	0.455048	0	0	0	1	0	Y	AT2G33730.1
AT2G43070	0.228445	1009.88	0.203922	0	0	0	0	1	N	AT2G43070.4
AT3G07160	0.252977	4165.341	0.301285	0	1	0	0	0	N	AT3G07160.2
AT3G08530	0.243122	6683.387	0.166151	1	1	0	0	0	Y	AT3G08530.1
AT3G18480	0.235854	925.8499	0.211022	0	0	0	0	1	N	AT3G18480.1
AT3G52760	0.253029	300.7976	0.224787	0	0	0	1	0	N	AT3G52760.1
AT3G54610	0.253067	361.3314	0.214432	0	0	0	0	1	N	AT3G54610.1
AT4G13780	0.254756	2199.598	0.216346	0	1	0	0	0	N	AT4G13780.1
AT4G16845	0.226352	641.6522	0.181492	0	1	0	0	0	N	AT4G16845.1
AT4G34270	0.197447	646.4378	0.2595	0	0	0	0	1	N	AT4G34270.1
AT5G06140	0.131886	1387.551	0.458909	0	0	0	0	1	Y	AT5G06140.1
AT5G08440	0.233144	430.1373	0.174154	0	0	0	1	0	N	AT5G08440.3
AT5G10630	0.249146	637.9463	0.478085	0	0	0	0	1	N	AT5G10630.4
AT5G15270	0.25253	800.824	0.697865	0	0	0	0	1	N	AT5G15270.1
AT5G19680	0.251706	676.8913	0.169188	0	0	0	0	1	N	AT5G19680.1
AT5G20200	0.258611	718.3506	0.239853	1	1	0	0	0	N	AT5G20200.1
AT5G37830	0.256207	1539.036	0.257233	0	1	0	0	0	Y	AT5G37830.1
AT5G42470	0.213156	541.783	0.7797	0	0	0	0	1	N	AT5G42470.1
AT5G48520	0.249213	525.4919	0.433492	0	0	0	0	1	N	AT5G48520.2
AT5G59710	0.255507	793.0905	0.428597	0	1	0	0	0	N	AT5G59710.1
AT5G60160	0.237897	1189.968	0.502109	0	0	0	0	1	N	AT5G60160.1

Supplementary Table S2

AGI	Gene Description
AT1G11060	Encodes one of two redundant proteins (the other is WAPL2) that are involved in prophase removal of cohesion during meiosis. Double mutants with wapl2 exhibit reduced fertility due to defects in meiosis and also some abnormal embryo development in rare cases where embryos are formed.
AT1G18700	DNAJ heat shock N-terminal domain-containing protein;(source:Araport11)
AT1G54080	oligouridylate-binding protein 1A;(source:Araport11)
AT1G54390	ING2 encodes a member of the Inhibitor of Growth family of nuclear-localized PhD domain containing homeodomain proteins. Binds to H3K4 di or trimethylated DNA.
AT1G60670	hypothetical protein (DUF3755);(source:Araport11)
AT1G64550	Encodes a member of GCN subfamily. Predicted to be involved in stress-associated protein translation control. The mutant is affected in MAMP ((microbe-associated molecular patterns)-induced stomatal closure, but not other MAMP-induced responses in the leaves. Arabidopsis has five ABCF proteins, which are all closely related by sequence to yeast GCN20. None of these five are individually required for GCN2 kinase activity.
AT1G64990	Encodes a GPCR-type G protein receptor with nine predicted transmembrane domains. The protein binds abscisic acid (ABA) and is predicted to function as an ABA receptor. It has GTP-binding and GTPase activity and binds to ABA more effectively in the presence of GDP. GTG1 binds to GPA1, the alpha subunit of the heterotrimeric G protein. GPA1 (in its GTP-bound state) affects the GTP binding and GTPase activity of GTG1 and may act to down-regulate GTG1 binding to ABA. GTG1 is widely expressed throughout the plant and appears to be involved in the regulation of several ABA-dependent responses including seed germination, plant development, and promotion of stomatal closure. GTG1 transcript levels do not appear to change in response to ABA or abiotic stresses.
AT1G71860	Encodes a protein with tyrosine phosphatase activity that is downregulated in response to cold and upregulated in response to salt stress.
AT1G71900	magnesium transporter, putative (DUF803);(source:Araport11)
AT2G16880	Pentatricopeptide repeat (PPR) superfamily protein;(source:Araport11)
AT2G26780	ARM repeat superfamily protein;(source:Araport11)
AT2G29080	encodes an FtsH protease that is localized to the mitochondrion
AT2G33730	Homolog of the DEADbox pre-mRNA splicing factor Prp28 which regulates abundance of miRNA. It plays essential roles in miRNA biogenesis and interacts with the DCL1 complex and positively influences pri-miRNA processing. SMA1 binds the promoter region of genes encoding pri-miRNAs (MIRs) and is required for MIR transcription. It enhances the abundance of the DCL1 protein levels through promoting the splicing of the DCL1 pre-mRNAs.
AT2G43070	SIGNAL PEPTIDE PEPTIDASE-LIKE 3;(source:Araport11)
AT3G07160	Encodes GSL10, a member of the Glucan Synthase-Like (GSL) family believed to be involved in the synthesis of the cell wall component callose. GSL10 is required for male gametophyte development and plant growth. Has a role in entry of microspores into mitosis. GSL10 mutation leads to perturbation of microspore division symmetry, irregular callose deposition and failure of generative cell engulfment by the vegetative cell cytoplasm. Also refer to GSL8 (At2g36850).
AT3G08530	CHC2 heavy chain subunit of clathrin. Involved in vesicle mediated trafficking. Mutants show reduced rates of endocytosis and defects clathrin mediated exocytosis Mutants have increased drought tolerance due to defects in stomatal movement.
AT3G18480	This gene is predicted to encode a protein that functions as a Golgi apparatus structural component, known as a golgin in mammals and yeast. A fluorescently-tagged version of CASP co-localizes with Golgi markers, and this localization appears to require the C-terminal (565?689aa) portion of the protein. The protein is inserted into a membrane in a type II orientation.
AT3G52760	Integral membrane Yip1 family protein;(source:Araport11)
AT3G54610	Encodes a histone acetyltransferase that plays a role in the determination of the embryonic root-shoot axis. It is also required to regulate the floral meristem activity by modulating the extent of expression of WUS and AG. In addition, it is involved in stem cuticular wax accumulation by modulating CER3 expression via H3K9/14 acetylation. In other eukaryotes, this protein is recruited to specific promoters by DNA binding transcription factors and is thought to promote transcription by acetylating the N-terminal tail of histone H3. The enzyme has indeed been shown to catalyse primarily the acetylation of H3 histone with only traces of H4 and H2A/B being acetylated. Non-acetylated H3 peptide or an H3 peptide that had been previously acetylated on K9 both serve as excellent substrates for HAG1-catalyzed acetylation. However, prior acetylation of H3 lysine 14 blocks radioactive acetylation of the peptide by HAG1. HAG1 is specific for histone H3 lysine 14.
AT4G13780	methionine-tRNA ligase, putative / methionyl-tRNA synthetase, putative / MetRS;(source:Araport11)

AT4G16845	The VERNALIZATION2 (VRN2) gene mediates vernalization and encodes a nuclear-localized zinc finger protein with similarity to Polycomb group (PcG) proteins of plants and animals. In wild-type Arabidopsis, vernalization results in the stable reduction of the levels of the floral repressor FLC. In vrn2 mutants, FLC expression is downregulated normally in response to vernalization, but instead of remaining low, FLC mRNA levels increase when plants are returned to normal temperatures. VRN2 maintains FLC repression after a cold treatment, serving as a mechanism for the cellular memory of vernalization. Required for complete repression of FLC. Required for the methylation of histone H3
AT4G34270	TOR signaling pathway protein; response to high temperature treatment.
AT5G06140	Homolog of yeast retromer subunit VPS5. Part of a retromer-like protein complex involved in endosome to lysosome protein transport. In roots it co-localizes with the PIN2 auxin efflux carrier. Involved in endocytic sorting of membrane proteins including PIN2, BOR1 and BRI1.
AT5G08440	transmembrane protein;(source:Araport11)
AT5G10630	Transcripts of this gene are alternatively spliced to encode either HBS1, a decoding factor translational GTPase, or SKI7, a component of the cytosolic RNA exosome.
AT5G15270	RNA-binding KH domain-containing protein;(source:Araport11)
AT5G19680	PP1 Regulatory Subunit3. Interacts with members of the Type One Protein Phosphatases (TOPP) family.Facilitates the nuclear localization of TOPP4 which is required for its activity in mediating ABA responses.
AT5G20200	Atypical nuceloporin-like protein.
AT5G37830	Encodes a 5-oxoprolinase that acts in the glutathione degradation pathway and in 5-oxoproline metabolism.
AT5G42470	BRCA1-A complex subunit BRE-like protein;(source:Araport11)
AT5G48520	Encodes AUGMIN subunit3 (AUG3), a homolog of animal dim gamma-tubulin 3/human augmin-like complex, subunit 3. Plays a critical role in microtubule organization during plant cell division.
AT5G59710	Encodes a nuclear-localized NOT (negative on TATA-less) domain-containing protein that interacts with the Agrobacterium VirE2 protein and is required for Agrobacterium-mediated plant transformation. It likely facilitates T-DNA integration into plant chromosomes and may play a role as a transcriptional regulator. The mRNA is cell-to-cell mobile.
AT5G60160	Vacuolar aspartyl aminopeptidase which also functions as a molecular chaperone.

Supplementary Table S2

AGI	Gene Model Type	Primary Gene Symbol	All Gene Symbols
AT1G11060	protein_coding	WINGS APART-LIKE PROTEIN 1 (WAPL1)	ATWAPL1, WAPL1
AT1G18700	protein_coding		
AT1G54080	protein_coding	OLIGOURIDYLATE-BINDING PROTEIN 1A (UBP1A)	OLIGOURIDYLATE-BINDING PROTEIN 1A (UBP1A)
AT1G54390	protein_coding	INHIBITOR OF GROWTH 2 (ING2)	INHIBITOR OF GROWTH 2 (ING2)
AT1G60670	protein_coding		
AT1G64550	protein_coding	ATP-BINDING CASSETTE F3 (ABCF3)	ATGCN3, GCN3, ABCF3, ATGCN20, SCORD5, GCN20, ATABCF3
AT1G64990	protein_coding	GPCR-TYPE G PROTEIN 1 (GTG1)	GPCR-TYPE G PROTEIN 1 (GTG1)
AT1G71860	protein_coding	PROTEIN TYROSINE PHOSPHATASE 1 (PTP1)	ATPTP1, PTP1
AT1G71900	protein_coding	(ENOR3L4)	(ENOR3L4)
AT2G16880	protein_coding		
AT2G26780	protein_coding		
AT2G29080	protein_coding	FTSH PROTEASE 3 (ftsh3)	ATFTSH3, ftsh3
AT2G33730	protein_coding	SMALL1 (SMA1)	SMALL1 (SMA1)
AT2G43070	protein_coding		
AT3G07160	protein_coding		
AT3G08530	protein_coding	CLATHRIN HEAVY CHAIN 2 (CHC2)	ATCHC2, CHC2
AT3G18480	protein_coding	CCAAT-DISPLACEMENT PROTEIN ALTERNATIVELY SPLICED PRODUCT (CASP)	AtCASP, CASP
AT3G52760	protein_coding		
AT3G54610	protein_coding	HISTONE ACETYLTRANSFERASE OF THE GNAT FAMILY 1 (HAG1)	HAT1, GCN5, BGT, HAG1, HAC3, HAG01
AT4G13780	protein_coding		
AT4G16845	protein_coding	REDUCED VERNALIZATION RESPONSE 2 (VRN2)	REDUCED VERNALIZATION RESPONSE 2 (VRN2)
AT4G34270	protein_coding	TAP42 INTERACTING PROTEIN OF 41 KDA (TIP41)	TAP42 INTERACTING PROTEIN OF 41 KDA (TIP41)
AT5G06140	protein_coding	SORTING NEXIN 1 (SNX1)	ATSNX1, SNX1
AT5G08440	protein_coding		
AT5G10630	protein_coding	SUPER KILLER 7 (SKI7)	SKI7, HBS1
AT5G15270	protein_coding	(ATKH25)	(ATKH25)
AT5G19680	protein_coding	PROTEIN PHOSPATASE 1 REGULATORY SUBUNIT 3 (PP1R3)	PROTEIN PHOSPATASE 1 REGULATORY SUBUNIT 3 (PP1R3)
AT5G20200	protein_coding	NUCLEOPORIN 82 (NUP82)	NUCLEOPORIN 82 (NUP82)
AT5G37830	protein_coding	OXOPROLINASE 1 (OXP1)	OXOPROLINASE 1 (OXP1)
AT5G42470	protein_coding		
AT5G48520	protein_coding		
AT5G59710	protein_coding	VIRE2 INTERACTING PROTEIN 2 (VIP2)	VIP2, NOT2B, AtVIP2
AT5G60160	protein_coding	M18 ASPARTYL AMINOPEPTIDASE2 (DAP1)	M18 ASPARTYL AMINOPEPTIDASE2 (DAP1)

Supplementary Table S2

AGI	UTR_promoter_seq
AT1G11060	TTTTGTCAACAAGAAATTACATAGTTTGATACATAATGATTTGTCAAAGCATTTTCAACAGAAAAAAATCAGATACATATTTTACTATT AATTTTATATCTGGTAAATTCTATATTTTACAGTATACAGTACATTACACAATTTCCTGAATCAGTTGAATATAAAATGCTCAAACTAGC ATAAATTTAAAACTTTCCAGCATTGGTAATACTCAAAAAAAAAAATCAATAACAAATAAATGTTTAGTTGTAATTCAGACACAACTTG AAACCCCAAAAATCAGACTGTGAAAATTTGTGTGCTCTTTTTAACTTTTAATTGCTAAAGTTAGTTAGTAATTTAAAACATTCTGGTT ACTCACTTAGTATTAGTTATTACTGTGGTAACAACAAAATTTA
AT1G18700	ATCATTTGTATTTTATTTAAAAATTATAAAGGTTTTGGCAATTTTATCTGTAAATTCGTAATACTATCTAATATTGTTTGGTAGTCGTCAC ATATTAATCTTTTTTTTTGTTTTTGTTAAAAATAACATATTAATCTTAAACCATAATTATATTTAATTTTAGAAAAAGAGAGATTTAGTAT GCTGTGATCTGATAAGATACGATTTTAACCATAATAAAAGAGTATAATTACTTACTTAATGCTAAGAGTCGATCCCTTTTTTTATTTATT AAGAAAAACGCGGAAAAATAACCGAAGGAGATTTACACTTATCACATCGAGACTCTTTGATCGGAGTTTCCCAAAATCCCCG
AT1G54080	CCGACGGCGTGAAGTTACCGGAAAACATAGATTACCTTAAAAGTTCTATTACTACTTTCACCTTGGAGGTAGATTTTTCGTTTTCAGT CACATGAATTTGGCAAGAAAAAAAGAAGAAGAAAAACTAACTACTTTAACCATAAACCTATAATATATTTTTTCACACTAAACTATAA ATATATAATACCTATTTTACCTTTTATACATAAGTAACAACATCATCACAACTACCACTTTCACTATCATCACCACCACCGTTGTGACCAT CATGTTATTACCATCCCTCCATCATCACTACCACTGTCATCAACGCCACACAAAAAAAAAAAACTCATGAAAATTATTGATAATAATCC ATAAAAAGAATTTATACTTTTAATATCTAAATTTTACTCAAATCAAATTACCAATTTTCTTCAATGCGGATATACTAACATCATGTTATAG CCTCCATGCAGAGTTTAAGATGAACATTTTTGGACTGCAAGCCCATATATTTTGTTTAGCCGACTTTTTAAAGAATAAATTGAAAAAC ACGAAAGATTGACGTTGAGATTCGAATCGAGATGCATGATTACTTAAAACCAGCTAAGTTTCATAAAAACTAGACCGATTAATATCC TCTGAAAACCGCAACCAAATACTTGTTCGCTTGTTCTCATGGACGACACTACCTCCATGGTAAAACCTTAATAGAGTCTTGATGTAAT AATTATTAGCAACAACCATCATAACTAAGACTATCTACATTGGTTACATGATGAATCAAATCTATAAGAATGACATTTGTGAGTATTTC GTTTCATTTTTTAAAGTGTAATTATCACTATCACTAATTTTGGTTTGGATAAATGATAGATCATAAATCTACTATTCTAAGTAAAATCTAA AATGGATTATTGGTATATGTAGATTTTAAATATTTTATTGGTTTATTGATTCTTAAATTCCATCCAAACTTCATTATTGTTGGTTCAAGCT TTTTTATATATTTTTATTCATATATTTTGTTATTATTGATATTCTAATTTTTTTTAAACCTATTGTTGTTACAATTCTCTGTATTTTGGCTCTC TAGTTATAATTGATATAAAATAACTTTAAATACTTAAAAATACAAATCTAGAATAAGAATGGAGAAATAAATGTTATTTATATCAATGAG AGAAGAGAAGAATGCACAAATAAAGAACATTGATATGTTCTTTTATCCACGTGAGAAAAAAAGAATTCATTTTGGAGTTAAAACAT TTCCATCAATACATGATTTACTTTGAAAATAAAATGTTAGGAAATGGTTGATAAAATTCCATATAAACTCACTAGAAATCATCGGTAAT AATATATAAGAAAAAATTATAAATTGTACACAATCAACTTACCAATAAGTACCATACTGTGAATTTGTGAAGATATTTGATCCTTTACAA GAATCTTGCTGACGTACAAAACGAACGATTAAATATAAGAGATAAACCAAACAATAAATTATATAAGAGGATAATTTTTATAATATTTA AATAATCTCTCCTATGCAGTTTCAGCTTAATCGCGTTGTAATCTCAGCTTAACTAGTATCTTACACTTGAAGATCAACAAAGACAAACC TGCAAGATGGAAATAATCCCTTGTTAAAGTGATGAACATATCAATATCCGCAACAACATCACCTTGAAGCACATTGCATATCGAGCTA CAAACAAGAGTTTGGTCTAGGTGAGGATTTTCATCATCCAGATGAATCTTTAAAAAGAGGTTCTTTATACTCGACTTGATTTTCCGAT TAAACAAACACTTCACAAAATCAATGAATTCCTTGATACTATCATGTTATTGTATAGCATAATCAAGGTTAAAACGCCGAAAAAGAAC AAACACATTCTTTCACTTTCTGGAGAATACTGTCGTTGAAACAGTTTTTTTGTGGAAAGAAAAGATAAAATGTATGCAAAAACATCG TCTTGTAAAAAAAATCTTTTAAAATTTAAAATTTTCTAAATCGTACATAATCAACTTACAAACTTGTGAAAAAATTTGGTAATATATAAC TTGTTTGTGAATATATTTGATCAAATACTCTAAAAAGTACTAATAAGTTTAAAAAAATCTTGCCGACAAACAAAACGAACAAACATAA GGGATAAAACAAACAATGAATTATGTAAGAAGATAATGTTTTTTAGTGCAGTTTCAGTCAATCAAAGAAAATCAAAATTTCTAAGTA TATATTAGTTTGGGCTGGGCCGTCAACAATTGGTCAGTCACTGAAACGAATTTGTTCGGGCCTTGGGCTTGATCCGTCAACAACATC AAGGTTCAGAACAAATGATTGGTCAGTCACTAAAACGAATATCACAGCTGCCGTGAGATTAGTATACAGTCTATCTCTCTCTCGAAG AAATCGTCACACTCCTATAAATGGTCACTTGTTTCGTACGGCCTTTCTTTACTCCTTTAGAGAGAGCGTGCCATTTTTATTTTTCTCTT TCTTCTCTCATTTTTATTTTCTTTTCTTTTTTTTCACCTTTTTTTTTTTCCTTTTTTTTTCTTTCTTCTTCTTCTTACTTGATTTTGAACCCTA GCTTAAGGGGAATTTTCTCGGGAAACAAAAGAGATATTTTTATCGCAGTGAGAAAGAAACACAAAAA
AT1G54390	GGTGATTGTGATGATGACCCATCTAATTGCTTCTATTTATATAATTTGTTCTGTGCATGCAGTTTCTTCTTGGATCTCCTATGTACTGAAT GATTCGTTTATTACCCTTAAAGATTTGCTTTAAAAATTAATTTATTTCACTCTTTCGAGTTTATATAGTTTGTTGTATCTTGTGAAATTGG TGTTTCATCACATGCACATATCTTCGTCTATGTGGTTCCAAATATTCTTTTGTTCTTTTTATTTTACCATGCATTGCTTTATTACTTACACA TAATGACAATTTTGCTGGTCATTGATTAGAACATCGTTTTAATATCGAATGGGATGTCGAGTTGAATATTTTGACAACGTTCGTTGGC TGGGTCATGAATGAAGTCAGAATACCATTTCTTTTTCATTAAACAATGCTATTTTTTTCCCGCAAAACAATGCTATTGGGCTCATTTTG GGCTTAAAATGTTTTGCGTAATCTCAATGTTTGTGGGGCTTTTAGTTTTTAGCCCATGTATGGTTTGATAACTTTAGTAACTCGTGAG AAAAAAAAAAGCTTCGTTCTCAATCCCAATTCGAAATCCTAGCTGAAGTCTATGCTCCATAATCAAATCCAGTTCATCAGAGATCAAA G
AT1G60670	TGATGTCTCATTTTTGTCTATGCACAAATTCAACTTGATTATTTAGGTATATAAGTAAATAAAAGATTTATCATAAAATTACACATGTAA AAGTTTGAAACAAATTTAATTATTCATAACTCACATAAGACATGTAAAAAAAAACAAAAAAAACAAATGTCACTTAGTGAAATGGAT ACCAAACTTTATGATTTATAATACCTTGTTGTAATTTAAATATCTGAATTTCCAAAGTGTAGGTGACTGTGTAATTCACAAAAATTGAG TAAAAGTTTAAAAAAATAAGGAGGAACAAACCATCTTTCTTCAAAGAGTGGACTAGACAAAAACAAAAAAGCAAAACCACCACA ATGGAATAAAAAGTGGTTCTGTTTCTTAATAGAATAAAAACCATTACAAAACAAAACCACCATTTTCTCTCTCTCTCCTCGCCAGC TTCTGCAATTCGAAAGTTCTTCGCATCATCTTGCTGTCACCACTCTCTCGCGAATCGAATCACGATTCTTCAATCTTCTTCTTCTCTATT GCGATCTCTCTGAATTTTGTAGCTTCACGATTTCAGTATCTCTACTTTCATCACCAATCGTTGGGGGTTTAGGGTTTTTGCATCGACG

	CCATCTATCGAGCTACGGAGAAGCATACTCATCTTCTATTTTGGCTAAATCCGAGGTGGGTTTTGCTCAGGTTTTGCTCTTTTAGGTT TCTGTTGGGTTTTAATCGGATTTTTTTCCGTATGATTCTGAAATTGCTAAAACCCTAAATTGAACATGTTTTTGGGCGTGCTTCGTTGT AGGTCAACTTAATCTCTACGAAATAGCTTTTGAGTG
AT1G64550	TTGGCGAAACAATTAAGAGGTCTAGTTTTGGTAAAATCCAAATTGCTAATCTATCATCTCTCAAACATGTGTTAACAAAAACAAAATT GAGGACTTTTGATTTCACAATGGCCTTGAGCCATGTTAAGAGGGAACAACGTTAACCGGATTAAATCATAATATTTGGTTTGGTTTA GTAAATTAGAAAACGTACTGTATACTCTCTCTGCCTGAGGGAGAAGGTTCCAAATAGCCCTAAAAAGCTTCGCCTTCTCTCCTTGCC TTCATTTCGGGTGTGGAGGAAAGTTATCAAATCAGTTTTTGTGGAATTTTGAATCTCGTCGCTTCTGTGTTGGCT
AT1G64990	AAATGTAGGATTGAAAAGCATAAAAACATTATAAACAAAATAGCTCAAAACTTTTTTTTGTTAAATAATTGTGATATAATAAGAAAAA CGTTTTGTAGTTTTTTTAATTTTGAATTATTTTTTCCACATGCATCAAATATAACATACACATGAAAATCAAATATAAATGTATGATAAAT ATTGTGAATGGTAAAAATAATAGAATCACGTTATGAAAAATGGTATATTATACATCATCGATTGGTGTCTGATAAGTTTGGAGGAAAC TTTAGCAATAATACAGGTTTTCTCTATTTATTCATTAGCTAATAGGATTATTTATACAAAAATCATAACTTGATTAGCAAGTTCTTTACAT TTATGATAATATTAAATTTTCTATCTAACAATAAGAAAATGTACACAAATATTAATTTTTTATTAATGATAAAAAAAATATTTATCTTGATC CCGTAATTATAAGATTTATCATGAACTATCGATATCCCGATCTTTCCCTTCCCACTTCCCGCAAATTTATATATGGTAACACGTGTAAGAT TGGACAAGCCTTCTAAACAACAATTTTAATGAATCGATCATTTTTCGTGTATATATGTAAAAGTGAAAGAAAAACAAAAAAATCAATC GACCAGATAAGAAAACATGTCGGAAAATGGAGTTAAAATTATCATCAACACTAGACTAGACGCCAAAACCTATAAACAAAAAAAAA GGATAGAAACACTATTGTTGTGCAGTTTCTTTCACATTACTCCGTAGCTCAATGGTTGGTTAGCGTTTAATAACGCCTTCCTCTCTGC GAGTTCGAGTTGAGTCGCATGGTGGTGGTGGTGATGATGAAGTTGATGAGGTTGTTGTTGATTAGAGTTTAAACCAAATAAACCCT AACCAAAAATAAATAAGGTATTCTTAAATACTTTTTTTTTAATTGTTTCATAAATGATCAAATTCGAATCGAATCGTACAACAGAAGCT CTGGTGTTGAATCGCCATGAGGTTTTTGCAGATCTGAGGTTTTTTCTCGAGAAGTAAGAGAGATCTGGAATTTGGCGACG
AT1G71860	GTGTGTCACTCGAGACTCGGAAGTGAAGAGTTTTTATTATTGGTTATGGGCTTTTAAGTTGGGCGTTTTAGGCCCACTGTAGTTTTT TCTGTGACGTTGCCCATCTTTTTTTTTATAATGATGAGCCCATCTAAAATGTTATTCTACAAAACAATAAAAGAAAGACCAGGATTCCT TTGCTTAAAAGACCAGGGTCCTTTGCTAAAACTATATTATAATGACCAATGAAATCATAATTAATAATGACCAATTAAAATGATGACAA AAAAAAAAAAAACCCAATTAAAATAGTAAATTTATATTATAGTCATCAACTTTAATACTTTGTTGCAAAATAAAAGGAAACAGTCGCA TAATAATCCACTAATCACAAACCTTTACCCTTTCTTCTCGGGGCCAATCAAAAACTTCAATTAACCAGAAAAGGTCAGATTCTCCGCC GCCTCACACGACGGGCAACTTCAGTTCCTGCCGTCGCGTTAGATCCCGCCGTCAATCGGTC
AT1G71900	CAATTTTTTTACTGTTTTTTTTGCTTCGGTGAAAGAAAGATCCAACTAGAGAAAATAATCCCAAAACAAAAAATCAACATTCAAAAT CTGAGCTCTTTTCTAAAATTAAATTCAAAACATGAACTTATTCTATGTAATTTTTTTTTTTTTGTTTTAAAGACATGGCCAATGGGCCTT TCACTCTTGTCAAACCCATTTGAGTCGGGCTTTAGTCATAACTCATTTTAAAGATTAAGGAATATGAGGCCAGTTCTTTTGTCACAAT CCTAGTTTAAGCTTGTCTCACTTACCAATTACGTCTCTCAAAAGACAAATTTTAAAAAATCTATAATCTCTACTCTCTAACATGTATATAT ATATGATAAACTCGTTGGCTATTACAACTTTTAAATCACAAATGGCCTGATTGGAAAAAAATGTATACATAAATGATAAGTTTGTCACT GATTAATGCGAGAAAATTCTAATTATTTGCATTAATAAAATTAGCGTAGTTAGGTTTTAAAAAGGTTTATACAATACATTATACATATTT ATAAATTTAAATGTTTTTAAATTTATATTAAATTTATTTTATTATTTAGATAAATCTTAAATAAAAAGAAACCAAATATTATTTGTTATTTT TTCGTGAAATTTATCTTTTTAATATATTTTTTATTTTAGTAAAAAGAAGGAATGGTGGCGCCGGAAACTGACCGTAGGGTCAGTCTAA TTACTCTCTCAGCGTTTCTCTTTTTGGTCCTCAAATCTTCACCTTCAAGATCTTCTTCGTCCCTCGATCTCCGATTTCAAATCTCGTTTC TCCGTTCAAGCAACCTCTCTGAATTCCTAGTCTCACAGCAGGAATTAGTACTCGGCGAAGCGATTTCTGCTTCGATCTCTACATCTGA AGTTCCGATCTGACGAAATTTAGGGTTCGTCTGGGTCGAAGTTAAATAGGAAAAAGATTTCGAATTGGCTTGAGA
AT2G16880	TTTAAATAAATATTTTAATACATACGTGTACAGTTTATTTTCATCGACACAAGTTAAGCTAAAACTCGAGTTTGAAAACCAAACTGAG GGAACCATTATTTCTAATTATACAAGTATGTGTCAAACATTAATCAGCTACAAGACAGGATAAATTCGGAAACGAATAGTGTCATCTT TAATGTTAAACAGAAAAAAAAAAATATACATAAAATTCATATTAAGTTACAAGTATATTGGTTGAGTGGAGTTGTCATAACATTAATAT TAAGAAAGAACAGTGTCATCTGATAATACTTTATCTGTATATGTAAGTTCAACTGATTCTAGTTTACGTACGAACTTAGTTATCGAGTT GAATTTAGGTAGAGAGTGGAGACACATCGATGATGATAGCAAAAACCATTACACTAGGGTAAATAAAAGTTTAATATTACGTATACA TATAGGGTAATAATAGACATCATAAGGTATGTGGGATATTTTGTGACTTAAATTAAATTATATTTGTGACAAAAAAAATTTTTATCTAAA GATTAACATTTTAAAATACTCATTTGTTACATATCAGTGAAATCACCTTCAAATCAAATTTCAAATCAAAAAGAAGAAGAAGAAGAA AGTGATAATCCAAAGCTGACACTTCGAGTAATTTCCCATATATTAATAATGATTAATGCTTATAGCCTTATGTATCAATATCACTCATGCA TGCGTAAGTTTAAACGTAAAACCTAGATGAAAAGAAGAAAAAGAATCACAAAGATTTGGGCAAGACAGGTGTTACGTATAACTTT CTCACGGAAGCAAGCAATGTCCCAGAAGTTTCTTCTAGAAACCAATACTCTAGAAAGACTATATATCAAATAAACAAATACAAAAAT TGACGAAGAGGAAAAATATATTAAAAGATGATAATTTGGATACACTATATTCATATTTTCACTGAATAAATATACGTATATATACGCTATT TGTTCACTTAAAGTATAGTTGTAAAGAACTATGATCTTCAAACTCTAAGTAGTATCATAAATCTTTTTTTTTTTTTTTCTTTTATTATTATG ACTTTATAGAAAGGACACTATGATGATGCAATTGACATGGACAAAAGTCACTACGACGGTCTATCGTTTCGCCCCACTTTTGCATTA GAGCCAAGTACTAGCCACAGCCTCTTGCAGTGACAATGTCGCGTTCTAGGCAATGTCTCGCCGTTCGTTTTTCTTTCTTAGTTTTATA TTTTGAAAATTTTGTTTCTTGTCCTTTTCACTTTTAGTCTTTAGACTTACTTTTTTGCACTACATCAACACCTATCTACAAGATATTTCT GTAGATACGTATATACACTTACTTCTTTTTCTAGAAGCTCATTTTATAACTTATATAACATATAAAGTATCCTAGCTTCATGTCATGTTGG CTCGTTTCGTTGTCGTCGACATAGAATATATAATGCTCCTTCAATACTTAGATTTAGATGTTGCTGAATAATTGAATAGCAAATGCTTAT ATCTAATGTAAATATTAGTTTATCACTGACCAAACTTATAATCTAGTGGTTAATACAACTTTGAATTAAGTGCTTTTGTATTCGTGCTGC TATCGTATCATCATCGAATTTAGTGCTATTGTACTAGAGTTCGTGTTTGTGAATTAGTGATAAAATATAATTTTACTAACTTATAGCTTAA CATTTCTTAAACGTTATATTGTAATTAGGGATGAAAACACAAATGCAACAGTTAGTTTTTTTATTGGAATATAATCACTACATTTATCAA CATTTCAATTTCGCAAAAGATATTTTCTAACTTGTAAATAAGTAAATTATTTTTTTCTATTTGTCTGTAGCTACGCAAAATAAGGTTTTG

	GAGAAATATGAATGTCATACAGTAAACCTAAGAATAAACAATAATAATTAAAACAAATAAATTTATTTTAATTTAAAAAGAAATCTGTT TTACTTTACTGATCCAATTTTTTTAAAATATAAATTATAAGAAAAATAATCATGCGAAAATTCGTATGTACTATGTAGCATGACCAAAAC ATTTGAGAATGAAATGATTTTTACCAATGAAAAAAAAAAAAATCATGAAACGACAGGGTCCTATTGCGTTTGGTTTAAGCCGAAAA CCCGAGTTCTCCAATTTCCGGTTCGGTGTAATCCCGGTTAAATATACGCTTCACAGTTTTAAACCATCACAGTGAAAAAGCTTCATAG CTTTCTTCTTCTTCTTCTTGACGATACAACGAGGATGATCATCAGATAAATTTCCATTGATTGCCAACAAAA
AT2G26780	AATTTAGTTTTGTTTAAAAGTTTTTCTGGTTTAGGTAATAAATTGTTGCATAATTTTCAACTCCGGTTGATGTGTGTCGCTGTCGGACT GTTGTTGTCTGAAACTTAAAAATCACAGAGCGGTCACGGGTTCTGTTTCGTGTTTAACTGCCACGATATTCTTTTTAATTGTTTTTTA AAAGGGAAAGTAATCAAAATGATGACACTAATAAGAGTGGTTTGGTGTGAAACTTAGACCTTTCTTACACAAATTTGGTAAAGCCA AAACTTTTTTTGGTCATGGGGATTTATTTTTTTTTACTTTTTATTTTTTTTTTTGTCATAAAGAAATTATAGAATCTTGTTGTGAAAGAC GTAAAATATAATGCCCAAACAAAAAAAACGTAAATTGGCGGATAAACTATAAGTTCCCAAAAAAACATAAGCTGGCGGAAAAAGA GAATTCTTCGGAATCTAGATCCTAAATAAAGAAGATGAAGTAAACTCAAATTATAATGCAGAAAATATCAAACCAATCTCGAAAAAG GACGATAGTACAAACTAAGATCCATATTAACAAGACAAAATCTTAAAAAAATAAGAGATCCAACAAAGCAGAAACTCCACTTCTTTT AACACACCTGCAAGTACGATTGATCAAATCCAAAATAACATTTACCACCTCCTATCACGTTGAGAGCCAATCTCTCGAGAACCAAGG ATATCATAACCCGACATGCAAGACCACTAAAAGAAACACAAACACCATTCCAGATACGACGATGACCACCAATTCGTCACTTAACGT CTAGATAGAATCCTTCAAAAAACAATCAGATTACGACAAAATTGTAAAAGAAGGACAAGAAGCGAGGCAAAACTTCTCAAAACCC ACCAATCACAAGATCACCAATCACGAGTAAAGCCACAACTATTTTAATTAAATCATACAATATCCTAAATTAGGCTTTTCTTTTTGATA AATTAAAAAAAAAAAAGATAGGATCATCCAACAAAAATAGTACAAACAAAAAATGTCAAAATCTTGCATCACCTCTAAAAACCTTTCA TGTTAACAAAATCTTGCATATTTTTGCCCTATACAAAGTCTTGCATATATACAAAGTCCTTTGGTATATTGTTGTTAGCCACCTTATCTA ACTCTTTTACGGGCTCTGTCCTCATGAAGAAGAACAATTTTTTTAATAAATTGTTACCTTACCACATATAAGCTTAATCTGGAATGGTT GTGATTTAATCAATGAAGATATTGCCTCAAACCTGAATAGAGGAGTGAAAGTGGATTTCAATTGGTTGGTTATCAAACCCATCAGCA TGATACCTGTTAGTTTCTCGGCCATACCACTCAATCATTGCCATAAATATCATTGAAGTCTACTAGTCTGTTTGAGACATTCAACTCAC ACAAGTGGTGGGTTATTTTCTCTGACAAACACGGCCAAACAATAAAATTGCTGAAAACAAGAAGTTGCCTAAAAAACCAAAGGAC TTGTGAGAAAAATGGCAAATGCGGGCTAGGATGGGCTGTGGATTAGAGGCCCAATATGCCTAATGGGATACAACACGACGTCGTT TAGACCAGCCATAGTTCTCTTGTCCGCCAAGACGGGAGAGACTAAGGAAAAGCATAAGAAGAGAAGAACCAAAGAAGAAGCAA AGAAGCACTGATAAATCTCTGCAACTCAACAAATTTCTGAGATTAGCT
AT2G29080	TTCCAACTAGCTGTTATTGCTATCATGTTGTGACAGTTGCGTTGAAATATTTGCGTTTGTGTTTGGATCCGAAGCAAACGCCAAATTT GGAATTTGATATTCTCTTTTAACCTCTTGGATCCTAAATGTTCGATTTATCCTTTTACATTTTCTCCGGCATCATTTTCTACAATGCAATA ATTCCGGATATTTGTTTGAGATTCCGTTCAGACATATAATGAGCTGAGGTAGACGTAATTGTTTGTACGGAAGCAAAAACGTATTATA ATTATTTACTTGTAAGTTTTTAGTTGGTAGTATTAGTAAGTTATACAGAGTTGTAAATAAACTAATAATTCGATTAGACCACCATTATCA ATATTGTTGTATGTAACAAACACGGAAGTACATAAATCGTTGAGCCGTGGAAAGGAGACAAAACATATTTTTAGTCTCGTGATAAAA CGAGTCGCCACGTGTCCACAGTCGGGTACGTGCATGAGATCTAACACACTTTATAACTAGCTTACGTGGGACTCTCGACTTCCATAA AAACGAAGAAGATTGAAATGAATCTAATTCATTGTCCCCGATTTGTCTCTGTCTAAAGTCAGAGAATCAGAGATTCGTTTCAGAGTA AAATGATTTTGCATTGGTCCACTTAAGAGTCTCTTAACCATTCCAAAGGTTTGGTGGACAAAACATGCTCCAACCCTTTCTTATCCAA TTTGTCCCCTCTTGTCTCTCTTTAATTAAGCCCTACCCCAAAGTTTTTCACATGATTTGTTCTTGAAAGTGTCTCTCTCAATCTTTGTTG TTTATTATTATTATTTGATGTGAGGCGACTAGAACGTGTACAAAGAAAAGAGGGGAGATTGGGGATTAACACGATCGATATTCCTTT TGGATTAGATACGTGAAGTACCCGAAATGTATAAGTTTTAACATACACTAAATTAGTGTCAGTAATAATCGGCGTAAATGAGAGGAG TCTTTCAAAAACATACACTACATCTTATATGTCGATCAATAGACATAGTTCATACCCAAAAAAATATACAATACGATGGTATAAGGTGCC ATGATTTCCAAATTGAGTTTTTGCTTGATATCAGTTCCTATACATTGTTTATGAGGAACAAGTTATGTATATATGTCTATAATTTTATCAT ATATATATATATATATATGATAAAATTAGTGATTATTTATTTGTTTCTCAATTATGATTTAAAGAGCCATAATTCTTCTTTAGCCAAAAAAA AAAAAAAAAAGATAGAACCATAATTCTTTCATTAACAACAATGCAGATAATTTCGTTGATGATGGTTAGTGTCCTTTTTAGTTAATATT GGTTCATTACTGAAAATAAAATAAACTGATATGCTACGGCTACCTCTTCACTTTTATTTTCAATTTATGACAAAAGCTATGCCCATATGT TCTCCAAGGCCCAATATATGTAAAATTAGGCCTATGAACCAAACACTGGTTTAATAAGACCGGTAAGCACAAGTTGGATTCGGACAT TATTTAGAACAAAACCGGTTCAAGAGAATCTACAGAACACAACAAGTACAGAGTCTTTTTGTCTTTGACATTTCATAAACCTTCTCTT CTCTCGATAAACTTCAGATATTAAAAAACCCAAACTTTTATTTCCCCCAATCAAAAACCCTAATTTCTTCTTATCTTGTTTGGATTTTTC ATTTCTTCTGTTTGATTTTATTCCACAAACACAATAAAAAAGTTTGCATTTTTTCCTCTTCTAAGCTCGTTTCAGATCTTCGATTTCACC ATTTTTATAAACTATTATGCGAACTTAGTTTTGTTCTTTGTGATTTTGT
AT2G33730	AAAGACAGAATTAGTGTAGTTGGTTTAATCCCTAATGATGAATGAACTGTGTCACACCACACTACTACAAGCGAGTTCTCTAACCCT TGAAATGGGTTTGTAATCTTCACCCAAGAGAAAATCTAACTACTCTACATAATCTCCGCTGAAAAAACACTAATGGGCCTAAAAACA AAATTAGATTTGAATTCGGAATGCAAAAGAAAAACTAATGGGCCTAAAACCCTTGAGAGTTTGCTAGAATTGGGGGGTTTTGGAA AAATGTAGAAAAAGAGAATCCAAGTGAAAAAGAAATCAGAGCGAGAGAAGAAGAGAAGAAAGGTTCTAGGTCTTGTTTGCTTTC CCCTCCTGCTAATCGACCAAGCGAATAATCTCCAGCTAACCACCG
AT2G43070	TAAATTAAGGTACTAATTACAAAGGGAACCAAACAGAAAAAAAAAAACAAATTGTTTTGTAATTAGCTCAAAACGAAATTCAAAGG GAAAAGATTTTAAGTACAAACTGTAAATTTCTTGTAATTTTATATAACCAAGAAGGTGTTTCTCTCCTGAATTATTCGTTTTTCTTGAT ATTTCTATTGTTAAATAAACAGATAAATCGAGTTCATATAAATTAGTTATCTAAATAATTTAGTTTTTCTTGATAGTTATCAAATGAAGA ACATTTAAACCGGCCGGTACGATGACATTTAGAATTTACAATCAAACTCCTCTACTTAACTGTTCTATTTGTTTGTTTGACATACTAGA CAAGTCAAAATACATCGTGGGGATCTGAATTCTGCTCAAAATACTTTTGAAACCATTATATTTTACATAATGCGACATATATTAATTTAT GATTTTTGATTAGTTGAGCAAAAAGAGCTGTCGTATATCATATAGTTTTAATGAATCACTCAAAAATTGGCACAATAGAAACAGTAAA

	AGAATTATTGAATTTAGAGGATAAAAAAGACTATAATATCGACCGTCCGATATATCCCAGAAAGATTTTAACAGCAACTCTTTGTCTC TTCTTCTTCTTCTTCTTCTTCTCTCCGATTTGCTCAGTCGATTGCGTC
AT3G07160	ACGTTGCGAGAGAGTATTCGGGAAATTAACATCCAATCAAAGTAACGGGTCGGGTCGGGTCGAGAATCAGGATTATGCGTTCAGA ACAAAAAAAACGATAACCGGTTTGACTGTTAACCAGAGTTAATCCGGCGATTCAAAAGACGTTTCGGTTCGATTTTTCTTTGACAA CGCGGCTAAGCTTTTCATTTCTTTTTTTTTGTTTTGTCAACACGCGGTTAAATTTTTGTTTGTCAGATTCATTTTTTTTCTTGATAAAA AACATCCAGATTCCTTTCTTTGTAATGTTTTAGATTAAAATAAACAAAAAATCCCTAAAAAACGAAAAAGTTTTGAAACTTAGAAAG ACGGACATAAGCGAGCTAGCTTCTAAGCTTTCTTCGCTAAATTCTTGGCTAATTCTCTTCTATGACTGATCCAATTCTGCTTCGACTCC ATATCTGAGGTTTCTGACGTCTGATTAGGTAACTCCCGCTTTTGTTTTTTTTGGGGGGTTTTATCCATTTTTAATGTCTTTGTTGTCTA CTTGTTTCAATCGTATGAACTTGAGACGCGAGAAGCTCAGATTCAGGATCCCTAAATTTTTCTCTATTTTTTTTTTTTAGTTTTGTTTCG TTGATTACAACTAGTCCTAGGTTAAAGAACAACGTTTACAACTTATTTTGCTTTTTTCAGTTGTGCCTAGGTCACTTGTCCGAATCTG TATAATTCACGTGTATCTTCCTGGATTTGATTACATTTCATTTACTTAGTGTTGGGTTTTACGTTAGTTTAAGCTGCTGAGTACTTTGAT GTTTTTTGTTTGTAGAAA
AT3G08530	AGTGATACTCTATGACATAATATTCCTTCTTCTTTTGGATCAGATTAATTAAGTTTTGTTTCTAATTATTAAGTCATGACACTTCTTCCAT TCTTAATTATTAAGTTTTTTTTTTCTTGTATCAAATCTTTAATTATTAAGTTTTTTAAAAAAAATTATTAAAAGTCATGACACTTCTTCCA TTCTTAATTATTAAGTTTTTTTCTTGGATCAAATCTTAATTATTAAGTCATTTATTGTTGTTTCACCTAAGGTACCTTCATTTAAAAGCAC CAATTAAAATATTTCACGTGTCGTCTCATTAGAAATTCGGATGCTTAGAGCCCAGTAAGGACCGATATGGGCCGGACCTGGTTATAA ATGTAAGACTCTAATTGGATCGCTGTTCGTCATTACTTACCCATATATATTTGTTATTCTTCTTTTCTCAATTTTCTTTCTCTTTCCATCAC CTTTCTCGTTCACTTGCTTCAGATCTGGCAGCGATTTCATTTCTTCTCCAGTCTTTCTCTCGTCTCGGTTCTATCTCCTGGCCATTGGC GGTCAGGACTTCTCAGCCGTATATTTGATTCCGTCCCCGGATTTACTCTTGCTGTGATCTGAGGTCAAAC
AT3G18480	GCTTTAAGAATGGAGGGTGGGTGGGTATATATATGGTTGATGGTGGATCGCTTCGAGAGATCTACAAATGTTTAATTTTTAGTACATT GCACTTTGGGAAAAATCATAAAAGTTGTAAAAAAGAGTAAAAGACTTTTAAAACTAGTACGTTGCACTTTTTGTTAAAAGGAGGA AGGTTTTTTGAAAATTGATTTTTGTTTAAGGAAGGTGAGAGATTTTACTGATTGTGAGAGATTTTAAGCTACTTAAAAGTAATATGT AAACGCGTTTTTGGGATTGGTTTAATCCACACATATATACAAGATACAAAGAATTTAATTTAATGACTAAGAAAATTAAAGAAATTAT GAATAAAAAAGAAATATACATCCTAAAAATGCGAAAGTAAAATAAAACAGGCTATGCTAAATTTTAAAATATTAAATTCAACATTTTT CTATTCATAAAATTTCTATTAAAATTATTTGTAGATAAGAGATAATATGTTCAAACATTATTCCTCTATATAAGGAATCAAACTTGCAAAC TTTAAAATAAAATTTTGAAAAATATATCTAATGTAATATTCTGGTTTTATTAGTAAAGTAATATTACTTAATACTAACAAATTACAAAGAA AAGAACTGAAGGAAAAAAAATAAACTTTTAAAATTTTCTAACCAGTGAAATAAATTAGAATTTAAACTTAAAATTTTATTAATTAATA ATAGTCATAACAAAATTTGACAAATGTTTTCAACTTAGATGTGCAATAAACATGAATTTATTCTACTAAATCAATACTGGGCCTATGAA TGGCGCAAGGAATAGCCCATTGGTTATTCTATATCTTTTGGACCTTACGTTTTTGAAGCCTGTTTAGGATCACACTCATAACAAAAAA ATAAATAAGAATTACGCTTCTACCACCATTAAACTCTAATGTATCTGAACTCAGAAGATCCGATCGGAGTAGTAAAAATTCAACAAAA CAACAATAAGAAAACTCATTAATTCATTCGTCGGATTCGAAGAAGAAATCACTTCCCAGAAACTTTCTTCGTATCGAAGTATTGCTTG TGTTCTTCGGAGAAACGAGATCGGGATCCGAATACGAGTACA
AT3G52760	TTTGTTGTTCAGTGAGCCTAATAACTTATAAAATTACGATAATTAGTCAGTAGTGAAAACTATTCACAACTCACAAGTGGGGCTATTG ACTTGTTTGTGTAGCAAAGGCTGTTCAAAACTACAAACTAATGGGCTGATAAAGCCTTATACGGCCCAAAACATTATTAATTCACTTA CCGAAAATGCCCTTATCACATATCACGAAGATACGAACACGAAGAATGCGTAGTAGTCGCCTGTTACTGGAATTATTTTCAATTTCTG AAATCAAATCCGATTGAGAGAAAACACTAACTTCAAATTCGAGATCTGAATTAGGGATTCTAGAGG
AT3G54610	CAGCAAATTGTGTTTCCTTGTTAAATACTAGTAACAGATTCCAAACACATTAAATTTTGTGAAGTTAAATTTAACAACATACTATAAAA TAAAACAACACATAAACGGATTAAAGACGTTAATGTTCATCCAAAAAGTATACTGAACTTTTAAATAATTTACCGACGAATTGTACTG TTTATATCCGACCAATATAGTTAATCACATTTTAATCACTGCATAAATTTACTCATAGGAATTACAATTACCGAATTAGTTTCGGTTTAAA ATTTGTTGGATCAATAAAACAAAAAAAAAAAAACGATGCCAGAATATTCCGTTCGTCGGCGTAGTCTTTCCCGCCCCCAAATTATCT GAGCAAAAACCCTAGACTCTCGATTTCGTCACGATTTTCTCAATCACCTTCACTATACTGCTTCG
AT4G13780	AATCATGAGATTTTTTTTAAAAAAGTTTTTTTTAAACTTTGGTGTTATTATTTAAAAATGTTACTAAAATCATCGTTCTCTGTGTCGAA ATCTTTGTGTGAAGTTTAATCATTAAAAAAAAATCTACTAGAGATGAAGTTTTAACGATGAAGAAAATGATATGATTTATCATATTTAT AGAGATTTAACATAAATATTCCTATAATTTTGTTTTATACGAAATAAATTATACAAAGAATCTTACCACTAACGAGTGCCTTTAAACACT TCTTTGACATGTTGTCAAATGTTAGCTTGTGATTATAGACATGATTACTCTATTGCAATTTGTAGGTAAATAATTAAACAATTAAATTTG CCTCCACATTTTTCCCTCTCTGACAAGTATACCTTGGAAGATTCTGCCTCTTAGTCATATGATAGATATTCTTCAACCGTTGTTTAGAG AATTGGTACATAACTCTCTTATTTTAACCATAGTGTTATCCATATGTAGCATTTGAAATATATGTTGTGAAATGATATAATAACAGAAAG ATGAAGAAGTCCAACAATATGAAAAAAGTTCGCTTCAAGTTTGGGCTGATAGACGTGGTAAATCTGTTTTGTTTAACATAAAAGATA GACCTAGTCACCTAGCAAAGAGTTTCTATAGGACGTGGTCATGTTGTAAAACAAAAGAAAATGATTTAGGAGAAAGTTTCCATGTA ATATAGGACGTGGTCATGTTGTATTTTTTTTTTTTGTTAATGGTAATTTCCTAATTAATATAGTCACTTTACTTTAGTTTCACAAAAGTTT AATAATTCACTAAACTATAGTAAAGCTATTTAGATTCCGGTTTAATTCCATCCATACCAGTAAGATCAATATCTTGCAAATGCAATTCAG TTATCAAACCGGAAAGTGACCAACTTTCTGTCTGGTCTAAACTTAACCGGATCTATCTGAGGCCGCCGTTACTTAGGTTTACCATCA GCTAAAGAGCTTCTCTTTTCTCAGCTTCTCTGATTCGTCTTTCTCCTCTGTCTCTGACTCTTCGCGAAAACCCTAATTAGCTTTGTGGT CTTGTTGTTTTCGCC
AT4G16845	AATCAAAATACATATTCAGTAATTTATTTTAAACTCTGACTATTTATGTAAAACACAAATGGAAATCAAAATTGAAGGTCATGAAGATT CTATTCTTAGTGTGAAAAGTATAGATCAATGATTCTTAATTTCTTCATCCTCCACGCATAGATCAATGGTGAATATGGTTTTAAATCCTC TAATACTCACTGTACTGCCATGGTAGAGTTAAAAAAACAATTTTAGAAATATTAGTGGATTAAGGCATTAAGCTGTCCAAGTTGCTTG

	TATTTTCTTTTCATTTTATTAATTAAAAAAAAAGTTCAACTATTTATTGACTAATAATAATACGTGTTAAATGGTTATCGGTTTAAAATAT GGGCCATAGGCCCAGACTTGAAGAAAAACTTGAAACCCAAAGTTTTATTTTTACTTGTTTTCTTTCTCAGTGAATATCTCCCAATCA AGCTTCTTCAATTTTGCTTGCTCTCTCTCTTACACGGCCAATCGGTGTTTTCGCAGCTTTCAGGTTTGTCTCAATCTCAAATTAAATCG GAGTCAAGTAATAACAATTGATAACCCTAATTGTTTCAATTATATTGTAAGATTTGAAATTTTGCAGTAGATCCGGAATCGTATTCTAG TTCTGGAATCGTTGATCTCGATGGAATTTTTTTTAAGATTTCTTCATACACATTTGGTTCAAAAGATCACATAATTTTATTTTAATTTGA TAAGTATGATGATTCTGCTAAGTGGCATTGGATAAAGTTTTCATTTTTGCAATACGTCTAAACTTGTCTATGTCTTGAATGAACTCTCT GAGTTGCTTAAAAAGTCTTGTGCTTTCTTTATTACACAGGCCTCAATACAAGACATTCTATATAAGCATATTGCAGAAGAGGCGGTTC TAATTGTTGCATGGAGTTGAACAATATGACGTAGGGAAATTCTAATTTAGGGGAGGCCTCAGAGTTTGCACTAACTTCATAATCAGC TCTGGACGTTGTTGATTGTATTTGAACAAGA
AT4G34270	TTCACAATTTTTGTGTTTTTAAAAAACCTATAAATAAATGAGTAGTGCACATAAATTAGTTTTGACCAAAACGGCTCCAACTTTAAATA TTAAAAAGATAGTGTGTCTAATTCTATCTGATTCACATTGACGCACACATGTGTTAATAAAATAAAGATACCATTATACAATTTCTGTTT GATACTTATATATGTCAATGTGTCTCAGTATTCTACTACGATATATAATGAAATATGTTTATGCATTAGAGTCTAGAGATTCATGAGGTTT CACCTTTCACTGCTTCCTGATTTTTTTACCACTTTGAATCGTTAATAGTGGCAATGACGACGAATATTTTGTTCATGAGGTGTATCGTG CAAATTAGTTGTGAATTTTTGTGATTGATTCATCTTTTGATGATTTATATGGATATTATTGAGCTTGAGAGAACAAGTGGGCTAGTGG TGTTCCACCATGTCAGGTTTCAAACCTGGTCTAGGTCCTCTAGAGGTTTATTTAGCCAAATCTACTAATAATGAAAAAAGAAACATTG AATATTGACCTTTTTTATAGTGTGTTCCAAAAGCATGAAATTCTAGCTATATTACATTATTATGAAAAAGGCCTACGTTATAAACTTAAA GCCCATTAAGGACCAAAATCAAATTGACCTTCTTTTTTCAAACTAAAATGAAACTAGGCCGGCCCATCAAAAGTCAAATGTCTTGGT TGGATAGTCCGGTTGGTCACACAAGTGACATCGTCGGAGGAGAAAAGGTAGTGCCAATATTAGCCGGCGATTCAG
AT5G06140	ATATCGTGAATCCCAATTATTTATGTTTTATGATTTATTTATTTTTTAAACTAGAGATGAAGGTTTTAATGAAAACTACTCTAATATAGCA TATTTTTACGAGATTATCCACTAAAAACCGGTTTATTTTGAAAGCATAACTACAAAATACTAGTAGTGTTATTAATTTGATAGTATATTT CATAATCTAGTAGTGAAAAAGTTAAAATTTCCATTTTTCTTTTGTCCCAATCTTAAGTGTCTTATCATTCTAATTTTTGTCAAGAATCTA TCTTTAATAATCATATCTAAAACGAACTAATTTATCTAATTTCATCAAATTTTGTTGTTAAATATGACAAAATGAAATTTTGGTTTAAAG TTAAAAAATCTGATAACCTTTCTCTTTTCAAAGGTTACTTACTTGTGCACCGAAAGTTGAACACAAAACTCGTGAATATTACAAATCT GCCACTAGAGGTGAATTTTTGACTAAATCGTTTTTTTTTTTTTTCGATCTCACGGGGTTGATTCTCGTTGTTGTCTC
AT5G08440	GTTAAAGTAGGATATATAAAATAGTTATTAGAAAGATATATGAGAGAAGAATCGAACTTTAGTAAGATACTATAATTACAGGTTGCCA AAAAAATAAGAAATATAAGTTACTATAATTACAGAAAAAACAAATAATAAAATTTAGTGTAATAATAATGATAAATTATTTAAGACAAT AATGAGAAGAAAACAACTTTAATAATTAGAAGAATTGGAACCAAAAGAAAAAATGAGAAACCCAATTCTGAGGTCACAACTCCAG AAGAAAAACTGAGAGAGTTTGCTCGAAGAAGAAGAAGAAGCAAAAGCTTTTTAAAATCCATTTTTATTTTCTTGTATGATTCCAAA CTTATCTTCATCATTCTTTCTCAATCATACCATTCTTTCTTAATCTCCAAGCTCGTATTCTATTTCTTATTCGATGTTTATTTCGTTGAAAT CTCGAAGCTTCTCTTCGGGTAATACCTAAATGTTGATTGGTTACTCGAAAGCCCAGATGAGATTTTGAGAGTAGGCCGAAATTGGA GCCGTGTTTTTGGTTCAGTTTCTCGAGGGTTCGATTTCGGATTCGTGAGCTTTTGGGGACTGGAGAAT
AT5G10630	AAAGAAAAAGAAAAAAAGAAGTGAAGTTTGTTTTGAGTACAAAACTACAACTTCAAACATTACCATTATTAACGATCCAAAACATA TGTTAATTGTTTATACCAAAACTATTTTCCATCAGATTTTGTCTCTTTGATTTTTCGGATCCTGGCGGAAAAAAAAAGAGAAGATTTT TCGGATCCTGGAAATATCAGATATAGGCCCGTTAAAAGCCCAACAATTTCATTAGTTATTATTTACGGTTTTGTTGTGTGTTTTCCATA TTCTTACTCTATCCAACCGAAACCCGAATCCGGATCCAAATCCAGATCCGAAGCCTTCATATCGTCGGCATCAACAGGCGACGAATT ACCTCAAACTCAAAGCTAAATTTACAGGCGGAGAAGTTGGATTCGAAGGAACGTCTTTCGTAAATCAATTCTGCTCTGTAACTAGTT TCTGTAAGTGTTTGGATCAATATTCTTTGTTAAGCTCAAAGTTAATGAATTTGGAATTGGATTTTCAATATCACCGTTCGTAAAATGCT TCAGAAAGATGTCTGAATTCTCTCTTTGTTACTCATTGCTTCTGATCACTGTCGGTTTTGTGGGTTTAAGTCCGACAAGTGTAGGAG CTTTTGCTTCTGAGTTATGTCGGTTTTATTATCTTAACTATATATGTGTTGTGTTTTTTTATACAGAATTGAAACTTAACTTACATACAAT
AT5G15270	CTCTGTTTTTTTGTTTGCCCTTGTGCGCAGAAACAAACG CATATATTTTTTGTTTGTTTGTTATATAACATATTACTCCACAAAACAAACACCAACATAATTTTTCTATTATCAAAATATATGTAACGAA GCTAGTTAATCTTATAAAAGTATTTAAACTGAGTAGAGACAGTAGACACAAATGTTAGAAGATAGTAAAAAAAAAAAACTGATTGAGT TTTAATATTGTTTCTATATCCAAACTGAACAAAGAAACAAGACAAAATCAAAACAAAAGATGGACACAAATATCAAACATGAACCGG TGACATTGTAAAATGTAGGTTATTTTATTTATAAAAGATGGTAGACATTTTACAATGTCACCAGCTCTAAGAAAACAAGGTTATTTTTT TTAAAAAGAAAAAGAAAAAGAAAAACAAAGTTATTTGATTAAATAATTTAATTTACATAAACATAAAATAATATGATTAACAGAAAA ATTCACAACTAATATTTTTTTTGAATTATTTAAATTAATTTATGTTAAATATATATGATTTAGATATTACAATATAATGAATAATATTTTTTT AAAGAAATAACGATTAATTTAAAGTGATTTGAATCATATTTTTTTTTGGGACATCAAGTGATTTGATCTTAACCCAATATAATTTATCTT TGATATCGAAACCATATTTTAAATATGTTAAAATAATATTCTATAATTAATTTCATTCATAGCCAATAGAAGTTAGAACAACAACAGATT AAGCTCAGGGCAAACCCTAGAAATACTGATTACCCCCAAATCCGAAGAAACGATTTCGTCTGGAGATATCAAAGGTTCCTACTTTTC TCTTTTCCTAAATCCAACGTCTCTGAGATTGTTGTTATCGGAATCTCTAATCGTTCCTATATTGTTGTTCTCAGTTTTAGCTACTCGATA GTTTCGTTGCTTGTTCGTTTGATAGACTAGAGTTCCCAGAGTTCTAATTTGATATGTCTCCTGGATCTTCTTAGAAAATTTGATTGATT TAGATTTATGCTTAGCTGAATCTTTGTGAGCTGCTACATTTCTGCTTTAATTGGTTAAAAGAACTCAGATGTGGAGTTCTGATTGTTTA ATTTTATATTTTCATTGCAAGCTACAGATTTGACA
AT5G19680	TCTATTTGGTATAACAGAGTGTTCCAAGCAAAGCGTAGATGATGTTACTATGGAGTGCTGATGGTGAAAGTTTAGACCCAATGTGCT TGTTTGTTGTGTGAGTAATGAATAGCTACAAAGTTGACTCTACCTTTGTAAATATTGACTCATAATTTTGTTCTAATTCGACTTGATGTT GGGCCTATTCATAGGTTTTCAGGCCCACCATAACAACCCTACAGGCAACAATGTTACAAGTGTGCCTGAAACGTTGACGTTTAAGTT TAAGAGAACTTCGCTTTAGAGTTTAGACCAGTTCCCCACAAAGCTCGTTGTCGTCATCGCGCGAGCTTGAACCAAAAAAACC

AT5G20200	AGCAAGAGAGTATAATCCATCTCTTTTGACCAAAAAACTTCAATTGAACTAGACCTATTTTTCATAACACTCTTTTAAGATTTACAAAT TAGAGAAATAAATTCCTTTTTTTCTGTTTAAACTTTTAGGTATTTTATAAGGTTTGTTTTAAAATGATAAAATGTATAGGAAAGCAAAT GCTGCGACAATCTTTCCTCTGCCTAGGCAACTTACCATCGCCCCTTGACGTTAGACGAATGAAGTTTCCCCATGGTGTGTTTAGCAA TTCCTAAGCCGCTTTGATCACTATTTAGTGAGGCTGAAGCGTTACCTTTCTCATTTACAAGCGTGGAACCAAAATCACGCAAACCTA AGACCAAAATGTGTTTGAATATTTAACTACATTTGAAAAAAAAGTTAGTAAGCAAAATATTTTATTTTAAAAAATATTCTCACGAAAA AAAACACCTAGAATGTCATATTAAAACCCTAAAAAAATTTCAAGTAAACTCTAAACGTTATTAATAACTCCCAAACCTAAAGTTCTAA ACAAATGGTTTATTTCTTTATCATTAGAGATTGTTTTCATTTTTTTGGTTAGATAATCTTTTAATTTTGCTTCCACAATCCACCTAACTCA ATTAAAAATGATCATGTTTTTATTTTCCAATTTCCTTCTTCTTTTAAATTTCTTGATTTGAGAATCTGAAAATCGATTAAAAGTATATATA TATACAAGATTACAAGTTGTAGAAAATTGTCAAAAGAAAAACACACATGAACAGAACAACCCTTTGACTTAAAAATAGCGCAATTA CACTTCAGAATATATTTTCTCGGGATAACAATAAACAACAGAGAGATTTCTTCTATAAGGAAGTGGTAAATTTTTTGAGAAAGGAAC ACTGAAACAAAGAGAGTTC
AT5G37830	GAAAGTTCGAGTCAAATTTTTACTACCGCCCATGAGCTATAATTTTAACATGCGTCCACCAATGTCTCAATTAAATATGCTTAAGCTAT CCAATCCCAATATATATCATAGTGTAGTAGCTAATTACGCTAACTCAATCCTAATGTATCAAACTTTATTAGGCTAATTAGATCCGGATCC AAAATTCAATCGGATATCCGGATGCGATAGATTCGAATCTGAGTAATAGTAAAAATCAAAATTGGGATATCACGGTTATTGGCCTACT TTCGCGCTAAACCCAATATCCTGATCCGGTCTAATAGATAAACATCGCACATTCACATAGTGGTGGGCTTAAAATACAGATGATGTAC CTATCTACTAGGAAATTCGAGAAAAGATGGAATAGGAAAGACGTGTTTCCAAATAATCCTATGACTAACCTAGCTCAAAAATTTTAG GACAAAAAACGGAATCCGCGGGATAAAAGGGGGATGGCAGTCTCATGTCCCCCTTGCCGGTGGGGACTCTGCTCCCGGGTTGGT GCCCATGTTTCGAGAAGTTTAAGGCAAATGGTGTGCCCATGTTTTGAGAGGTTTAAGGCAACTTGTAACTATATGACATTTTAGTTT AAACATATATCCCAAACCTAGAGTGTGCCTATAAAAGAAGGAAAATTGAAATTAAAGCTTTAATTAGTCATTTTTCGTCCTTTAAAAA AGAAGAAGAGTGAAAAAGATATATGTTTTTTTGTAATTTTTTTACTTAACACTCAATTTTTTGGTTTTAACAAAATATGATTTAAAATA CTAGAAAATCATTATAATTAGTGTATTTTACAATTTTATATTTTCCCAACATGTAACAAAAAAAAACGATTTTTCTTATGTGCATATTAGT ATGTATGTAAAGTTTCGTAATATGGATATACACATCTGTTTTAAGAAATAATTTTAAAGTATGATATTATAGGAGATTTCATCATATAATC TAATTATATTAATTAAAAAGTAAATAAACAATAAATTTATATAGAAATGTGGCCTCTAAAATGTTAATAATTCGTTATTCAAAATCGAATA TCCCGATCCGGTCCAGTAGATAACATTACACAAATTAGGCCAACACATTCATTCATACCGTGGGGTTTATGATTATATCCGTTTTAAAT TCTAACCAACCAAAAACCGAAATGGTTTAGTAAAACCGGAAAAGCAAAAGAAGATCCTTCACATTCTCATCATCATCACCACCTGAT TCATCTTCTTCATCATTCACCCACTACACTGTTGACTCTTCGCTAAAACTCGATCCTTCAACGATCTTCGCCACCAAAGCAAGGTACTT TCATAAATATCTGGATTTTGATTACAAACCCCTTCTCTCGTAGTTTCATTTTAATCGATACTATCTAAAATTGCTGTTATGATCTCTCAAT ATCGTTGTTGTGAATTGGTAATTTAGTGTAATTTCGTTGTTGGGTTATGCAGAGAGAGTGAG
AT5G42470	GTCTCATATTTCTTCATCACGCACATACTTTTTTTATTTCTTTTATTGTAAAGGGTAAAATCACATGTTCACTATTAACTTACCACTAAAA TGCGGTAGTTTTGGATAGATATTGTAGAATGTAGTAGATTTAGAAAGGGTTTTCCAAAATGTGGTAGATTCAACAAAATTTCTTTTTT TTTAAAAAAAAAAATCTACCACATTCGAGAGTTGATACTTTGTTAGCTTTGCTTTGTATTGATAATTTTTTTTTTTTTTTTAATTGATAA AAAGAAAACTAGATATTGACATTCGGATTTCGGATCGGATTTGAATCGAAGTTTTTCGGATTCAGATAATTTGGATGGGAAAGTTTA TTATCCATTCGGATTTTATGACTATTCGGTTTGGATGATGTTGGATTCAGATCGGTTTTAGGAAAAAAAATTGGTTAAATTTTACAAA AATTTGGTTAAATTTTACAAAAAATTGTATAATTGTAACAAATATGACTAAATTGTTTTGGATATCTAGGTTATTTCGGATATTTTTTGG TATAAAATTATAACTAAATATAACTAATATTTTTTAAAATTATAACTAAATATCATAATATTTTTAGAGGTATAATTTTATTTTGGGTTATCG AATATCCGTTTGGTTTTTGGTTCGAATCAGATTCGGTTCGAATAATTCGGATAGAAGAAGTCATTACCCATTTGGAGTTTTGTGATTA TTCGATTTGGTTTCAAAACCGGATTTTTGGATCTGTTTCATATTGATTTTTGGATTTCGGATTTTATGTCTACTCCTAAAAAAACCCAT GTAGTTTTCGTTTTAAGTGTAGAATTTATCTTTTGGTTGTAAGTTATATCTAGTCTATCTTTTGGGACGTTTGGTTTCCCATGTTTGGAT TGTTTGTTGTAAGTCCTATGTACAACTCCATCTTCTATCGAGAATCTTTTAGAAGGAAAAAAAAAAACTCGTATTTACGAAATTTGACA GTAGAAAACAAAAACCTAAGTACAATTCAAATTGACAAAAAAAAAAAAAAAGTAAAAATTCTTAAAAGGGGACTACAGTAAAAAT TCATGTTAAATCTAACTGCTCGAACGTTCCAGTTTCCACCACCCGGCGCGTGAGTAGTACCACCGTCGTCACTTTCACAATCTCCCAA CAGAGATCCTTCCC
AT5G48520	TTTCGTTAAAATGTTCAATATTTTAACAAAAACAAAAATGGAATATATTTGTGTTTTGAGTTTTGGTTGTGATGTTTTAAAGTCAAGAT AGTAGCATATTTATATGTGCTTTTACTAATGTTACATAGTACTAGTAGTTAATGTTTTATCAAACATATTTTCTTAGAAATCATAGTCAAA GAAACTATAAACTTTAATTTACAGTTTTTAATTTAAAGTATTTCACCCGTTTCACAAAAAGTGTTGTTTAATTTTTTTTTTTTTTTTTA TGTTTTAAATCAAATCTCTGTGATGTAGTTTTTAAAATCAAATCTCTGTTTTATCTATGTGTTTTCAACTTATTAATTAAAAATAAAGTGT AAAAGAAAAATTTAGATGTTTTTATTAATTTGTGTGAAAAATGGTACAGTTGTTAATAGGCTAGAAACTATTTACCTGTTGTTACATG TATATTGGCGGTCTAAGCATTAGGCTTAAGTTGGTCGTTTTGCTCTTTTATATGAATCACAAGTTATAGGCTAGAAACTATTTAGGGTT CTTGGTGTGAGTGAATAGAGTGTTTAACATATCTATTGTTTTGGTTGTGGGGTTGAGAAGATTAGGAATTGGTGACTGACTAATTGT ATTAATGTTGTAAATCTGATTAAAGAGAAAAAGTTCAATTAACGTTGTAAATTTGATTAAGTGATTTTTCAATTGCGTCTACATTTACA TAATAACATAATTTTCTATCGTTCGATCCAAAAAGAAGTTGATCTCTTCCCTCAAAACATTATTCTGCTTCTTGAGGGTCACGTTCTCT TGTAGCATCGATTCTTTTTCCTTCTCCAAATTTATGGCATCAGTTTGGAAACTAGCTATGGTCTGGTCATATAGTCTTATACTATTTTTC TAAGGTGAATAATCTCTTTTTATCGTCAAAATCATGATTTAGGCCCATCAACCATATATGCATAATTGCATTCAAACAAAATAAATAAAG AAATACTTACTTTACATAGTGAAATATAGACGTCGTATGGAAAATGGAAACAACATCAACAAAAAAAAATCAAAAATCAAAACTCTT CACAACCTATCTATTGAAAACGAATGCTTATTGTTTTTTACATAACTTTCAAAATTGAATTAGAGGCAATAACGACAGTGAGAATTTC AAATTTCATTTCTGAGGCAATCGCATAATTCCCACCGGCGATCGGAG

AT5G59710	CTTTTTAATCTGTAAAGAAATAGCGATAATTTAACTTTTTCTCTACAACAATACTATATATTTTTCTTGTTAATTATAATTATAATTATAATA TCATTTTTGTATGCAATCGTCGGTTCGTCACTCACAGTCACATATGCTAAATCTCAAATATATACTTTCTTTATTCCAAAAAATTAAGGA TGGTAGTGGAGGTTGGAAAAAAAGTCAACATATGTAGGTTGTGTAATTTTTTTTAAGGATATACCAAACTTCCGACCATATTTATGAA GAAAGCACGGGACACATGATTTTTTTCGGCATCTGTTCATGCATTATAAATAGATTATCATGAGTTTACTATTCCATGTTCGTTCTGTT GGAAGTTCATTTCAAGTTTTTTTTTTTTTGTCACAATAGTCTCATTTCAAGTTTTTGTTACTAGTTAATTGAGCTATTAAGTTTTATACT TAAAGTATCATACACTACTTAGCCTTGAGATTTTTCATAAATCAAAACGAAACAGAAGCTTAAAGATTTTAAATGAAATGAGGTGGT CACATACATATATGTGTCAAAGATCTAGAAGTTCTAGATTGGATATAAATGAAGAATGACGACATTAACCGCCCCTACTTTTCATCTAA ATGGAACATGACATACAATAGAGTTTATCTGAAAGGATCGATAAGATGTCGCTATATTTACTAAAATTATCTATAATATGATGAGCCCA CTATGTTTTTCTTGACAATTGGGAACTTATAATATAATTCTTGGATATCAACGTATAAATGATCATTTACATAATCTAGTGCTTGTGTTAG TGTCGTGTGTAATATCCAAGTTCAAGCATGTATCATATGTTGTACAAAATAGTTTCAGCCTAAAATGATATTTGTGGACCTGTGTTGTC TACCGACACATTTGTTTGGACAGTGGAAACAACAATTCTATACTATGTTAAAGCATAATCAAGAAGAATAAAATAAAAATTGGTTGA GATGTCGCAAGGAGAATGACTTTTAGATTTGTTTATATTGATTCACTTAAAACAAATACAACATAAAATTCAAATTTACACATATTAAC ATAAATTGAGTTGTATGTTGTGCTAAGATTACTTAGTGATATTTTAAAGCCCATGCAAAAATAACTTATCTTCTAAAGAAACGGCCTTT TTCTTATTTTTTGTAAACTTTGACAAGCTTAGATCATATGATTGAATTGCATTAATGGACGTTATTTTAATTATCTTGAAAGATCTAACG TCATTTTCATGTTAACTAAATACAAAACATCTTAGCATTCATAGCTGAATAAGAGGCATAGTTACTATTCAACACTTTTAGCTTTAACTA CAACCTTCAAAACACACAATTTTGTCGTTTTGTTTGATTTAGGCTTCGTAATCTTGTAAACAGAATTCACCTACTTCAGTATTTCTTCT TTTACACTACGATTTCATGATTCCACAAAGTAAAACATCAACATCATTCTTGATTTGGAAGTAAAATTGTGAATCGTGTATCTAATGGG CCTGGTCAAATAGGCCCAAACATGGATCCGCAAATAAGGAATTTTTCTCCTTACAAAAGACACTCTTTCTGAGTAAGAAGGTTTCGT GGGGCCTGTCGTGACCTCTGAAAGTTCCCTCCATCACCAATCATTCGGCACAAAACTAATATTTATATTTTGTTTCCTTTATTTATTTTA TTTAATAAAATATTTATCGATCCCAATTTTCTAAATCTTGCTTCTTCTCTCGTTCCTATATCCGTTTTCCGATTTAGGGCACCAAATTGAA TCAAAATCTCCGCTTCTTCCCAATTTACGAAACTTCACGAAAAATCTTCTTCGGATTCTTCTTAACCCTTTCTCTCAATTTTTCTTCCA GAGTTTTCTAATTGGGCACTGTTTCTCTTCTTCTTCTTCTTCCTCCGTACGCATCCTTCTCCTTTTTACTGTTGAGTGTGAACAGA TTTAAACAACGTCATTTCTAGGGTTTTCGGAGTTCAATTTTGTTATCTTACTATTGAATTTGCATCGCGATAGCTGAATTTAGGATCCT GTGTTGTTGCTCAGTCGTATTATTACAACTGTTGAATTACGCTTGCGTTTGATTGTTTTTTTCTTTATTGGATCCTCGTGAGGTTTTTT TTTCTTGTTCTGAATCGTTTCCTGGACTTCTGATTTCCTGCTAGAAGTTTATTTCATGGTTGCTGATTCATTATCTTCTTGTTTCTAGAT GAAAGTTTACTGTGTATAAACCTTCAAAGGATGAGAATTTCAATCTTACTTTGAACAATATGTAGTTTATACAGTAGCTTCGTCTGTTG TTGCACCATGGTTGATAAGATTTATGTCTCCTTTGTTTGGTTTATTGAATTCTTCCTTTTTTCCAGGCTATTCATTTCTGAGTAATCGTT TATTCCATCTCAGATTGTAGGCGGTTTCGCGTCTTGTCACCGAATGATACATTGTCTCACGTGAGATTGTCAGTCAGTTAGTGCA
AT5G60160	GCCTTTTTTAACCCTCTTAATATTATTTATATATATGATTTTGTCTGCCAGATCCACGTTGCTGTCTTGTTAAAAAAAAAAAAAAAGG GTTTCTAGAATTTATTACCCTTCGATTTCTAATTTTATTACCACTTAGTCCGGTTCGGTTTTGATTGGTTTAATGGTGTTTGAATGTGTT ACTACCGAAATTACCCAACCACATAAAATCTTCTAGTTTTGGGAATTCTTATGTTATTACCAAAATTATCCAACCACATATAAGGAAGG GTATTTATAGTATATGACTACAAAATTTATTTGAAAGTTTTTAAAATGAACAAATTTTTAAATAGTGCGTTAATTCAAAATTTCAAATGA AACAAATTTTTGTATAATAATTTATTGAAAGTTTTTACATAAAAGTATAAAACGTTTATATGAATAAATAACATTTTTTTTATGTACTCAT GAATAAATAACTATTTATTTAAAAAACCAATTGGGCTTTGTAATGGGCCAATAAAACTTCTCGGGTTACTAAAATGGAAACGGAACC TCCGGCAAGACTTACACAGGGACGATCGGAAACCGAACATCCATTGGAGTGCTGCCTCACAAAAACCCGAAGAACTCAAGTTAAG AGATCCAGCC

Supplementary Table S2

AGI	UTR_terminator_seq
AT1G11060	AAAGGGATCAAAAAGGGCAATTAATTCTTCATTCTCCAATTTCTCTGTACAGCTTCTGCAATAGAAAAGAAAAAAGAACTTGGAATGCTTTGGGGTGATCACTAACTCTGGTGATCGCATTGATGTATGATATAAGATTATAAGGCTGCTATTAGTACGTTAGTTCTCTCTTTTAGAGTTATACTGTTTGTCTCGGCTTAGGTTTTGTTTAGTTGTCCTTTACAGTTTTACTCATACACTCCATTCTTTTTGTAGTTTTCTTGACTAATCCCATATATACAAGTTTGCTCTACCACTTTTTCTTTATAAATCTGTTCCATCGTTATAGTTTATATATAATTGTCATTACGTTATGATATTTCGCAGCATATTACAGTTTTGAAATGTGCATCTATGATCATTGAAGCTGTAAGGAAAAGTCGGCAAAAATTTATTAAATAGTTTTAGCTCTCTGAGATAAAACTCGTCTTTTC
AT1G18700	TATATTAAGCCTTAGGATTCTTCTGACTTTCTTTTTTTTTGACATTTACGTTTTTTATTTTTTCTTTTCTTTGTGAAAGTCATAGGTTTCAATTTTAAGTTACTAAATAGAGTAATTTACCTTTATCATTTTGAACACTTCAAAATATTCGATCCCCTTTTGCTTATACAGTACCACATTTTAGATTTCCTAAATTGAAAGAAAAAACACAACTTATAAAACTAGCAATAATGCCAATATTTTCCATATATGTTGTATCACTCTTTGTTTATACTTTTATATTATTTGATTATAGACACGAAAGCTCTACAGAAACATCTCAAATCCAAAAAAAATAAAATATATTATATATATGCGTGTGTGTGAAAATTTTTGAGAAAAATTTCTAATTTAAAGAAAACAATCATAAACTAATTTAGAAACCCAGATCAAAAGTAGGAAATTTACATAAGACATATCATAGGGGTAAATCCAGAATCGGATTTATTTAGAGATGAAGTGCTCTATTTCTTGAAAAATGACAAAACTTTTTGGGTGGGCCTTCATTTTTTAAAACTTTTTTTGATAACTTTATTGCCCATATCTGTTTGTCGTCAACATTTTTGGAAAATTTCTTTTAAATGATTTTTGGTAATTAAAATCTTAAAAACTTGATTTTTTTAATTTAACATATTGAACTGTCTATATATATTCCTGACCAACCGTGCTGCTCCATATGGGACAACAACCTCTTTAGTGCGGTCCATGAAAGGTTTCCAGAGGCTATGACAACACCAACCTTATCTCGTGATTCAGTTTCACTTCAACCACAGAGTATAAAGGAAAAAGATTCAGACCCATTAGCTGATGGCGATTCTCAGGATATCATGTCTCCATATAAACATTATTTCGTGATTTCTCCAAAAAAAAATCTTGTCACTTCTATCTACTTTTTTTTTTTTTTTTCACTTCTATCTACTTAATTGAGTGGCATGCATAAATGAGTCTATATTTTTTTAACATTAATAAGAGAAAAAATATTTTACACTCTATATATTGTTTTCTTTTGGTACATTGTATACATCTCAACAAATACTAACAAAAGTGAAACGAATTTTGCAACGTATGATTGACGACAGCAATAACAAACGATCTCTTATGTTCTTCTTCGTTATGTTTTCTTCAATGTATGAAACGCTGAGAAATATAGCTTTCCAAAATTTCGATAATTTCTTGTGTAAAATCTTTCTCGAATTTTTATGTTAAACTATAACAAAAAAAAAAGGAAGTACAGATACATGTACTATTTTCATTACAGTATATACATTAACATTGTTAACGTTGATAAGAAAACGAACGGATGTGGCTTTGGCTTATCTCTACGCGTCCATTTGTGTGAAACTTCGTATCAAACTATCGTTATCCTAAAATAAGACTTTTTATTTCTTTTGAAGCACGTTAAAATTTAACGAACAAGTAATTTAAGCGCGTGAGCAATGATGATTAACGCGGCATCAAATCATCAACTAACAAGTGGGAAACTAAAAAGGACAAATTGATATTTTTGTAAGAATTTTTAAGAAATAAGTGAGCTGTCTTAGACCATCATTATTGATGGTTGCTTAGTGTAGTTTTAGGTTAGTTCTTAAGAATATATATATATAATTGATTTTGAAAATTTTGATGTTAAGAGTTTAGTTAAGAGATTGAAATTTTAGTAATATGCAATGGTGGGATCTTAGCTTGAGTTCTTATAGCAAATAAATCAATTTAATCATTAATTTTTACATCCAATTAACCTAAGAACTACACTAAGAACCCTCCAATGATGATGCTCTTATATACTACTATATCTTTTCTTTTTGTTTCTAATAATATTTGCTACGATAACTTCCTTGTTCAGACTATATAGTAACTCGTCCAATTAAGGAGATATATTCTAAAATACCAAAATATATGCTTTCTTAAAATACTATACGTATAATTTTCAACAACAACAACAAATGATTAAGAAGATGTTCTGAAAAGTTTAAAATCATTCTTAAAACTGATATACACAACGATAACATAAAGACTACTCGTCTCGACTAATAAACAACTATGGTTGGTTCTATTATAATTACTTGGTCCCCAACGTTTATTACTCAAGATCCAAACATTCTACCACAAGGTGTTACGTTGCAGCGTGGTTGTCATATCTGTTTTCATGCGGTCACAACCACATGATTGGATTCTTTTTAATGTGGTCCAATATTAATTGTATTTCTTTTATTGAAGAGATAAGACATTGGATGAACGGTAACACCTTTGTCCAGTATTTTTTGTAGCCATCGTAGAAGATCAAAACAATAGCTTTTGAGAAGTTATTATCTTTGTCAGAAAGGAGTGAAGCAACGTTGAGAACAAGAAGAAGAAAAAGTAGAAAAAGAAATTTTATTGCAATCTCGTGTTTGCAAACAATATATTGACTTAACAATTTTTCGTGTTCTAATATCACTGTTTTCTTGTAGTTTACATTGGTCCATACTATTTTTACTCTCTCTTTTTTTTTGCGTAAAAGTAATCATTAATCAATCTCTCTCACTCTTTTTTACCACAAAGAAACAATTCTATGTTGTATCCATCATTATATATTTAACATCAGATTTTAATTTTCTTTTCTATCTTTGATATAGATTATTTTTACTATATCTTTTTCTTAAGAGCAATTGGCAATATATATAATTTTGGCTGATTAAAAGAGACATATAATTTTTATTTTGTAAAATGTACATCTCTTTCATACTTTCATAGTATTTTTAAAATAAAATTTATACTCCTTTGAAACAATATAATGACGATTTTAAAAAGGAAAACTCTTTTGGCCAATTCATATCTTTGTGTTCGACAACATTCCATGGCTTTGATTGCAGATTAATTTGGCCCTTCGAGCCACTGACTTAGGTTAGCTCACGTAAATTTTAGGAGTTTGACATCTTGAAATCGATACTTTTTCTTTTGTTTTCGTTATGTACATACGTTT
AT1G54080	TAAACCCTCTTCACTGGCTCTGAGATACCTTTTTCTGTTTCTTTCTTTTTCTTCTTCTTAATTTTATAACTTTCTTGCTTTTTCTAGACCTTCCTTGTTCAAGAGTCTTTATGTATGTGTCTCTTTCATTTAAAGCCGTTGGTTTTATTTATGTATGCAGAGCTTTATGCTCAGTTTGTAACCTATAGGTCTTACTTGGATTGTAAGCCAAGCAATAAGACAACATCAAATAAAAGGGGATTTGGTTTTCTGGGGTTAATGTTGTTTGTGGTTCTGTAATGATAGGTTTGAAACAAAGTAATTTGTCTTTTATAAAAGTTTTATAGTTTCATTTCTCATGTTAGTACTTTATTTTGTCGGAAATCCAGTTTAACTAAGCATCAAGTGTTCTGTCCACGAGTGTCCTTTGCCATGCA
AT1G54390	GCTCTCCTCCTGCAATTTCCATGCATTCTGTAGCTAATATAATACGTATGGAGATTTGCACGTACTTATGATATTAATTCTCATAGGTGCCACTGGCATCTTCTTTGGCACCTTCCCTTCATGACCAATGCCACCAAACTCATCTTTCTTGTTACCCAACATCTAGCAAGTGCAAGTAGAAGCAAACTTTAGATGGAGAAATCATCATGTTTTGTAAACCTCGGGTTAATTCGTGGAGCATATATTTTGTATATCAATTGTATTTGGGTATTGGTGGGCATTTTCCGTTTTATTCTACTAGTATCGATATTTACATCTGCTTCTTTCTTTCTTTTTTCCCCTCAAACCATGAGCCTTTGTATTGATCAAATGACATGTACAAAGGTTTTGTTCATCAGAAGGAATTTCACTAAAATAAAAAATTGCTTGATATTTTATTCTTGCTGTTTATATATGAAAAAAGAGAATCCATTGTGAGCAGCTGGTTTGATGTGTCTAGATCGATAGAGGGTGTACATATATCAGTGCTAGATAAGTTAGTTACTCCCTTAGTTTGAGAAGTACTACTTGTTTTGACTCTATTAAGGAATACTGATATTATTGTAGTTTAGAGAGGAGAGATGCTTCACAGAACAGACCCACTTCTGCTTCAGAGGCAATGCTGTGTTTGGGACTAGTAAACTAAACTTCTACAAGGCCAAATCGATTCCTAAACATATGTCAAAATGGATCCATTAGATTAGTCCTTTTTAGCCTTAAAACCATTAGTACAGTATAAAATTTGTGGTCGATCTGT

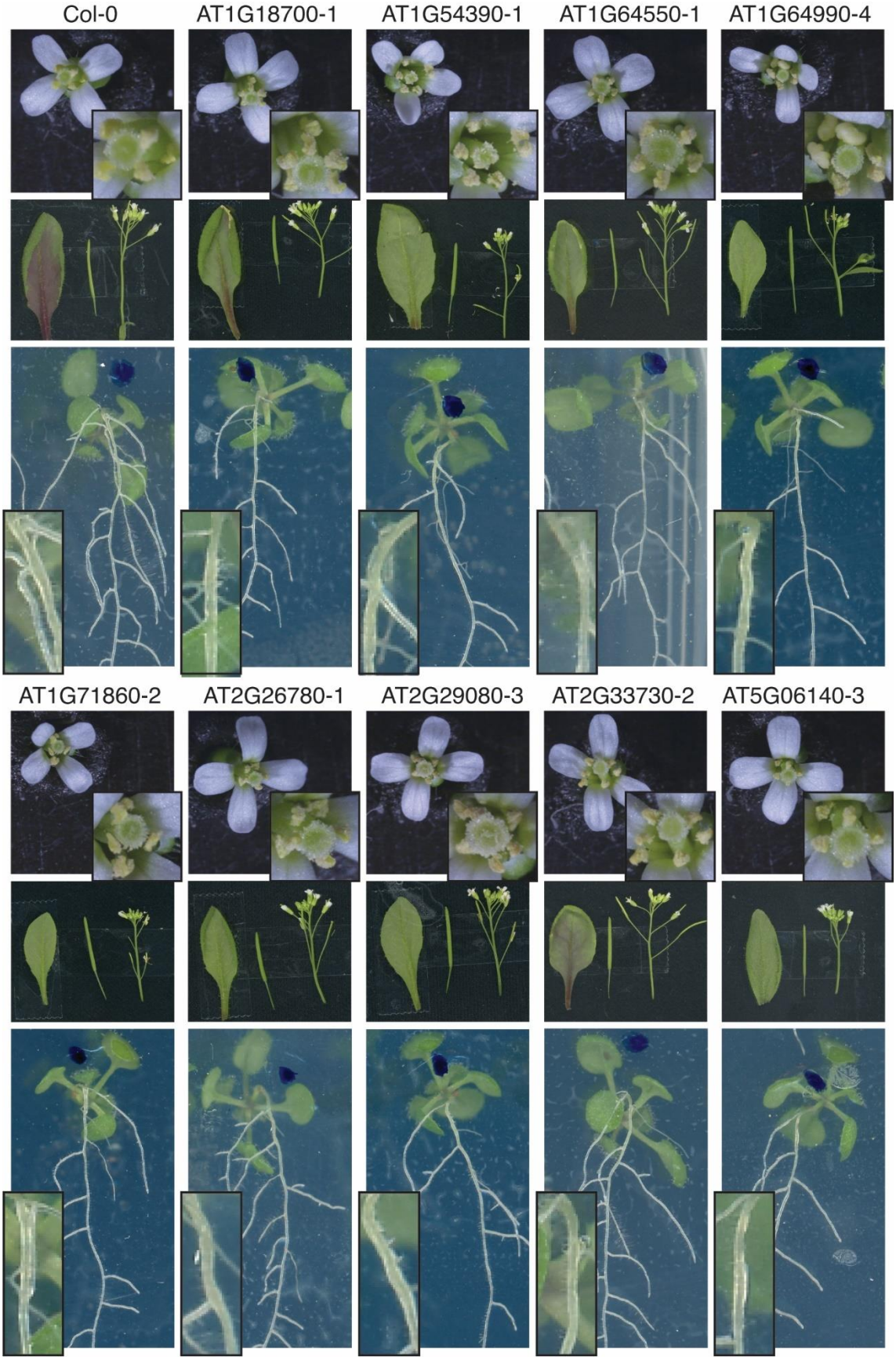
	AATCATCCTCTTAGTATAATAATCTCAAACTCAAATATAATAACACCACAGTAGCATTTATTTGTTCACGGATATTCACACACCGTCTCA ACTAAAATAATTTGAAAATGCAGATCACTAGATCTAGAGACTTGAGAATGGGACCCAATGCAACAATATAGAGCGGTGAGATTTCG AATAGAGAGATTTGAATTTAAAAATCTCCGATCTATTTTTTGTTAAACACGAGAAAATATAATTATTTAATAAATAAAAAAAGACAAA A
AT1G60670	CAATGATCTTGCAAGCAGTTTAGTGACAAGTGCCACACAGGTAAGCAATCTTGCATTTTCTCTGAACGTGGATATTCCTGAATATGT AAGGGATTAAAAACCTCAAATGCTTCCCTTACCAAATTTGCCTCTGTTTCATGAAAAAATGTCTAATCGCTTCAAATCATAGGACTTC CATAACCCCTAAAAATCGGCTTCTAGTATTTTAGGGAACTATGATCTTGTACAAGAACATTGCTTTAAATCTGGTGAATGATGGAAAC ACAATGTTTGATGGTAAAATATAATTTTTTATCTGTCTGTCTCTCAGCCGAGATCTTATACCATTCCCTCGAGCATCTATCTGAAGCAG GAGCCAAGAAACTGATGGGGGAGGAAACTGGATTTTGACCAATGCTAAAGGAACTAGCACAAGAAACCTCTGGTAAATGGCAGC TGTTCACCGTGACCCACATGAAAGATGATGGTTTTGTCAACTGCTAAACCTCTTGCAAAACACAGGATTGAAGAAGATGGGATCTT GTGGAACCCTTTGTGCTGTCGGCTAATATTAACAATGAACCGCAGGCTAGCCAAAAACCCCTTTTACAAATCCTTTCACGTCTCCAT AACTGATTATTTTTCCGTTGCCAAAATGCTTTGGCGGTCAATCCAAGATATGAATTCGTTGCACAGATCCATGTTAAAAGTTGGTTGT ATAGATTAGCTGTTTCTGTAAGGTTTTAAATGAGAGTTACTTTAAAGTTTGTATCAAGAACACATCTTTGTAAGGTTTGGTTTTCAGA GGCATTTTCATTCTTCAAATCTTATCTTCAAATCAGCTACCATATGGTATGGCTCCTCTTTT
AT1G64550	GCTGCGCCGCATTTCTTGTATTCCCCAAATCCATGTTGCGTTTTTGTTTACCCTCAATACGACACAAGAGTCCTTTATACCTGAGAAT TTGGCGACAGTATTTTGCTTTATGTATTAAGGTTGAAACTAAAAAGAAATACTGTAAAATCATACACTTGGTGGAGCACGAAAGGAT TCATATCCAATGTGAATCATCCGTAAAATTGTAAACGAGCCAAATTAGAATAAAAGAGCATGTTTCGCAGAAATTTATGGCAACTAA AATTCATTGTACAGGTACAGGAATAGCAAAAAAAGAACATGTAGGCCCAAAGACAAGATTACAAGTTGTGTTTTTTTTTTCCCACA TATGTTTAACCGAAAGAAAAACCCAAATCTTCGAATCGAGCACCGTCGAACCATTACCGAGGAGGAATATCATTCCACAAATCGATA TCCTTAAGAACCTCTGCTTGTGGATTCGATTGTTCAGAAGAACTACAAGACTCTTCAGGCCGTACTAGCGTCGTTGCAAAACAATG GAGATGCTGCTGCTTTAAACAACTCCGCCATAGATCTTGGTGCAGAGCGTGGGAATGGAAGATGATGACGACAACAAAATGGATG ACGATGATATAGAGATGGATGGGGGTGAAGGAGAAAGTTATAGATGTGTTGAAGGAATGTGGTT
AT1G64990	GTTTCTTATTCAATAATGGTGCCTGAGAGAGTCTTATGATTGTTTGGCCTGACACAACGAAAATCAATTTACTACTAGTAAAGGTGA CTTAGTGTATAGCTTACAATTGGCTATTTGTTTGTTTTTGACGCTTATGTGAAAGATTAAACTGATCAGTTCAAAATTTCTTATACCATA AACGTTGAATTATGTTTATTTGTACATTGCTCCTACTGAGGAAGTAATGATACGATGAAGTCAAGAAGATTTTCGATATTTGTATGTTC GACTTATTCCACAGGATCCAGGTTTCTCCATTTGTTATTTATTAACATATACAAATTGTATTATAAAATGCGTTGCGTGAGTAACCTTAG TAAATAAGTAATAACACATTCAGTATTGGATGCTCGTAAAACAATTTTGAGTAGTAGAAAGTCTAACGATTTTGTTAGATTGAAACTG CAAGACTAGGGAGCTAAGAATATGAATATCTCAATCCCACCAAATGTAATAAAATGTTGAATGAATCATAAACCTCTAGAGTAGGGT CTGGTGGTCTGATTCAGTTAATCAATCCATATTACAAGTCGATGTGGAAACGACCAATCATTTAGACTAAAGAGTCTGTAAATTTTTC AATATATTTCTAACCATGCAAACTTTAGACTAAACTAAAAAAGAGTCATATACTTATTAATGCATGGAACAAAGAAAAAAAGTAAAA AAATAAAAATAACGAAGAAAAAAGATTAAAAAAAAAACAGAGCGTGTAGTGGAGGCCCGAGCACATGGGAAGAGAAAGAAGCA ACGCATAAAGCTCACGCCGACGCCGAGTCAACACGTGAGTCAACTCGAGCCTCTGTTTTTATAATTA
AT1G71860	AGGTACGCAATTTCAACAATCAAATCTGAGTGAGCCATACCTGTTATAATCTATGAATACTCTTATCTTGTCTATCTTTTGATCCAGCTG AAGGGTTCGTCTGCTTAGAGAGGGAAAAAGGCTTCTACCCAACTATATCAATTTCGTAATGTGATTCAAGAAGAAACCCGCGTGAA TCTCTACATTCAAGACTTTTCAATTTCTGTAAATTTCCATATCTGATAGGTTTCATTGTTACTTTTGTTGTGGAAATGCTGATTAAAAA GCAAATGAAGTCGCCAATGCGAGATTCTTCAAAGACTATGAATTCGGTTTGATTTGGTTTGAAGTGTAAAAGAACCGAAGATTAAT TTAATGGTTTAGACTTAACCGAGTTTCTTTCGAAAATAAATAAATCAGAAGCGGTTAAAATTGGCGTAAAATGTCCCGAGTCATCGT T
AT1G71900	AAGCTTTTTCGCTCTGTTTTCTGGTCCACTGCGATAACTGCAAACCAGAGAATCAACCTTTTTTCCTTATCTCGACTTATTATGAACAT GGACCATGGTTCTGTCTAACACTTACAAATGGTTCAAGAAGATCGAAGAAGGAAAGAGGAGCAGCGAGGATGAGATCATCTTCTC CAATATCGGTAAATATGCTGTGAGTTCAAGCACACGCCATTAGCACACACCCACGCCTAAACCCTTGGTGTGACAGTTAAAATTTTA TAAACAAGAGATGGGTTTTGCCTAAATTGGTCTTACAAGCATCTATAACTAAAGGTTTCAATTCAAGATTTGCATTATTTTTTATTGA ATCTTCTATTTTTCTTCCTTAGTTGGTTTGAAGCTTTAAACCCTTTGAAAAACAAGAAATGGAGACTTTGAAAGTTGACGAAACACA CATGTTTTATTATTAGTCAATAGTAATTTGATGAGGTGGAGTTATAACATTGATGATAATAGATAGTGACAACGTTCTATGTTCATCTTT TCCTCAATCAATCACTCCACCGGTAGATGAGATGTCAAATTGAAACTGCGTTATGTCACTCTCATTACTCTTTTTGTATACGTTTCTCA TCTTATAGCTAATCCGAAACCAATTTATTTAAAACTTTATAATAGCATAAAACCTCTTAACTTCAAACAAGTTTCAACTTTCACATATCT TATAGGGATCATGGTAAAAATGGTTAAAAGTTAAACCTGATTTGCATTACAAACAGAGGATTACGTGTATAAGGGCTATAAACTATAT ATTACCTAAAAAAGAAAACTAGAGTTGACTTATTGAGTAATAAAATCGCCTATAAATATCGTAACGAAAGAAAAAAGATCCCAA
AT2G16880	AGGAGAAACTCCTTTTTTGTGGTTTCTTTTTTGTATTTCGATTTTCTCAAAGCATACTTATTACTTGTTACAGAAAGACATTGTAATGT AGTTTTTATGTGATGATGATGATTATTATGTTTAGATTGATTAAAAAATTGTGTATTAACATATCAGGAAACTTTCAAAATGTCGTGTA CTCGTACGTATAATATTTTACTGACAACAAAAGAGAGAAAAAAACGATCTGAATTCCTTTTTTGTAATCGGTGACCGGTTCGGAAG GCTGGACCGATCGGGTTACCGAATTTTTTTTATAATTTTGGTGAATTGTTTCCATTTTTGGGTTTAAAACCGAGCCTAACCGTGACA AATTCTCCAAAAACTACCGAAATATATCAAAACTACCAAACATAAACCGAAATTACCCACATAAACCGATATTGTTTTTTAAACCGAA CTTGGGTTCCAATTGAGAATTATTT
AT2G26780	GCAAGCACAGAGAAGATTTTAGTTTTATGCCTGAAGTTGATATAAATGGGAATTGAATCTCCTACTTTGTATTTGTTTACTTATATACA AGTTCACATTGATATATATATCTTCTAAGTTTTTACTTTTCTTCCCTGATGGTAAAAATTTATTCACGGCTTTTTTAGCCAGTCTATCTAC TTCTTACATTCGAGGTTTGTGAGCTGAGCTCGGAAACCCAAATTTCAAAATTGGCCAAACTGCAAATCACGATCAAATCACCATGTA

	TTTAGAAGCTTGACCAGTTCCTTGTGAAGAGATCGTGTGGCTTGCAGTCATTGCGACTACTACGAATCAAGTGTATGATTATTGTAG AGGACCACTGAGCTTGGGACTTCTCTTCCAAGTCGATCGACACTATAGATTAATTATGGTGTGATACTTAGGTCATTGTAATTTTTGG ATTCCGTCCTCAACAAGATAATAGTTTTGTATATCTCAATCCCTTTTAAAAAAAAAGTGAATCAGTAAAGGCGCGTAAACCAAACCC AAGAGTAATTGAATTTTTTTTTTTTTTTTTTTT
AT2G29080	GAGAAACGTCACTGTTACAACTGTAAATTTGATCAAGTATGCAATAACACAATTCAAAAAAAGCATTCTCTTGTTCTTCAATGTTTTT AGTGACGACGGAATTTGTTAGCAATGAACATTTTTTGTTTTCTAAATTATTGCTTCTAATAAGATTTTTATTTGCTTTTCAGTCTCATTC TTATAGTTTGCGTTTTGTAATAAAGGATTGAATATGATGATGATATTGTTACATCTTCCAATATTTACCAAATACCATCATGTAAGAATG ATGGCATCATGTGTCTAAACATCTTAAG
AT2G33730	GTCGATAGGCTCTCTGGATTTGTTCTTAACTTTCCCAGTGACGTTTACCTTTTGGGATTGATCTCATGCGTGTGCTTAGTCGCATGTT TTATCTTTAAGATTCGTAGTTTGATAGTGAAGGAAGCAGTAGCAAGTGTAATAGTTGAAACTTGAGTACCTTTTTAAAACGATTTAG TACTGGTCTTGTTCAAGAAAAGTATCTTTGATTCTCTGCCTAATTTTTCCACATATATATCTCCTCTTTGTGTTTTAAATTGAAAATTGA TGCTCGAGACGATGCTAAGCAAGAGACAACAATGCAATTTGGTTATGTGGAAAATGACTTCGCGAACTCGACCTTTCCTTTATCCTA AAACACTAAGAGAATGTGAAATATTCTTTTGTAATATCTAATGCCCCGTGGTGGATTTGTCTACAACTACTATGTTGCATTGTTATATG TGGGTTTGTAACAAATAACTCAACTACTTTTACT
AT2G43070	TAACATGATTCAACAATATTCTAATGATGTGTTATCAATCAAACTGAATTAGCTTTTCCCCAAAACATCGAGCATAGCTTAAGCAGATT TCATATGTGAGAGGTACACAAAAGCTGATGATGTAAGTTCTCTGCAAAGCCAAAGTTAATCGGATACCAAAAGTAATTCAAGACAC ATCGGAAAGTACTGTATTATATCGTTTTTTTTTCCTTGTTTTGGGGGAGAAGAAAGAAAGGTTCTCTAGTAAGACTCATGACATGTT GTACAGAGTCTATATTAACCTTCTCTTGAATTCAGACTCGTTTTTGTAGAATTTCAAGTATAATGCAAATTCCCTTTTGTTAATCTTTGT GTGTTTTCTTTTTTGGGAAAAATAAGTCTTTGAAAATTCCATAAAACTTGAAAGTTTGTTCAAATATTATTCGACTGATGGAATGAAA TTCTGACGAAAACGAATGCACGAATCAGATATTTGATGTGAAATATGTGGAATTGACCAATTTGTCTGTTTACGTACACGTGTATAAT AATTGACAAATTAAAGTAGGACACTCGTATGAAATATAAGTATGACCAACTGCATAAATAAGACGTGTAGATAATCCTTCCAACGGA AAATGTGATTATAAAGAAAACAAGTAATCAGAGTCTTGTAATTTTTGTTAATAATGTTGTTGTGAGCGACCATTTGAATTCAATCCGA CATAAATCATGAAATGTTAGAGTTATACAAAAAAAAAAATCTCTCTTACCTGAAAAAAGAGTTATACAAAACAAAAAAGGAGTTAA ATTATATATATTCTACAAAATTCTTTAACCTATATAATGTCAAGTATTTTTAAGACTTATTCAAGTAACTCCGTGTTAAACAAAAAAAAA TCATTTAATTAAATGATTCATAATTCTATAAATATATATTTTTAAAATTATTAGATTTATCATATATTTTCTCAAAAACATAAACAGTCGTAT ATGTATAGGAGATATATAAATACAGAATGTATATTTTATGATAGTTATA
AT3G07160	ACTTTTGGAAGCTGTTATGTTTATGTTTTTCATACATTTATTTGTGTTTAGTTTTCCGAGTGTAGCAGTAGTAAATGATGGGAGCCTCT CTGTAGCTTTAGGCATTTTAAGACTTCTGTAACATTTGTCCCCAAAAAAGAAAAGACTTTTGTAAAACATAACCTAGAATCTTTAAG ATATTACGTAGAAATGCCTTTGGAAAGAAAGCGATTCAAGGAACAAAGATTAATTTATTCATTAACGGTAAAAGAATAAACCTAATT ACTGGATCCAGTCATTAACGCGCATGTTATGTTTGTCGTTGTTGACGAATCCGTGATTTGGATCCTTTGAATGTTTTTGGAAGTATTA GTGGTTAGGGATCAAATAAAGCCCAATTATTCGACAGGCCCATCTAACAAAATTGGGGACCTAATTCC
AT3G08530	AGCAGACCAGCTTATGCTTTAATATCCGCATTTGGAAAAGTATCAGTAGCGCATAGAGTCGGCCCAATTTGCTAGGGCGGTGTTCAT CTGCTAGCTAGCTTATATGTCTGACCAGAATTTGTCTGGGAATAATGTGGATAATAACCTCCTTTCGTTGTAATATTCTGTTGGGTTTA GGTGAGTAATTTTTTTTTAACTTTTTGTCAAAAGCCATTTTTGAGTATATATACTCAACTCTCTTACTCATATGAGTTTTGTAGTTGTAC TTTCCTTCAGCATTTCATTCCATTCGACTTTCGATTTATTAAACCGTTGATCAGTTTCACCTTGTTGTTATTCATATATGAAACAATTAAT ATCGAGAAGGACTAAATAAATCCTTCATGAGAATTTGGTTGACCTCAGAACAATCATCATCATTCATCAGCATCAGAAAGACATAGA TGCAACGTGAATCTCATTTTCATGAAATAATAATGAAATCTTAGGGTAGGTTTATTTCACATGTACACATAAAAACCTAAGTACCAATA GTCATCTGAACATCATGAACTAAAAATTGCATCCCTCCGTAAACACACATTTCTGCTACATTACTAAGAAAACGTTCTGGATAGAAAG ATTCAATCAACCCAACATACTTTGTTTTTATTAGTATACTTAGATGACATACAAATTCACTTGGTATTTCTAATATCAAAAGTTCAACCT TCACCAGGATAAAACTCATCCAGCGCAGACACTAACTATAGATAATCAATTAAACCAAAATGATAAGCAGATTGAAGAACTTTAAGG AGAATCAGTAACTCATAAGGATCATTCC
AT3G18480	GCTCTCACCCTAGGGACTTATTCGTTTTTGGTCGTTATGTTCTTTCTTGTCCTTGTAGTGTTCCGTGGGCAATTCTAAAACGGTTGAA TATTTGTACAGAGGGATCCAAATCACTCGATACATATATAAAGCCACAGAAAAAACTTCCTTCAGAAAACTGCAGAAATCTAAATAA GATCTGGTTTACTTTTCTTCGAAACATAAATAACCATGAAAAATGCACCAAAACGATAACAAATCCGACAGGAATAATTGTTGCTATC CAAATACAAAATAAAATAAAATGAAAGTCGTTGCTTGATATCTGTTAGGCAACAACA
AT3G52760	TTCCTTGCTTTCGAATATTGATAGATTGTTGAACTTTCCGGTGCCATTACTGAGTAGCTTAATGTGATATCTTATAAGTGTTTGTGGTG AAATAGCTGAGACATTGTAAAGGTTTTTGATAACTTTAGAGGAAGTTGTTATTGATGTGGTTTAAGTAATTGAAGATGATGTTCTAA GATCTCTTACTATTGGAAATCTCGAAACGTTTATGCCATATACATCCTTTTGTGGTTCTGTTTGACAGTTTCTTGACATGAACTGTTCT TTGAATTAGATTTGGTGAGAGTATCAATTGCTCATTTTTAGCGTACATGTGTTGCAGTTGTAGTGTAGATTTTCCGAACGTAGAAGC GAGTCACTAAGATGTTTTGATTAAAGATCTTAGGTTGCTTCAACTACCAATGATAAATAAGTGACTGGGGAGAAGTGTGAATGGTT TAGCTTGAAATTGTTCTATAAGCGAGAAAGTAGTTGATCAGAAGATACACATGACGAGCTTTGGCCTGGCAATGGCATTGCATAAG ATAGCGAAACTATCATACTTCGATTGTCTCAATCACAATTGATGCGTTCGAAACAAATCTAAACGCAAACAAACAATCTGAATTATAG AATTGTGGCTTGTTTACGTTCATCATCATTCATGGTTCCATGGACTTCACTGTTCTACATGAGCAAGCAAAAGACTCATAAAATTGTG CAAATTTTTCGTATGCTTATTATCCGGCTAAAGCCCAGAATTCGTTGTATGGGCTTTAAGACCAACATTATAGGCCCATTTATATTAAA CATTT
AT3G54610	AAGAAGATGGTAGAGGTTATATTAATATCATTTTCTAGATCTTTCATTTTAAGATACTTTCATTGGAACTGGGAATACATGTATACTATT TTAGGCCCATGTATACTGGAGCTTAATTGTTTGTATGTTTTTTTTTCGGATAGGTAAATATTTGATTTAGTTAGATTCAGAATGTTTAA

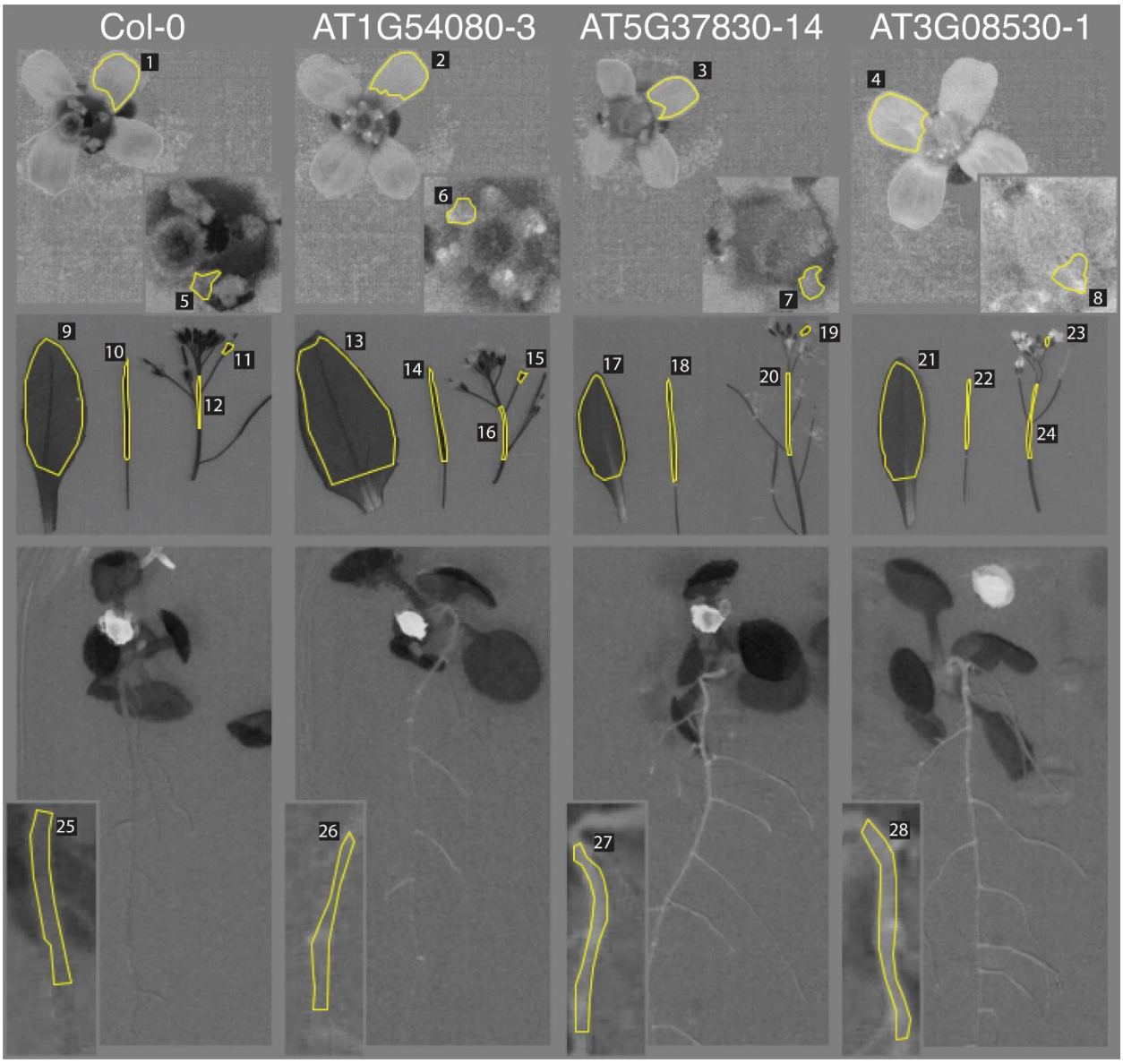
	GTTTTTACTTTGTAGGTTAAATTTGTTCTATAAAGCTTATTTCCAACTGTAAAGACACATTTCATTTTACTAAATGTTTCTGATGAAAAT CACATAACCAAATTTTGATATCGAACGGGATCACCAGTTTCGGTGGTTTTGATCGGATCAACAATGGTGAAGAGAATGAGCGTCAT GGGTCATTTGTGTTTACACGCTTACACTCACGCTTACGCTGCACAGATTTCTTTAATCTGATGAATGTGAGTTGTTGTTGCTCAACAT TTCGTTAAAACGGTTACGTTTGCCATAAGTAACT
AT4G13780	AAGCTTACTTGAGCCATAAGCAAACTAATATTCTCCATTTCTATCTTTCTCTGTAAGAGAAACGAAATTCGGAGATTTTCATTTGAAA ATTTGATTCTAATTACAAACATGTTCTTGTAAACTTAAAGCTTGGTACAATTGTTGTTATTGATTGATGTTGTGAGAAACATTCAGCTT TTTAAAATTGGGCCCAGACGCAAAGCAACCACACAACTTACGATTTCAAATCTTTTATGGTACTAGTTGAGGTTTTGTGAAGGTTG ATGTTTTTCTTAATTTTTTAATATTCTAGTTCTCGTGTTTAAAAATAGTCGCATCATTCTTTTTTTCTGGCGGTTCGTAATTGTGTTCCCT TGTAATACTGTATAGGGTTGTATACAATCGAACGTTTGAAGTGATCATGTAATTCAATCATAAATTTAAATCTTTGGTTATTGAAAACA TTTTAAGCATTCTATTTTTATTGACCTTTAAGAGATTTTGGCATTTTATTGGCTTGAGTTCACTGAGCGTTTCCAAGTTACCAACTACA CCAACCCTTCGGATTCAAATCTGTGTTGATGTTTTGCGATTCCATATCATTACTATATCAACGAGTTGATTTTTTTTTTTGGTCAACCA GCAAGTTGATTTTTGAAAAACTCTTACTCTACAATTTGCGACTGACGACTGTTTAGAATTTTGGGTACGTACCAATAATCCAATATGT TATATTTTCGATATAGAAATGGATTGTGATTGATATAAGTCTTGGGTTCGTCGAACAAAGTTAAACAAAGATTTGTGTAAATAACAAA TAGATAAAAACAAGAATGAGTGTGACTGCATGAACCACTAAAACATATACGACGTTAGAGGGATGTTGGGTAGGTGAAGGTGTTT ATTAATTCTAACACGTGAGGTTTCGTAAGTAGGTATAAGGGAGTAACCATTTG
AT4G16845	TTAAATAGGAAACACTCCGGTTTAGATGATACCGATCTATCGGATTGTAACTTATTCTTCTTTCTTAAAAAAATTGTTTAGGAGCAAA CAAAGATTTTATTTGTTAGTGTATTCAACTGATTACATTTTTAGTTAAAAAAATGGATTCTCCTTAATAACTAAAGACTGAAAAATAAG ATAAGTTTCCTTAATTTTTCTTTTTGACTTGAGAAAAAGCTCCTCTAGACCTCTAGTAAATAGGAGTTATATATTAATCAAGTACATAA CATAAAAATATATATATTAAGTGCAAATAGATTGAAAACAAATCAAGAAATTA
AT4G34270	GCTGTTTATAAAGTTGCAAATTATATCTAAAAGAGCGTACCGATGAGCTTTCTTTACAGCCAAACAAACTACTTACTATTGAATTTGT TAATGTAAGGAGTTACTCCACTGACTTATTGTGGTGGTTTCATATGAACATTCAAACCCTTTTCACGTTGTTCTTACACTCTTATTGTT CCCTACCTTGTTCCCACCATAATTCCCCCTCACTGTTGAATTTCTATAATTAGTGTATACACAATAACGCAGTGACTCGTATGCATAAGT TCAAATGTTGTAGTTATTTAAAAGCCTGCGGCCAATTATCACACTT
AT5G06140	ATCCAAAACTAAACTTACCTGTGTCTGTGTGTGTTCCTTCATGAGAGGGTTTGTGTCTAGAGGCAGTTTGTTTTTGTGAGAAATTCT GGTGTTTTTTTGTCAACTTCAAGTTTAGTTAACTTGGCTTGTTTCCACTTGTAAAGAAGATACATATGCGTCATTATTACTCTCCTAAA CGTTCATATAACAATTTTATTTTAGATTACAACTTTCCCAATTCACTAACATGATA
AT5G08440	GAGTTTTTGGTAAATAGTATTGTTGGCTCTTTTGTTTCTTCTCTGGCCTCTCCAGCTTTAGAATATTAGAGATAGTCTTATGAACGATT CAGAGTTATTTATGTAAGGCTTCTGTAGCTGATTAAGAAAAAACTATTCAATCAATTTCAGTATATGAAAAAAAAAATATGAGCAAAA AAAAAAGTAGTAATAAACCAAATCTTTTGTTGATCTGAGTTCCACTTTACATTTCATTACAAATGTCGGTTTTCAGAAGAAGATGAA ATGAAATCATTTCTAAAGAGAAGCCATTCTATATTCTAAGTCAACATGTAATTACCAAAATGCCGTCTTTGGCATATGTGATGTTCATT GAACCACATCATACAAAGCTAACAAAGCCCAACATATGTGATGTTCACTGATGGCATATGTGATGTTAAAGTGGGCTAGCAGCCCGC GGGCCTGAACGGGCTAAGTATAAAAAGGAGCTTAGAGAGGAGGTTATGCAACTAGGAGTGAGCTCTCGCTATATGTTGATGTTCT TGCTGCATGATGAAATTCTAGTGTGTTGCAGTTCGGCTTGAAGTTATGGAGAAGTGTTAAAAGCGGCCCGAAAATTGGTAGTCCCG TGGGGGAAAGGCTGCTTTGGTTATCCATCA
AT5G10630	ACATTCTCTGTTGTTGTACGAGTTTTACTTGTTCACTATAAATCAAGTCGTCCCAAGTTTTGTCTGCAATACAGTATCAGGAAATATCC TAAAATAGATCATGCAAAATTCGAATCAAATTTTATTGTGTAATTCAATTTATCCATTTTTTTTTGTGTTTAACACAATAAATCTTTCTTC TTAAAAATAAGTCTTTTTATTTGCAGTAATTCTGTAGACTATGATGAGTTGATGACACTATATAATCTTCGTAAGGTAATAAACTCAGT ACAAAAAGAGTGGAGGTCAAAGTTAAGATAAAATCTCTCTACAAGATCAAAGTTAAGACAAGTGATACGAAACATTCTACAATGCA AAACATGAAAACATAATATTTTGTTTTCTGCAATTTCATGCAAATAACTTGAGACACTTTTTGAACAACACTGATTTTTACATATGAAT TCATCTTGCAATTTTTCCTAACATTTCTATAGAGTGACATGTTAGTATTCCCGTATAATATAACTCAG
AT5G15270	AAGCTATTGGTGGAGGTGATTAAAAAGGTTAGATTATAATATTGTTTGTTTAAGTCCTTTTAAACTTTAGTATACCTAGCTATATACCC CAAGTTCATGAGATAATCTGATCTGATCTGTCCCAGTGTCAGAATCTGATTCACTGCTTTTTTCCCTTTGTGACCTTTTCCCCATCCCA CTAAAACAGTGTGAGATCCTGAAGCCTGAAGCCAACCCACAGCCCACAGTTGAAGCCTTTTCACCAAACAAGATGACAAAGCTGA ACCACCAAACTCTACCTAGGAGAACTGATTACATGTGGATGAGATATTTAAGTTTGTTTCAAAACTTTCCATGTCCTCTTGTAACCAA AATACTGCCAAGTTCCTACCCATTAAACCTTATCCAGTTGGCCTTGGCTGTAATGCTGTCTATCTTTCAGTTTACCTTTGTCTTTCCTT GGATTAGATTCAGAAAACCTTAGTTTGGTTGTGCGTAGTTCCGTTTTGTATGAATTATATATGTTAAAAGCTCTCATGTACTGGTTAA GCTTACAGCACGCTTAAGCCTCTCTCCCTGACAGGTTTCGGAATCACAAACTATTTTACCTTATAATAATCTCTCAGATTTTTAAAAG ATCGTATGCAGGGAAAGAGATAAGTGAGAATCTCTTGGTTCTTGTCGACATTGTTCAGGTTAATCAACCACTTGGGCATATCAATCT GTTGTGCCCATTTGGTCCACATGAAAGGTATATACAATTCTTCTTCTCTCTTTGTTGTTTCTTTGCACATATGCTTCTCTGCTAGCAGTT CAAAGACAGGTTGCTATTTCTGGAAGATTGTCGTTTTTGGAAACCAATCTTTTGACCATTAGATTGTCGATAGCTCCATCTCGCCATC AGTTTGCGCGGTGAACACTCCGAAAGATTACGATCCGGTTTTCTCGACTCAGCCACCCCGGTTTTTTGCTGCCTAAATTAAAGAAA TCTCTTGATTTCTACAATGTTTAATGATGAGTTAAATTGGTAGCAAGAACTATATTTTGTTTCAACTTCAAGCAGTTAAAACTTAAAAC TACTGAAACAGTTTTTTTTTTTTTTTTTTTTTTTTGTAAAATTATGTAGATCAAACAATATTTGTATAAATTTTCCATCTTTCCATTTTAAT CCAAAGGCTTGTAACATTAGGCCTGTTTATGAATAAGTCAAGCAGTCACTTCG
AT5G19680	AGATGGATATATTCCTTGTCGTGCTGTTTGTATGAATGTCTGCTTCCTCATCCCATGGGCATGGGCATCGAATTTAGGAAATCAGGTA CAACACGGATTCCTCAAAATATACATGTGTCTGCTGTTTATCATAAAGAAGTGCAAATGAGCAAATTAAATCGCTCTGGTCATGGATA TTTATTAAAACGTTTTTGAAATTCTAGTGTTCCCTTGCAAGTAAACCAATTCTTGTTTTTGGTCTGTAATAGTATGTGGACTGTGTCTG

	TAAAAGGGAATCCCCGAGTCTGAAGTTGTCTTTGTTTCAAGTTTCAACCGAGAACACTAACTTGAATACTTTGATCCATATTTTTGA CTTTGCCCCTTGCTTTTTTATCTTTATAGTTTCATGTGATTGTTCAGTATTGAAAGGTTGCTATATTACATATTTTACTACTTTTTGGTAA CTTGTTACAAGTTTTGGGCTTCTAATCTTTCTTTCCGACAGAAGTACTTTTTGTTAATGAACTGAAGAAATTTGCTGTATAAAAGAA GGTACAAGAGTGGTATAAAAGAAAAAGAAAAAAGTCATTTCACCAATACAGATCTGCACTTCTGCAATCACATTCTCTAGAGAGTT AGCCGCAAAGTATCACCTTAAATTATCTTACGTACCAAAATATTAATGGAAGTAGTCTATGTTTATACTAAACCGATTTGAACCGAAA ATGTTTATAATAATGGTGGTATTTTGAATTGAGGTT
AT5G20200	AAGACAAGGACAACGGTCTTTGTTTCTTGTATGTTTTCCAAAAAAATTTCCTTTTTTGTTTTCTCGAATTGTTTTTGCTTCTTTACTTT TGAGAGTAGAATTTTGTGCCTAGATCATGAGTTTTGATTTCTATTCATTTGAGGATTGATTTCTTTATGAATCCGATTTTTATAAAAAT AAAAGGTTAATCGAGATTTATAGGTCGAAAATGTCATTTGATTACTTTAAAATAATACTAACTTTTTACGTTTTAAGTATT
AT5G37830	ACAATTCATCTCCTCCTCATAAGTCTCTTCTCGAACTTCATTACTCCTTCATTTAACGTATAATCCAAATTAGTTGTTTATGTGTCTTGTT GTTGTTTCGGTGTCTCTTAGCCGGATATTCTTTGGAGGTTTGGCCATTTTCAGTGAAATAAGGATCAGATTCTTGATTCATTAGAAAC AAGGTAATAAATTGTCTTGGTGAATAAGTTTTAGTTGTTGTTGTATGTACTTATGTATCACTATTGCTGGTCCAAGACTTTTGGCTGG ATTTCATTGATGCCATTGTCGACATTGAAAAGAGATTTGTGTCTGCTATGATCGGTCCAACAAAACTAGTCGTCTGAAGCAACCAAA AAAAAAAGTGGCCCCTTGTATAATTTGTCGTAGTAATGTATTTTTTCCTATTTTGTGTAAATTTGAATTCTAATTAATTTTTAAAAGGA ATACAAAAGAAAATAAAAGGCAAAAAAAGGGATAGAAAGATGAGAAGTTAGGTAAGATACAATTAAATGAAGGATTTTGCTTTTT TTTTGTCCACCATATTAAAGAAGGAAAATTAGAAATATGTGGCCTTCACAAACTATTTTCTAAAATAAACGATATTGTTTTAAATGAA ACCATGGGCTTATGAATATTGCTCCCTAAAAGGCTAAAATCCATTATATTTATCAAAAAATATGACCTTCAATAAACTCAAAAGTTTGG TGGTCATACGCTTGAACTTGCCCGTCTATCCTACGGACGGGATCTAGTATTTGACATTTCAAATGTCACATTTGTTTACATATAAGCTT TTAACTCAAAACTAATTGAGAATGAATATAGAACATTATCTATTAATGCATAAATTACCCAACTTTTATGTTGGATTTTTTTAATATTTTA ATAAATCTTCTTAAGATATGGGTGAAAAACGGTCCAATAAAAATGAATCAAGTCCATTATCTCAGAAATATTACAACAAAATGATTTA CTTGTACATGTCATAGATGAAAAATCGTCTATGTGAAAAATCGTCTGTTGCGCTAATATTGGACTCTGATACCATATTAATGCTATTTTC CATTAGGGAGAGTTCAAGAGAAAACATAATATGAGAAAAGATATTTTCTTTATTCAAGTCAATGTATACATGATTAAGCTTAAATACA TAGCAAACTAAACCAAAGATATTAACCAAATTAAACCATAACTAAACCAGCTATATAAACTTTAATATTCCTCCTCAAGATGAAAGGA TTGGGAACGATGAACCTACATGATCAAAACTAAATTACAATGATCCTCTAGTGGTCTTGGTAATGAACATTAAAGAAATCCAAGTAT GCATTGAAACTAACCAGAATGAGCCACAATAATAGATTCAATAAAACTTGGTTCTGATACCATATTAGGAATCATGGCATCCAACTCA AAAACAATTGAGAATGAGTGGAGACACCCATAAACATTATATACTAGTGTAGAAATTATCCAATTTCCCATGTGAGACTTTTCTAGCA CTCTAATAGTATGAATTAGTTAAAATAAAATAAAATATGACAAAATCTTTCTTTGTCAAATCTGAGTGTTAGAAAATTAGAAAAGTTTA CCCCACAAAAAGCAATGGTACAATTTCCCAATTTCTTTCTATAACAAATAAAATAAAATAAAGAAAAAATTAGACTAAAATCTTCGTT AGTATGTAGAATTACAAAGAAGAGATATTCAGCACACTACGACATATAGCAAGATTAAAATGATCTTTCTTTTGGTGGAAGTTGGTC ATAATCCATAACCATTACATATTGGGAACATCTTCTTCTTTATGCTCTTCTTCATCTTCTTCAAATTTGTTGTCCAATTCGATTGGTAGA GATTTTTCTCCAAGACAAACATGATGAGCATTCATACGAGCGGGAAGGGTTTCTTATGGTTGTTGATACTTGTTAATGTTCATGAGA ACCGATTCTAACAAGCCCAACAAGCCTTGCAAAATATAGGTCGGACTAGGCCTTGCTAGCATGATTAACATGCCTAAGTGTCGTTGA TTTTCTATTCTTCATAGGTAGTAGAAAAGCTAGAAAATTAAAAAAGAACAAAAGAAGTATCAAAAAACCTAGATAGTAGGAGCATC TTTATTGGTTCGAATATCTTGAAATCACATAAAAGTAAAGTAAAATAATATTTTTAATCATATATATGTATATATATTATTTTAATAATTTAA CTATGAATGTCATATGTTATCTAAGTCTCTATCTTATCTTCTTATTTTTCCTCTATTTAACATTTTTTTAATAAAAATATCTACTTTACATATC TTGATAATGATATAGCTGCTCTAATATACTATAGATTAATTATAAGTTTTTTTGATAAAGTAGATATTTTTCAAAAATAATTGGTAAAGTA TGTAATCTTTTATATGATTTTGTACACCGTTGAGTTGAGTTAAAACAAAATAGAAGATTACGTATCTACAAATATATAACAAAATTAGG TTCTTTAGATTCACAAAAATAAAATTATTATTTTAAAATGTTACTGTGGTAAATATTTTAGGATGATTTATAACAAAATCAAATTAAAAA TGTAAAGGAGAAAAAAAAAGTATGATTTTATAATTGATTTTTGTGGGATCTTTATGTGAACAAATGAAAAGAGTAAAAGTTTTTGAT AAAAAAAGAAAAAAAGAGTAAAAGGTTAAGTTGGTCATGTTACAATTGTGTTGTATTATAAGAGGAAAAGACAGCACTACCTATTT TACAATTTTTTTTAAAGTACATAGTTACAAAATTATTTGTAATAAAAATGCCTAATTTGCATAATTTTCTAGTTTATGTGGAAGAGTACG TAAAATGGTACATGCACATTAGTAAAATAATTAATAACTTTGCACAAAATGTTCCATTTCACTTATAATATTTTAGGCCGACACTTTTAT AATGGTACTTTTTAATTTTACTAGGTAAGGTCCGACTAAATAAGCGGATGTAAGATATAAATATAAATTTGTAT
AT5G42470	TGTATCTATTTCGAAAGTTAGTATTGTCCAAATTTTATGTGACACTCCTCTTGCTATGTAAACTGAAATGGATTACTGTTCTGCATTAA GGAGAATACAATTATAAGCAATTTGTCTTGATTTCAACAAGATTTTGCTTGGCTATAGGATTCATTGGCTCTGTTTGCTTTTACATTTA CATGTCATAATAGTTTCGAATTTTACACATTTCAGTTGGATGTTAAGAAAAGAGAGGGAATTGATGGGGTTTTGTGGGTTTAAACTT TAAAGTAGTCAAGAATTAAGTCATTGGTTTACTGTTGCTCTATATGTGTAAAATGAAGGCAACTCCAACGGTTCTTAGGTGGAATAG ATTA
AT5G48520	GTTCTCACTTCTCACCATAAGTACATAGCACAAGTGATCTCTCTTAGTTTTTTGTCATTTCACTTTGTGATTGTATTATCTAGTGGTTGA TTGATTTTACTTACTGCCCATCTTCTATTGTTTCTGAGTTTCGTGGTTTGTTGGGGTTATTTGAGCTTTTCTTCTTTCGTTAGTATTTTC AGTTGTATAATTCTTCTTTTGCCTCTTATTTTTTTGTGTGTTACATTTGTTCTCCTACTTTTGGTTTTAAATTCATCATTCTCAAAAAAAT AATTTCCAGTTCAAAAGCAATACAGGTTATTATGAAATGAGGGCAGACAGCCAATTAATACACGAACTATGGTTTTAGTCACTACTA TAAGTAGATGAACTATAGAGATGTGTAGATTCATACACCAACTTCCAAATGATGTCAGTTTGTGTAAATGAATTGGCCAAATGTTTAA ATGTGATGTAAAAACTAAAAATTTCGCAATATGCTATCGTAACAAAATGCTAGAATTCGTGTATTAATTGGCCATTTCCCCATGAAATG AATACATACATGCATCTATTGAAATGTCATCCATTACTCACATTTCTCAAAACAATTGATGTCAAGATTCGTGCTCGAATCTGTGTTGT TTAGTTTCAAGATGTCC

AT5G59710	CGACAAGTAGGTTTGTTCAATTAGTAGCATCTTTACAATGTAAAGCTTTTCTCTTCATTTCTCTTTCTTTCTTTCTTTCATGCAAGTTCTATTGATTAGAAACCAAAACCATGACATATCACATTTTCAATTTTTAACCCCAAGTTTCCCAAAAAGGTAATCACAATTACTTTCCATGGTCATTTCTTCTGGTTCAAGCATGACATGAACAGGCAATAAATAAGTTGAGATTTTGATCACAGTAACTGATACTTGAATCGAATCATTTAGATTTTTTTTTTTTTTAGTTTACTTGTTTAGTAAATATGTTGTCTATGTTTGTCACAAAAACGTGGCTCAGTTCTTGTATATATGGAGACAAAAAAATCCATTAAAAGATTGTTGACATTCTCGGAAATTTAGTGCCAACTGTTATTGCGAGAACTTACTATAGTTTTCCTTTGGCGAAAAGCTAATAATCTTAAATCTTGATTTTGTCCTCTTTTCTCTGAGTTAGATTTTCTTAAATTCCACTTCCGACCTATTAAGAAATGGGCTTTTGCAAAGAAGATCCGCTTCACTGAGCCCGTATCTCGAAGAGGATAA
AT5G60160	GAGATTCTTAAATCTTACTGCTGTTTAAGTTGGAACTGAAAAGATCCTTTGTTAGGAAAAACTACATTAATAAAAGGATTTATAGCCATGTTATGTGTTGCCTGCGGTTACCATCAATTCACCCTCCCGAATATTCATCAACTTATTACTTAAGCGATTCATTTAATTGAGTGGTTTATCTCAATTATCTTGAAATGAAGAAGATCAAAAGACATATATATGATACTTTTGTTAGAGCTTCGAAAAAGACAAAGACAGAAAGATAACAACTAGCAATGCGGTTTCAGGCAAATCCGTCATGGTCCATGATGATGAACTAGCCTGATCTGGCTCTTAACAAATCTCAATTTTGG

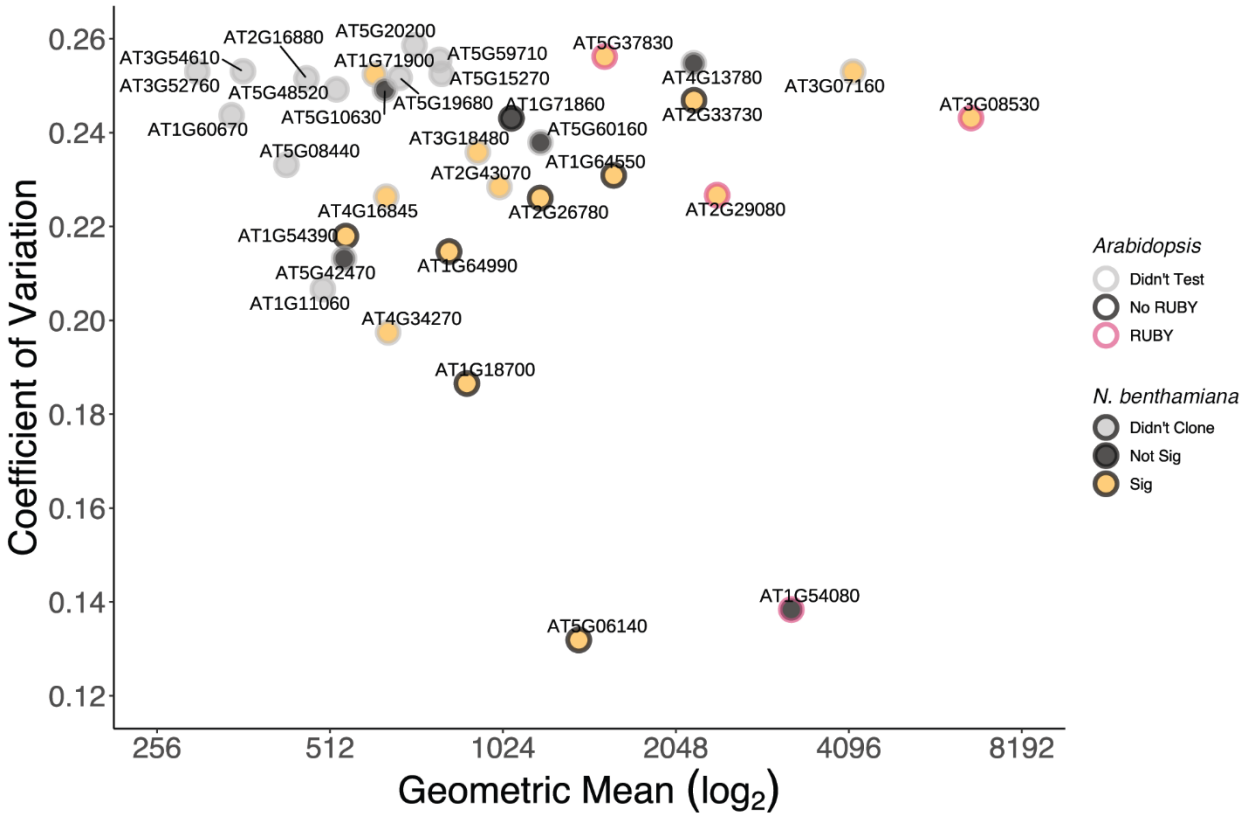

Col-0
AT1G18700-1
AT1G54390-1
AT1G64550-1
AT1G64990-4
AT1G71860-2
AT2G26780-1
AT2G29080-3
AT2G33730-2
AT5G06140-3

Supplementary Figure S3. *Arabidopsis* T2 plants transformed with various promoters driving reporter RUBY. The flowers, siliques, and leaves are captured on day 34 while the seedling images are captured on day 12. The inset boxes are zoomed in pictures of their associated images. Areas where there are RUBY expression visible by eye is marked by the red arrow

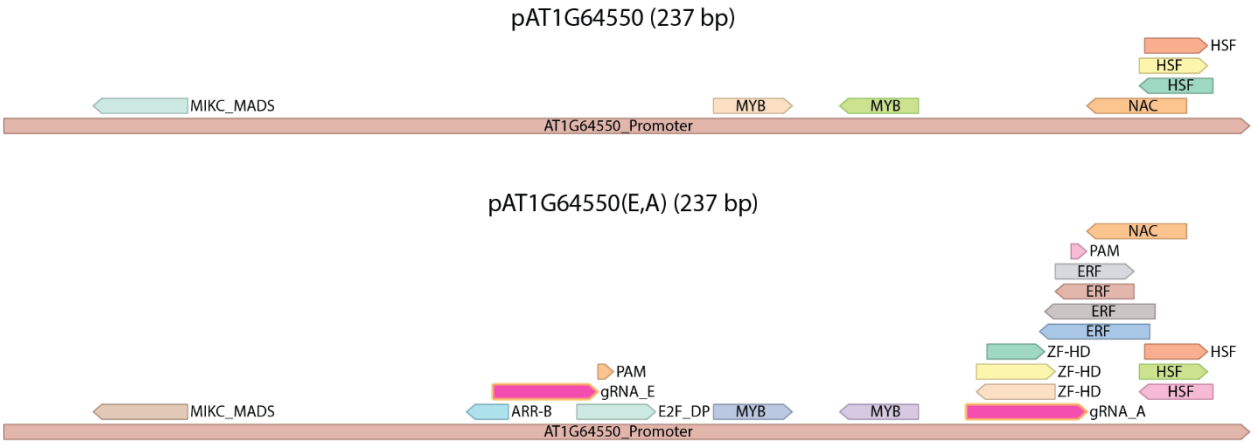


	Col-0	AT1G54080-3	AT5G37830-14	AT3G08530-1
Petal	1 : 148	2 : 160	3 : 158	4 : 184
Anther	5 : 107	6 : 147	7 : 131	8 : 175
Leaf	9 : 61	13 : 61	17 : 62	21 : 71
Silique	10 : 34	14 : 40	18 : 71	22 : 88
Sepals	11 : 42	15 : 48	19 : 72	23 : 75
Internode	12 : 43	16 : 43	20 : 70	24 : 76
Root	25 : 89	26 : 113	27 : 127	28 : 131

Supplementary Figure S4. Quantification of red intensity of RUBY expressing *Arabidopsis*. Figure 3A was converted to CIELAB color space and the a* axis was extracted and presented here in grey scale. Lower values in a* axis are greener while larger values are more magenta. ROI are highlighted in yellow, and the corresponding mean intensity of each ROI is shown in the table below.



Supplementary Figure S5. Final 33 candidates and the summary of experimental results. All final 33 candidates that passed through the pipeline were shown. Y-axis is the coefficient of variation, and the x-axis is the geometric mean on a log base-2 scale. The points were colored in grey if the construct wasn't cloned, black if the construct was cloned but showed no significant difference from negative control in transient *N. benthamiana* infiltration experiments, or yellow if the injection was significantly different from control. The points were outlined in grey if the RUBY construct wasn't tested in *Arabidopsis*, black if the promoter did not give visible RUBY expression, and red if at least one part of the tissue displayed visible RUBY expression.



Supplementary Figure S6. The introduction of gRNA_E and gRNA_A target-sites (highlighted in pink) in pAT1G64550 did not disrupt any predicted motifs but introduced additional ones.

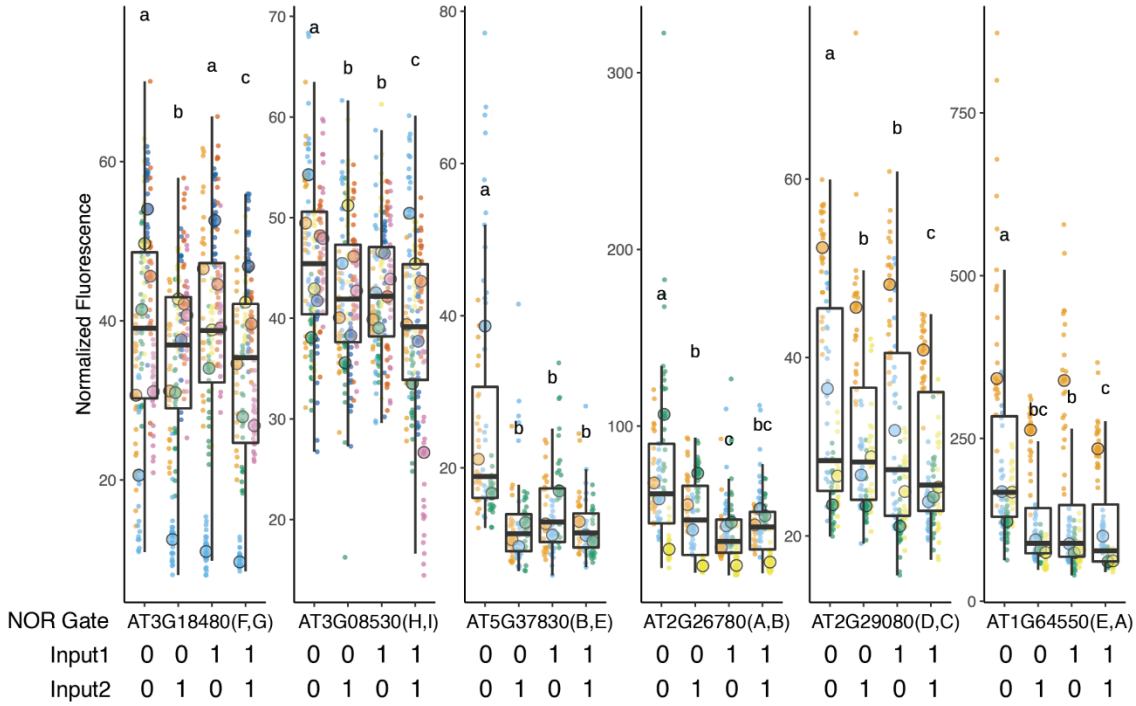
Supplementary Table S7

SRR	Sample Description	AT1G54080	AT5G37830	AT1G64550	AT2G29080	AT1G13320 (PP2AA3)
SRR3581336	Seedling Hypocotyl (S.H)	2791	1880	1294	1711	924
SRR3581345	Seedling Cotyledons (S.C)	2906	2139	1782	3002	486
SRR3581346	Seedling Meristem (S.M)	2703	1103	1849	2449	1.15
SRR3581347	Seedling Root (S.R)	2434	1982	2097	2827	628
SRR3581740	Seedling Hypocotyl (S.H)	2844	2202	1354	1958	552
SRR3581831	Seedling Meristem (S.M)	1186	1139	1103	2310	776
SRR3581833	Seedling Cotyledons (S.C)	2890	1853	1829	2682	246
SRR3581834	Seedling Root (S.R)	2158	2012	2566	3070	610
Mean		2489	1788	1734	2501	527

Values correspond to normalized read count.

Supplementary Table S8. Guide Sequences

gRNA	Guide Sequence + PAM	Source
A	GCAAAGGTGATTAACTGCAAAGG	(Bao et al., 2017)
B	AAAGGGGAAAAGAGTATTGGTGG	(Dahlman et al., 2015)
C	GGCAAGGCTGGCCAACCCATGGG	(Dahlman et al., 2015)
D	ACCCTGGCGGAGCTGATGGGTGG	(Dahlman et al., 2015)
E	TCTCAAGCTAGACTCTAGTGAGG	(Dahlman et al., 2015)
F	CATTGCCATACACCTTGAGGTGG	(Gander et al., 2017)
G	GTGGTAACTTGCTCCATGTCTGG	(Gander et al., 2017)
H	CTTTACGTATAGGTTTAGAGTGG	(Gander et al., 2017)
I	GAAGTCAGTTGACAGAGTCGTGG	(Gander et al., 2017)



Supplementary Figure S9. Normalized Repression Data. Each biological replicate represented by a beeswarm plot and their median is marked by the large circle. The boxplot represents all the replicates together. The y-axis is normalized fluorescence by having the mPromoter:NLS_YFP signal divided by pUBQ10:NLS_mTURQ signal within the same construct. Each input for a given condition can be either ON (1) or OFF (0), and each NOR gate can accept four possible combinations of the two inputs.

Supplementary Table S10. List of primers used in the experiment.

Supplementary Table S11. List of plasmids used in the experiments, plasmid names correspond to Genbank files in Supplementary Data S11.

Supplementary Data S12. All the scripts used in the experiment.

Supplementary Data S13. All the plasmid maps of used in the experiment in Genbank format.

Supplementary Table S10

	Promoter+UTR		UTR+Terminator		
Gene	F	R	F	R	
AT5G06140	TTGAAGACAAGGAGatatcgtgaatcccaattatttatg	TTGAAGACAACATTgagacaacaacgagaatcaaccccg	TTGAAGACAAGCTTatccaaaactaaacttacctg	TTGAAGACAAAGCGgtatcatgttagtgaattggg	
AT1G54080	TTGAAGACAAGGAGttaatatctaaatttactcaaatc	TTGAAGACAACATTttttttgtgtttctttctcactgcg	TTGAAGACAAGCTTtaaaccctcttcactggctctgag	TTGAAGACAAAGCGctgcatggcaaaggacactcgtggac	
AT1G18700	TTGAAGACAAGGAGttttatttttgtttatttcacaatatgtgttcc	TTGAAGACAACATTcggggatttgggaaactcc	TTGAAGACAAGCTTtatattaagccttaggattcttctg	TTGAAGACAAAGCGctaatgggtctgaatctttttccttttatac	
AT4G34270	TTGAAGACAAGGAGttcacaatttttgtgttttttaaaaaacc	TTGAAGACAACATTctgaatcgccggctaatattggc	TTGAAGACAAGCTTgctgtttataaagttgcaaattatatc	TTGAAGACAAAGCGaaagtgtgataattggccgcaggc	
AT5G42470	TTGAAGACAAGGAGtctcatatttcttcatcacgcacatac	TTGAAGACAACATTgggaaggatctctgttgggag	TTGAAGACAAGCTTtgtatctatttcgaaagttagtattgtcc	TTGAAGACAAAGCGataatctattccacctaagaaccg	
AT1G64990	TTGAAGACAAGGAGaaatgtaggattgaaaagc	TTGAAGACAACATTcgtcgccaaattccagatctctc	TTGAAGACAAGCTTgtttcttattcaataatggtgcc	TTGAAGACAAAGCGgtttccacatcgacttgtaatatgg	
AT1G54390	TTGAAGACAAGGAGggtgattgtgatgatgacccatctaattgc	TTGAAGACAACATTctttgatctctgatgaactggatttg	TTGAAGACAAGCTTgctctcctctgcaatttccatgcattc	TTGAAGACAAAGCGgtttactagtcccaaacacagcattgcc	
AT2G26780	TTGAAGACAAGGAGaatttagttttgtttaaaag	TTGAAGACAACATTagctaatctcagaaatttgttg	TTGAAGACAAGCTTgcaagcacagagaagattttag	TTGAAGACAAAGCGgacctaagtatcacaccataattaatc	
AT4G16845	TTGAAGACAAGGAGaatcaaaatacatattcagtaatttattttaaac	TTGAAGACAACATTtcttgttcaaatacaatcaac	TTGAAGACAAGCTTttaaatagaaacactccggtttag	TTGAAGACAAAGCGttaatttcttgatttgttttcaatctatttgc	
AT2G29080	TTGAAGACAAGGAGttccaactagctgttattgctatcatg	TTGAAGACAACATTacaaaatcacaaagaacaaaactaagttcg	TTGAAGACAAGCTTgagaaacgtcactgttacaactg	TTGAAGACAAAGCGgcttaagatgtttagacacatg	
AT2G43070	TTGAAGACAAGGAGtaaattaaggtactaattacaaaggg	TTGAAGACAACATTgacgcaatcgactgagcaaatcgg	TTGAAGACAAGCTTtaacatgattcaacaatattctaatgatg	TTGAAGACAAAGCGatgtttttgagaaaatatatgataaatctaataatt	
AT1G64550	TTGAAGACAAGGAGttggcgaaacaattaagaggtc	TTGAAGACAACATTagccaacacagaagcgacgagattc	TTGAAGACAAGCTTgctgcgccgcatttcttgtattcc	TTGAAGACAAAGCGgcctgaagagtcttgtagttc	
AT3G18480	TTGAAGACAAGGAGgctttaagaatggagggtggg	TTGAAGACAACATTtgtactcgtattcggatcccgatc	TTGAAGACAAGCTTgctctcaccctagggacttattcg	TTGAAGACAAAGCGgtgttgttgcctaacagatatc	
AT5G60160	TTGAAGACAAGGAGgccttttttaaccctcttaata	TTGAAGACAACATTggctggatctcttaacttgag	TTGAAGACAAGCTTgagattcttaaatcttactgctg	TTGAAGACAAAGCGaccaaaattgagatttgttaagag	
AT1G71860	TTGAAGACAAGGAGgtgtgtcactcgagactcggaag	TTGAAGACAACATTgaccgattgacggcgggatc	TTGAAGACAAGCTTaggtacgcaatttcaacaatc	TTGAAGACAAAGCGcaacgatgactcgggacattttac	
AT3G08530	TTGAAGACAAGGAGagtgatactctatgacataatattcc	TTGAAGACAACATTgtttgacctcagatcacagcaag	TTGAAGACAAGCTTagcagaccagcttatgctttaatatcc	TTGAAGACAAAGCGctatccagaacgttttcttag	
AT2G33730	TTGAAGACAAGGAGaaagacagaattagtgtagttgg	TTGAAGACAACATTcggtggttagctggagattattcgc	TTGAAGACAAGCTTgtcgataggctctctggatttg	TTGAAGACAAAGCGaagtaaaagtagttgagttatttgttacaaaccc	
AT5G10630	TTGAAGACAAGGAGaaagaaaaagaaaaaaagaagtgaagtttgttttg	TTGAAGACAACATTcgtttgtttctgcgcacaag	TTGAAGACAAGCTTacatttctctgttgttgtacgag	TTGAAGACAAAGCGactgagttatattatacgggaatactaac	
AT1G71900	TTGAAGACAAGGAGcaatttttttactgttttttttgcttcggtg	TTGAAGACAACATTtctcaagccaattcgaaatctttttcc	TTGAAGACAAGCTTaagctttttcgctctgttttctgg	TTGAAGACAAAGCGtttgggatcttttttctttcgttac	
AT3G07160	TTGAAGACAAGGAGacgttgcgagagagtatttcgggaaattaac	TTGAAGACAACATTttttctacaaacaaaaaaacatcaaagtactcagcagcttaaac	TTGAAGACAAGCTTactttttggaagctgttatgtttatgttttttc	TTGAAGACAAAGCGtagtaatatgttgctctctgcccc	***
AT4G13780	TTGAAGACAAGGAGaatcatgagatttttttttaaaaaagttttttttaaacttttgg	TTGAAGACAACATTggcgaaaacaacaagaccac	TTGAAGACAAGCTTaagcttacttgagccataagcaaac	TTGAAGACAAAGCGagtaagagtttttcaaaaatcaacttgctgg	
AT5G37830	TTGAAGACAAGGAGgaaagttcgagtcaaatttttactaccgc	TTGAAGACAACATTctcactctctctgcataacccaac	TTGAAGACAAGCTTacaattcatctcctcctcataagtctc	TTGAAGACAAAGCGtgagtttattgaaggtcatatttttttgataaatataatgg	

The list contains primers used to clone the "Promoter+UTR" and "UTR+Terminator" regions. The primers contains overhangs with BbsI cutsites for insertion into a MoClo lvl0 vector. The overhangs are capitalized.

*** The terminator region accidentally included portion of the next gene. The portion was removed with Q5_mutagenesis using CGCTTGAGACCACGAAGT and TGGAATTAGGTCCCCAATTTTG

Supplementary Table S10

Gene	Short Fragment		Long Fragment	
	F	R	F	R
AT1G64550(E,A\	TCTCAAGCTAGACTCTAGTGAGGagccatgttaagagggaacaacgttaacc	CCTTTGCAGTTAATCACCTTTGCtaatttactaaaccaaaccaaaatattgatttaatc	GCAAAGGTGATTAACTGCAAAGGtgcctgagggagaaggttccaaatagc	CCTCACTAGAGTCTAGCTTGAGAtcctcaattttgttttttgttaacacatgtttgagag
AT2G26780(A,B	GCAAAGGTGATTAACTGCAAAGGacacaagtggtgggttattttctctgac	CCACCAATACTCTTTTCCCCTTTcaacttcttgttttcagcaatttttattgtttggcc	AAAGGGGAAAAGAGTATTGGTGGagaaaaatggcaaatgcgggctagg	CCTTTGCAGTTAATCACCTTTGCagtagacttcaatgatatttatggcaatgattgag
AT2G29080(D,C	ACCCTGGCGGAGCTGATGGGTGGatgctacggctacctcttcacttttattttc	CCCATGGGTTGGCCAGCCTTGCCaaccagtgtttggttcataggcctaattttac	GGCAAGGCTGGCCAACCCATGGGtggattcggacattatttagaacaaaaccgg	CCACCCATCAGCTCCGCCCAGGGTatgaaccaatattaactaaaaaggacactaaccatc
AT3G08530(H,I)	CTTTACGTATAGGTTTAGAGTGGtgacacttcttccattcttaattattaag	CCACGACTCTGTCAACTGACTTCcatccgaatttctaatgagacgac	GAAGTCAGTTGACAGAGTCGTGGatgggccggacctggttataaatg	CCACTCTAAACCTATACGTAAAGaaaacttaataattaaagatttgatacaag
AT3G18480(F,G	CATTGCCATACACCTTGAGGTGGatactgggcctatgaatggcgcc	CCAGACATGGAGCAAGTTACCACaacgtaaggtccaaaagatatagaataaccaatgg	GTGGTAACTTGCTCCATGTCTGGctataacaaaaaaataaataagaattacgcttctac	CCACCTCAAGGTGTATGGCAATGtttattgcacatctaagttgaaaacatttgtc
AT5G37830(B,E\	AAAGGGGAAAAGAGTATTGGTGGttataggagatttcatcatataatctaattatattaa	CCTCACTAGAGTCTAGCTTGAGATatataaaatttattgtttatttactttttaattaata	TCTCAAGCTAGACTCTAGTGAGGataattcgttattcaaaatcgaatatcccgatccg	CCACCAATACTCTTTTCCCCTTTaaaacagatgtgtatatccatattacgaaactttac

gRNA target-sites are introduced using primers with the target-sites as overhangs. Overhangs are capitalized.
MoClo level0 plasmids containing the native version of the promtoers were used as templates.
The PCR generates a **short fragment** containing the sequence between the two target-sties, and a **long fragment** containing the rest of the backbone. The two pieces are assembled using Gibson assembly.

Supplementary Table S11

Gene	Promoter+UTR	UTR+Terminator
AT5G06140	pich41295pro5u-pat5g06140.gb	pich412763uter-tat5g06140.gb
AT1G54080	pich41295pro5u-pat1g54080.gb	pich412763uter-tat1g54080.gb
AT1G18700	pich41295pro5u-pat1g18700.gb	pich412763uter-tat1g18700.gb
AT4G34270	pich41295pro5u-pat4g34270.gb	pich412763uter-tat4g34270.gb
AT5G42470	pich41295pro5u-pat5g42470.gb	pich412763uter-tat5g42470.gb
AT1G64990	pich41295pro5u-pat1g64990.gb	pich412763uter-tat1g64990.gb
AT1G54390	pich41295pro5u-pat1g54390.gb	pich412763uter-tat1g54390.gb
AT2G26780	pich41295pro5u-pat2g26780.gb	pich412763uter-tat2g26780.gb
AT4G16845	pich41295pro5u-pat4g16845.gb	pich412763uter-tat4g16845.gb
AT2G29080	pich41295pro5u-pat2g29080.gb	pich412763uter-tat2g29080.gb
AT2G43070	pich41295pro5u-pat2g43070.gb	pich412763uter-tat2g43070.gb
AT1G64550	pich41295pro5u-pat1g64550.gb	pich412763uter-tat1g64550.gb
AT3G18480	pich41295pro5u-pat3g18480.gb	pich412763uter-tat3g18480.gb
AT5G60160	pich41295pro5u-pat5g60160.gb	pich412763uter-tat5g60160.gb
AT1G71860	pich41295pro5u-pat1g71860.gb	pich412763uter-tat1g71860.gb
AT3G08530	pich41295pro5u-pat3g08530.gb	pich412763uter-tat3g08530.gb
AT2G33730	pich41295pro5u-pat2g33730.gb	pich412763uter-tat2g33730.gb
AT5G10630	pich41295pro5u-pat5g10630.gb	pich412763uter-tat5g10630.gb
AT1G71900	pich41295pro5u-pat1g71900.gb	pich412763uter-tat1g71900.gb
AT3G07160	pich41295pro5u-pat3g07160.gb	pich412763uter-tat3g07160.gb
AT4G13780	pich41295pro5u-pat4g13780.gb	pich412763uter-tat4g13780.gb
AT5G37830	pich41295pro5u-pat5g37830.gb	pich412763uter-tat5g37830.gb
AT1G64550(E,A)	pich41295pro5u-pat1g64550ea.gb	
AT2G26780(A,B)	pich41295pro5u-pat2g26780ab.gb	
AT2G29080(D,C)	pich41295pro5u-pat2g29080dc.gb	
AT3G08530(H,I)	pich41295pro5u-pat3g08530hi.gb	
AT3G18480(F,G)	pich41295pro5u-pat3g18480fg.gb	
AT5G37830(B,E)	pich41295pro5u-pat5g37830be.gb	

Plasmid maps are also annoated with predicted trascription factor motifs from PlantRegMap (Tian et al. 2019)

Supplementary Table S11

Gene	YFP
AT5G06140	pich47732lvl1-pat5g06140yfp-tat5g06140.gb
AT1G54080	pich47732lvl1-pat1g54080yfp-tat1g54080.gb
AT1G18700	pich47732lvl1-pat1g18700yfp-tat1g18700.gb
AT4G34270	pich47732lvl1-pat4g34270yfp-tat4g34270.gb
AT5G42470	pich47732lvl1-pat5g42470yfp-tat5g42470.gb
AT1G64990	pich47732lvl1-pat1g64990yfp-tat1g64990.gb
AT1G54390	pich47732lvl1-pat1g54390yfp-tat1g54390.gb
AT2G26780	pich47732lvl1-pat2g26780yfp-tat2g26780.gb
AT4G16845	pich47732lvl1-pat4g16845yfp-tat4g16845.gb
AT2G29080	pich47732lvl1-pat2g29080yfp-tat2g29080.gb
AT2G43070	pich47732lvl1-pat2g43070yfp-tat2g43070.gb
AT1G64550	pich47732lvl1-pat1g64550yfp-tat1g64550.gb
AT3G18480	pich47732lvl1-pat3g18480yfp-tat3g18480.gb
AT5G60160	pich47732lvl1-pat5g60160yfp-tat5g60160.gb
AT1G71860	pich47732lvl1-pat1g71860yfp-tat1g71860.gb
AT3G08530	pich47732lvl1-pat3g08530yfp-tat3g08530.gb
AT2G33730	pich47732lvl1-pat2g33730yfp-tat2g33730.gb
AT5G10630	pich47732lvl1-pat5g10630yfp-tat5g10630.gb
AT1G71900	pich47732lvl1-pat1g71900yfp-tat1g71900.gb
AT3G07160	pich47732lvl1-pat3g07160yfp-tat3g07160.gb
AT4G13780	pich47732lvl1-pat4g13780yfp-tat4g13780.gb
AT5G37830	pich47732lvl1-pat5g37830yfp-tat5g37830.gb
AT1G64550(E,A)	pich47732lvl1-pat1g64550eayfp-tat1g64550.gb
AT2G26780(A,B)	pich47732lvl1-pat2g26780abyfp-tat2g26780.gb
AT2G29080(D,C)	pich47732lvl1-pat2g29080dcyfp-tat2g29080.gb
AT3G08530(H,I)	pich47732lvl1-pat3g08530hiyfp-tat3g08530.gb
AT3G18480(F,G)	pich47732lvl1-pat3g18480fgyfp-tat3g18480.gb
AT5G37830(B,E)	pich47732lvl1-pat5g37830beyfp-tat5g37830.gb

Supplementary Table S11

Gene	YFP;mTURQ	RUBY
AT5G06140	pagm4673-pat5g06140nls-yfp-ubq10nls-mturq-basta.gb	pich86966-06140ruby.gb
AT1G54080	pagm4673-pat1g54080nls-yfp-ubq10nls-mturq-basta.gb	pich86966-54080ruby.gb
AT1G18700	pagm4673-pat1g18700nls-yfp-ubq10nls-mturq-basta.gb	pich86966-18700ruby.gb
AT4G34270	pagm4673-pat4g34270nls-yfp-ubq10nls-mturq-basta.gb	
AT5G42470	pagm4673-pat5g42470nls-yfp-ubq10nls-mturq-basta.gb	
AT1G64990	pagm4673-pat1g64990nls-yfp-ubq10nls-mturq-basta.gb	pich86966-64990ruby.gb
AT1G54390	pagm4673-pat1g54390nls-yfp-ubq10nls-mturq-basta.gb	pich86966-54390ruby.gb
AT2G26780	pagm4673-pat2g26780nls-yfp-ubq10nls-mturq-basta.gb	pich86966-26780ruby.gb
AT4G16845	pagm4673-pat4g16845nls-yfp-ubq10nls-mturq-basta.gb	
AT2G29080	pagm4673-pat2g29080nls-yfp-ubq10nls-mturq-basta.gb	pich86966-29080ruby.gb
AT2G43070	pagm4673-pat2g43070nls-yfp-ubq10nls-mturq-basta.gb	
AT1G64550	pagm4673-pat1g64550nls-yfp-ubq10nls-mturq-basta.gb	pich86966-64550ruby.gb
AT3G18480	pagm4673-pat3g18480nls-yfp-ubq10nls-mturq-basta.gb	
AT5G60160	pagm4673-pat5g60160nls-yfp-ubq10nls-mturq-basta.gb	
AT1G71860	pagm4673-pat1g71860nls-yfp-ubq10nls-mturq-basta.gb	pich86966-71860ruby.gb
AT3G08530	pagm4673-pat3g08530nls-yfp-ubq10nls-mturq-basta.gb	pich86966-08530ruby.gb
AT2G33730	pagm4673-pat2g33730nls-yfp-ubq10nls-mturq-basta.gb	pich86966-33730ruby.gb
AT5G10630	pagm4673-pat5g10630nls-yfp-ubq10nls-mturq-basta.gb	
AT1G71900	pagm4673-pat1g71900nls-yfp-ubq10nls-mturq-basta.gb	
AT3G07160	pagm4673-pat3g07160nls-yfp-ubq10nls-mturq-basta.gb	
AT4G13780	pagm4673-pat4g13780nls-yfp-ubq10nls-mturq-basta.gb	
AT5G37830	pagm4673-pat5g37830nls-yfp-ubq10nls-mturq-basta.gb	pich86966-37830ruby.gb
AT1G64550(E,A)	pagm4723-pat1g64550eanls_yfp-ubq10mturq-basta.gb	
AT2G26780(A,B)	pagm4723-pat2g26780abnls_yfp-ubq10mturq-basta.gb	
AT2G29080(D,C)	pagm4723-pat2g29080dcnls_yfp-ubq10mturq-basta.gb	
AT3G08530(H,I)	pagm4723-pat3g08530hinls-yfp-ubq10nls-mturq-hyg.gb	
AT3G18480(F,G)	pagm4723-pat3g18480fgnls-yfp-ubq10nls-mturq-hyg.gb	
AT5G37830(B,E)	pagm4723-pat5g37830benls-yfp-ubq10nls-mturq-hyg.gb	

Supplementary Table S11

Construct	GeneBank
dCas9_TPL Repressor	p2301y-tdtomato-dcas9_degrondead_tpln188.gb
SelfCleaving gRNA_A	pich86988-35sselfcleaving_a.gb
SelfCleaving gRNA_B	pich86988-35sselfcleaving_b.gb
SelfCleaving gRNA_C	pich86988-35sselfcleaving_c.gb
SelfCleaving gRNA_D	pich86988-35sselfcleaving_d.gb
SelfCleaving gRNA_E	pich86988-35sselfcleaving_e.gb
SelfCleaving gRNA_F	pich86988-35sselfcleaving_f.gb
SelfCleaving gRNA_G	pich86988-35sselfcleaving_g.gb
SelfCleaving gRNA_H	pich86988-35sselfcleaving_h.gb
SelfCleaving gRNA_I	pich86988-35sselfcleaving_i.gb

References

Bao, Z., Jain, S., Jaroenpuntaruk, V., & Zhao, H. (2017). Orthogonal Genetic Regulation in Human Cells Using Chemically Induced CRISPR/Cas9 Activators. *ACS Synthetic Biology*, *6*(4), 686–693. https://doi.org/10.1021/acssynbio.6b00313

Dahlman, J. E., Abudayyeh, O. O., Joung, J., Gootenberg, J. S., Zhang, F., & Konermann, S. (2015). Orthogonal gene knock out and activation with a catalytically active Cas9 nuclease. *Nature Biotechnology*, *33*(11), 1159–1161. https://doi.org/10.1038/nbt.3390

Gander, M. W., Vrana, J. D., Voje, W. E., Carothers, J. M., & Klavins, E. (2017). Digital logic circuits in yeast with CRISPR-dCas9 NOR gates. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15459

Tokizawa, M., Kusunoki, K., Koyama, H., Kurotani, A., Sakurai, T., Suzuki, Y., Sakamoto, T., Kurata, T., & Yamamoto, Y. Y. (2017). Identification of Arabidopsis genic and non-genic promoters by paired-end sequencing of TSS tags. *The Plant Journal: For Cell and Molecular Biology*, *90*(3), 587–605. https://doi.org/10.1111/tpj.13511

Chapter2: 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.

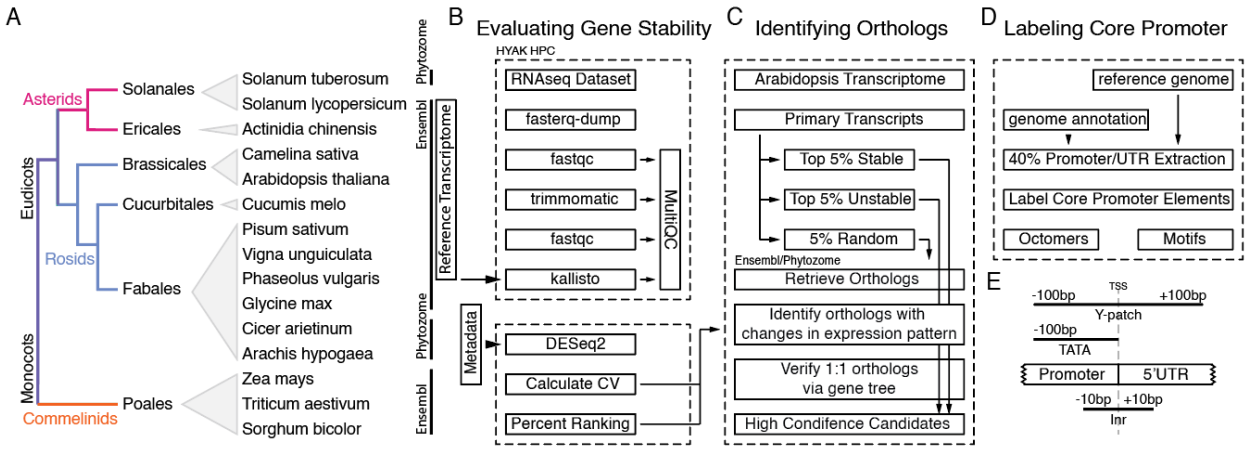


Figure 1. An outline of the bioinformatics pipelines. A) The fifteen angiosperms included in this study and their phylogenetic relationship. B-D) The three major data processing steps performed in the study. Detailed parameters are included in the Methods section. Reference genomes, transcriptomes and 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.

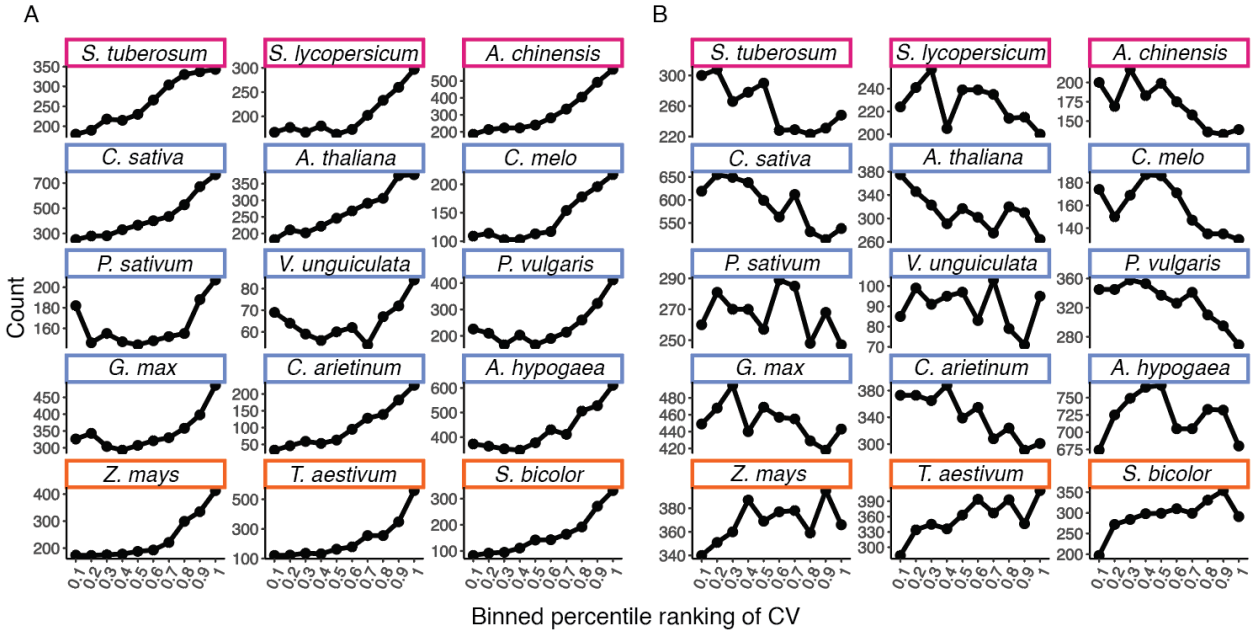


Figure 2. Distribution of relative specificity or uniformity of TATA-box-containing and Coreless promoters. Higher Coefficient of Variation (CV) rankings indicate more specificity, while lower CV rankings indicate more uniformity. A random subsampling of forty percent of promoters from each species are 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.

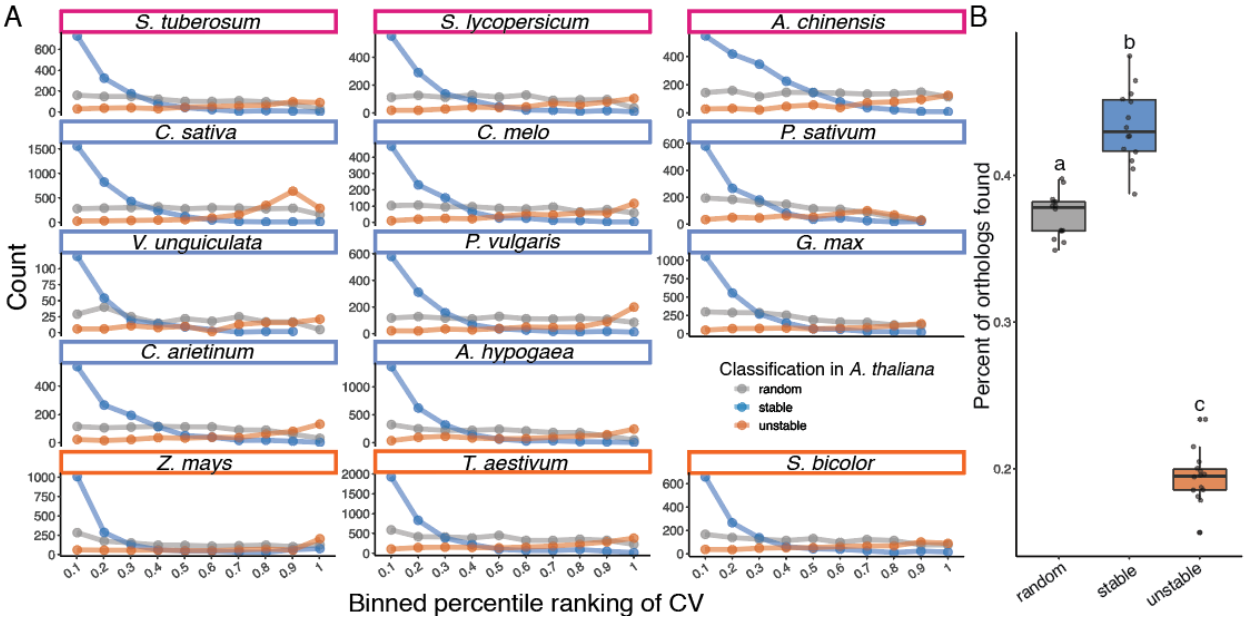


Figure 3. Genes that show uniform expression in *A. thaliana* tend to behave similarly in other species. A) Distribution of CVs for orthologs of stable (blue), unstable (orange) or random (grey) *A. thaliana* genes. The color of boxes around species names corresponds to Figure1A. B) Percent of orthologs 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.

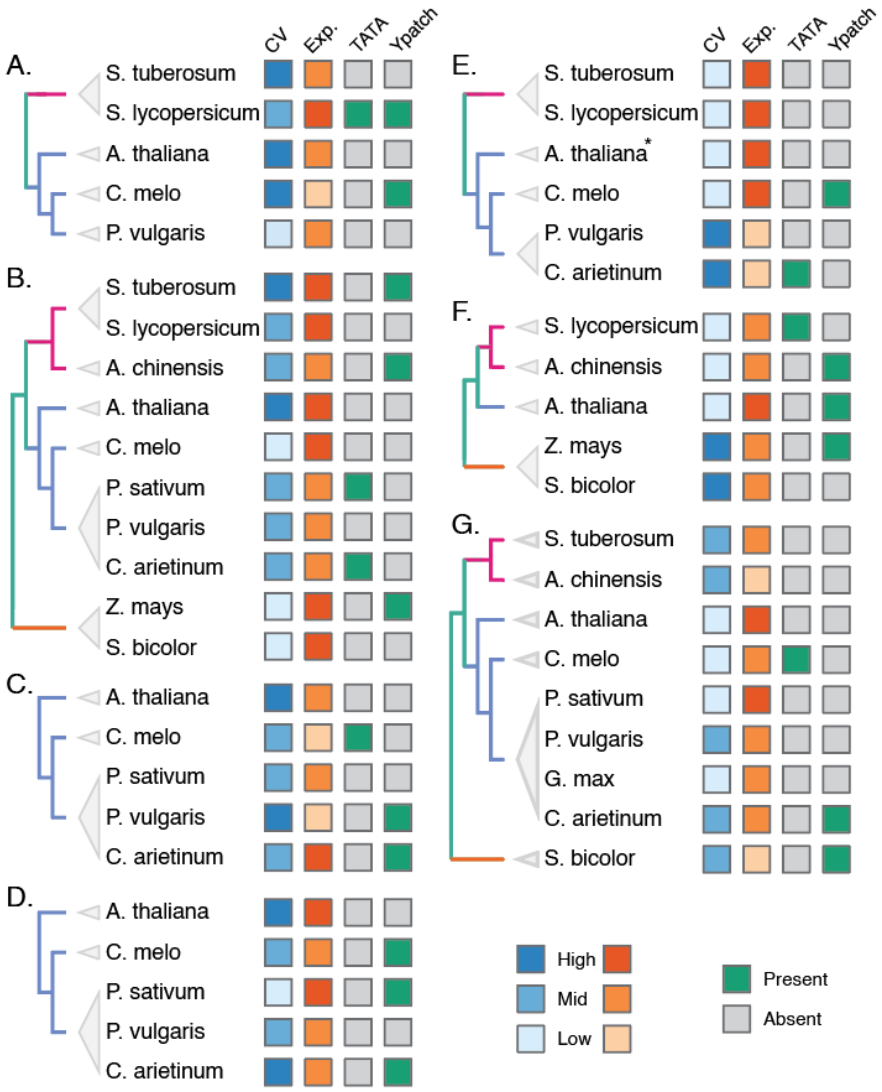


Figure4. Individual gene trees where expression stability changes can be observed. A-D) The gene is unstably expressed in *A. thaliana* but stably expressed in another species. E-G) The gene is stably expressed in *A. thaliana* but unstably expressed in another species. CV and expression strength (Exp.) 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 50^{th} 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

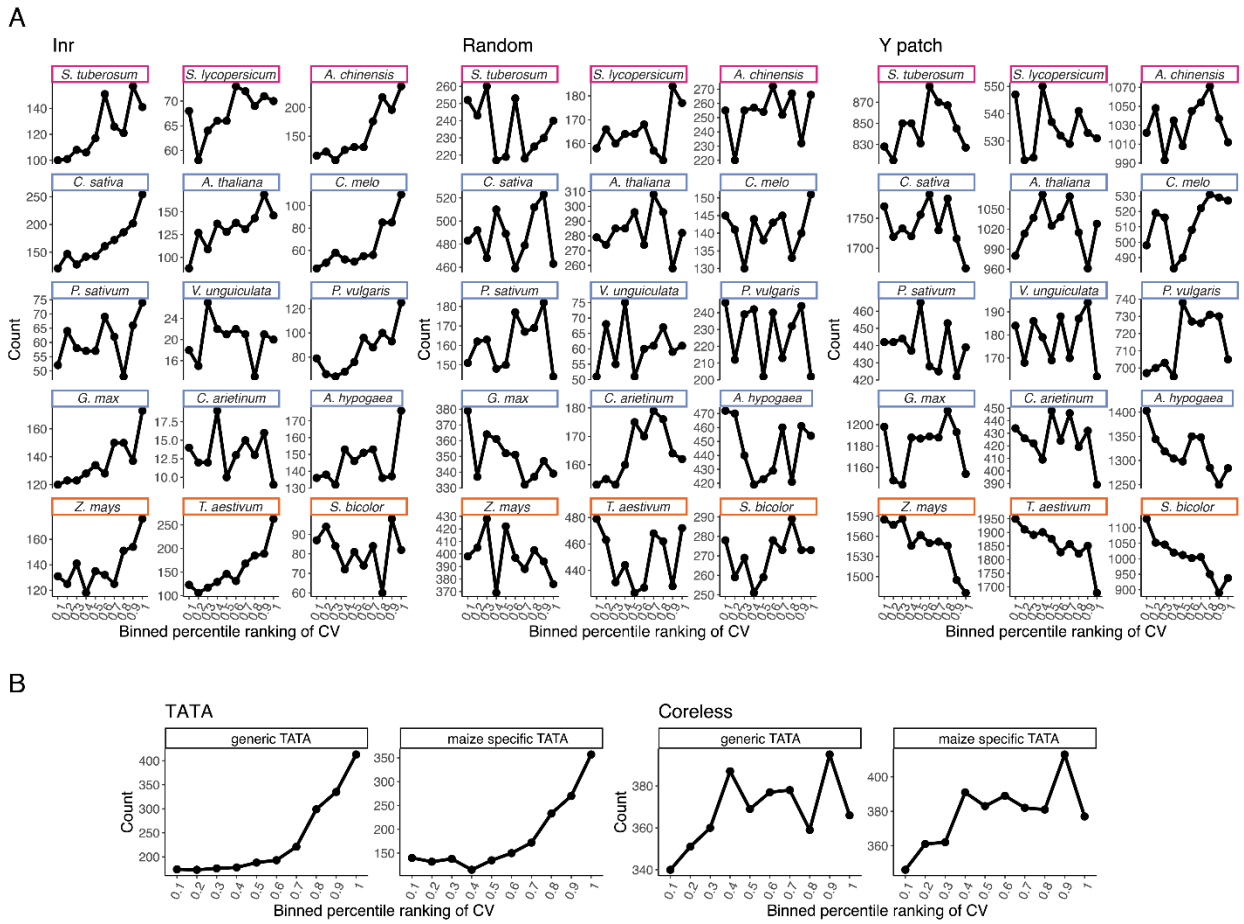
Chapter2: Supplementary Materials

Supplemental Table S1.A summary of the species included in this study along with details regarding the RNA-seq dataset including BioProject ID, library layout, number of unique samples used in this study, source of reference transcriptome, and references to the original study whenever available. File also available at https://doi.org/10.5061/dryad.9w0vt4bmk

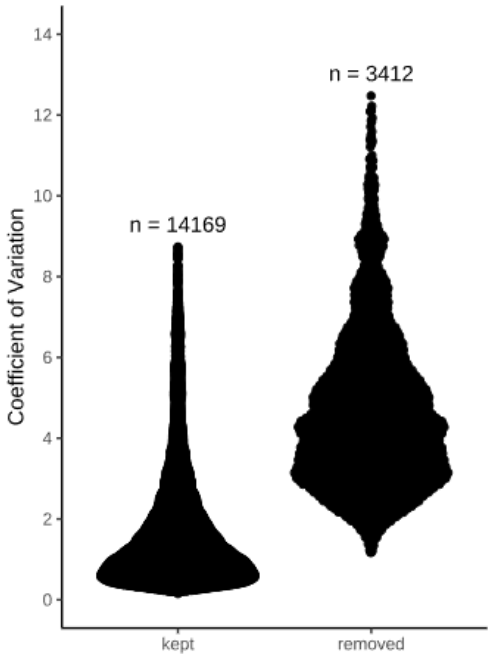
Supplementary Table S1

Species	Common Name	Layout	BioProject	Number of Unique Samples	Reference Transcriptome	DOI	Reference
Arabidopsis thaliana	mouse ear cress	Single	PRJNA268115, PRJNA314076, PRJNA194429	80	EnsemblPlants	10.1111/tpj.13312 ; 10.1104/pp.112.211441	Klepikova et al., 2016 ; Loraine et al., 2013
Camelina sativa	false flax	Paired	PRJNA231618	12	EnsemblPlants	10.1038/ncomms4706	Kagale et al., 2014
Cucumis melo	melon	Paired	PRJDB6414	30	EnsemblPlants	10.1093/pcp/pcx193	Yano et al., 2018
Arachis hypogaea	peanut	Paired	PRJNA291488	22	Phytozome		Peanut Genome Consortium
Cicer arietinum	chickpea	Paired	PRJNA413872	27	Phytozome	10.1111/pce.13210	Kudapa et al., 2018
Glycine max	soy bean	Single	PRJNA79597	10	EnsemblPlants	10.1111/j.1365-313X.2010.04222.x	Libault et al., 2010
Phaseolus Vulgaris	common bean	Paired	PRJNA221782	34	EnsemblPlants	10.1186/s13059-016-0883-6	Vlasova et al., 2016
Pisum sativum	green pea	Paired	PRJNA277076	11	EnsemblPlants	10.1186/s12864-015-1815-7	Sudheesh et al., 2015
Vigna unguiculata	cowpea	Paired	PRJNA389300	11	EnsemblPlants	10.1111/tpj.13279	Yao et al., 2016
Sorghum bicolor	sorghum	Paired	SRA558272, SRA558514, SRA558539	47	EnsemblPlants	10.1111/tpj.13781	McCormick et al., 2017
Zea mays	corn	Single,Paired	PRJNA171684, SRP010680	116	EnsemblPlants	10.3835/plantgenome2015.04.0025	Stelpflug et al., 2016
Solanum lycopersicum	tomato	Single	PRJNA507622	30	EnsemblPlants	10.3390/genes10010050	Penin et al., 2019
Actinidia chinensis	golden kiwi	Paired	PRJNA691387	13	EnsemblPlants	10.1186/s12870-021-02894-x	Brian et al., 2021
Triticum aestivum	bread wheat	Paired	PRJEB25639	75	EnsemblPlants	10.1126/science.aar6089	Ramírez-González et al., 2018
Solanum tuberosum	potato	Single	PRJEB2430	14	Phytozome	10.1038/nature10158	Potato Genome Sequence Consortium et al., 2011

Note: Number of samples indicates the samples included in this study. Sometimes not all the samples provided from the original study were included.



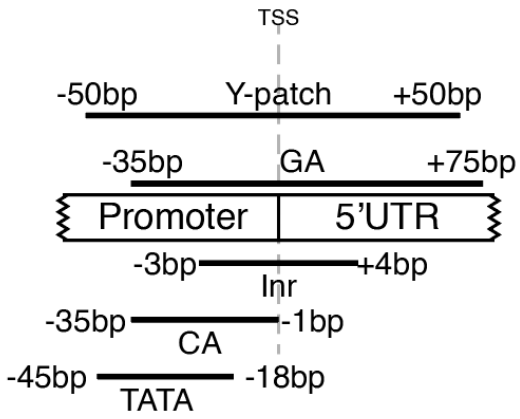
Supplemental Figure S2. A) Distribution of relative specificity or uniformity of Inr and Y patch containing promoters for all fifteen species. B) Distribution of relative specificty or uniformity of TATA and Coreless containing promoters in *Zea mays* using a generic TATA motif from Jores et al. and a maize specific TATA motif from Mejía-Guerra et al.. 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 further random subsampling of twenty percent of the promoters were used in the Random set as a control. Colors correspond to phylogeny shown in Figure1A.



Supplemental Figure S3. Distribution of coefficient of variation for transcripts in *Arabidopsis* that contained only a single read (removed), and the rest of the transcripts (kept). “n” above the beeswarm plot represents the number of transcripts in each group.

	Random		Stable		Unstable	
	Target Orthologs	*Arabidopsis* Transcripts	Target Orthologs	*Arabidopsis* Transcripts	Target Orthologs	*Arabidopsis* Transcripts
Highest expressing target transcript only	22,339	1,197	25,427	1,323	14,350	934
Single ortholog in target transcriptome	4,066	984	6,505	1,224	2,149	682
Changed expression pattern in at least 2 target species	NA	NA	189	27	345	58
High confidence candidates	NA	NA	17	3	21	4

Supplemental Table S4. Number of *Arabidopsis* transcripts and target orthologs after each filtering step. The random set of genes were not used in analysis involving changes in expression patterns from *Arabidopsis*, and the counts were left as NA.



Supplemental Figure S5. Regions searched for the various core promoter octamers. The regions were defined in Yamamoto et al. 2009.

Supplemental Table S6. Summary of the core promoter scanning results. The transcripts are grouped by orthologous gene groups. “Change” denotes whether there is a change in expression pattern compared to the *Arabidopsis* transcript (*Arabidopsis* is always unchanged. i.e. “Unstable_to_Unstable” or “Stable_to_Stable”), and those that changed expression compared to *Arabidopsis* are highlighted orange. CV_PR and Gmean_PR are the percent ranking of CV and geometric mean within the species, respectively. At_transcript_id is the column used to define each orthologous gene group and it is the *Arabidopsis* transcript that the gene group is related to. TATA, Inr, BREu, BREd, Ypatch, and TCT are core promoter elements screened using “Motif Scan” method, and the columns are highlighted in yellow. The numbers are relative motif scores and scores above 0.85 are considered positive. TATA_result, Ypatch_result, Ypatch_resultUTR, Inr_result, GA_result, GA_resultUTR, CA_result are core promoter elements screened using “Octamer Scan” method, and the columns are highlighted green. Details regarding the two scanning methods can be found in the methods section. The presence of a core promoter element for either method is highlighted in green. Gene symbol (Symbol) and Description are from TAIR. File also available at https://doi.org/10.5061/dryad.9w0vt4bmk

Supplementary Table S6

Figure4	Transcript_name	Change	Species_name	CV_PR	Gmean_PR	At_transcript_id	TATA	Inr	BREu	BREd	Ypatch	TCT	TATA_result	Ypatch_result	Ypatch_result	Inr_result	GA_result	GA_result	CA_result	Symbol	Description
	Soltu.DM.09G028480.1	Unstable_to_Unstable	Solanum_tuberosum	0.891	0.535	AT1G04700	0.733069	0.318759	0.207685	0	0.731948	0.426132	FALSE	FALSE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	Solyc09g091460.3.1	Unstable_to_Unstable	Solanum_lycopersicum	0.572	0.986	AT1G04700	0.907175	0.71798	0.317637	0.724077	0.855536	0.739587	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
a.	**AT1G04700.1**	Unstable_to_Unstable	Arabidopsis_thaliana	0.955	0.501	AT1G04700	0.784217	0.853692	0.330742	0.638174	0.818008	0.723167	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	RAF16	Raf-like kinase required for stomatal vapor pressure difference response.
	MELO3C006335.2.1	Unstable_to_Unstable	Cucumis_melo	0.729	0.33	AT1G04700	0.732972	0.713677	0.588677	0.881281	0.853364	0.74318	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE		
	ESW16164	Unstable_to_Stable	Phaseolus_vulgaris	0.313	0.401	AT1G04700	0.806025	0.594446	0.218372	0.466652	0.688992	0.579679	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Soltu.DM.01G019780.1	Stable_to_Stable	Solanum_tuberosum	0.017	0.987	AT3G17020	0.792482	0.718102	0.408979	0.140619	0.673814	0.783035	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Solyc01g059930.3.1	Stable_to_Stable	Solanum_lycopersicum	0.3	0.956	AT3G17020	0.796076	0.658994	0.193494	0.501878	0.720552	0.607384	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
e.	**AT3G17020.1**	Stable_to_Stable	Arabidopsis_thaliana	0.046	0.99	AT3G17020							FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		Adenine nucleotide alpha hydrolases-like superfamily protein(source:Araport11)
	MELO3C006711.2.1	Stable_to_Stable	Cucumis_melo	0.302	0.957	AT3G17020	0.76118	0.665083	0.530258	0.232547	0.798679	0.668942	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE		
	ESW19354	Stable_to_Unstable	Phaseolus_vulgaris	0.826	0.183	AT3G17020	0.709689	0.682418	0.456513	0.298604	0.71591	0.482048	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Ca_22768	Stable_to_Unstable	Cicer_arietinum	0.911	0.125	AT3G17020	0.858515	0.731539	0.565576	0.409171	0.825007	0.532829	FALSE	FALSE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	Solyc12g040460.2.1	Stable_to_Stable	Solanum_lycopersicum	0.206	0.633	AT3G18215	0.882866	0.664393	0.119873	0.330215	0.638324	0.740411	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	PSR684703	Stable_to_Stable	Actinidia_chinensis	0.29	0.48	AT3G18215	0.779648	0.829741	0.317661	0.838685	0.937764	0.739693	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
f.	**AT3G18215.1**	Stable_to_Stable	Arabidopsis_thaliana	0.011	0.853	AT3G18215	0.763233	0.573702	0.2741	0.571479	0.677842	0.883773	FALSE	FALSE	TRUE	FALSE	FALSE	TRUE	FALSE		transmembrane protein. putative (Protein of unknown function, DUF599);(source:Araport11)
	Zm00001eb156540_T002	Stable_to_Unstable	Zea_mays	0.666	0.448	AT3G18215	0.740395	0.7548	0.482955	0.353108	0.89358	0.720188	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	EES01278	Stable_to_Unstable	Sorghum_bicolor	0.827	0.339	AT3G18215	0.635276	0.636057	0.406795	0.566447	0.803443	0.618394	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Soltu.DM.11G005290.3	Stable_to_Unstable	Solanum_tuberosum	0.62	0.446	AT4G40045	0.802153	0.592092	0.63241	0.479339	0.758522	0.682593	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	PSR698770	Stable_to_Unstable	Actinidia_chinensis	0.526	0.273	AT4G40045	0.734997	0.691565	0.428378	0.152031	0.748084	0.694512	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	AT4G40045.1	Stable_to_Stable	Arabidopsis_thaliana	0.036	0.808	AT4G40045	0.776262	0.843503	0.314626	0.652987	0.796959	0.492234	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		transmembrane protein;(source:Araport11)
	MELO3C012366.2.1	Stable_to_Stable	Cucumis_melo	0.259	0.528	AT4G40045	0.859898	0.826534	0.383089	0.174215	0.669688	0.672598	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
g.	Psat6g094320.1	Stable_to_Stable	Pisum_sativum	0.317	0.76	AT4G40045	0.761319	0.635611	0.45792	0.60713	0.707142	0.348788	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	ESW17514	Stable_to_Unstable	Phaseolus_vulgaris	0.584	0.533	AT4G40045	0.810687	0.848293	0.551286	0.416755	0.685382	0.708818	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	KRH71739	Stable_to_Stable	Glycine_max	0.059	0.665	AT4G40045	0.769727	0.639711	0.425687	0.536112	0.754151	0.533674	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Ca_20854	Stable_to_Stable	Cicer_arietinum	0.452	0.432	AT4G40045	0.772932	0.658669	0.360852	0.545893	0.878279	0.381564	FALSE	FALSE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	EES15823	Stable_to_Stable	Sorghum_bicolor	0.46	0.262	AT4G40045	0.621554	0.700402	0.408288	0.688497	0.964977	0.890239	FALSE	TRUE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	Soltu.DM.06G025410.1	Unstable_to_Unstable	Solanum_tuberosum	0.715	0.674	AT5G17400	0.830861	0.590671	0.407276	0.343043	0.871353	0.410706	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Solyc06g069410.3.1	Unstable_to_Unstable	Solanum_lycopersicum	0.609	0.715	AT5G17400	0.782936	0.424999	0.268893	0.654405	0.843108	0.589992	FALSE	FALSE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	PSS08010	Unstable_to_Stable	Actinidia_chinensis	0.473	0.403	AT5G17400	0.798477	0.654975	0.57762	0.376285	0.968857	0.913601	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE		
	ATSG17400.1	Unstable_to_Unstable	Arabidopsis_thaliana	0.952	0.802	AT5G17400	0.759959	0.633459	0.287181	0.372953	0.780725	0.843896	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	ER-ANT1	ENDOPLASMIC RETICULUM-ADENINE NUCLEOTIDE TRANSPORTER 1
b.	MELO3C002052.2.1	Unstable_to_Stable	Cucumis_melo	0.164	0.708	AT5G17400	0.819234	0.653108	0.508095	0.378127	0.766259	0.608081	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Psat4g001520.1	Unstable_to_Stable	Pisum_sativum	0.382	0.6	AT5G17400	0.852534	0.54128	0.326744	0.623503	0.831963	0.480632	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	ESW32531	Unstable_to_Stable	Phaseolus_vulgaris	0.36	0.504	AT5G17400	0.670727	0.868916	0.508465	0.407683	0.795952	0.638702	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Ca_10349	Unstable_to_Stable	Cicer_arietinum	0.338	0.396	AT5G17400	0.901417	0.439979	0.370675	0.41959	0.757049	0.693772	FALSE	FALSE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	Zm00001eb239090_T001	Unstable_to_Stable	Zea_mays	0.079	0.791	AT5G17400	0.797297	0.836277	0.459796	0.321355	0.909569	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	EES04810	Unstable_to_Stable	Sorghum_bicolor	0.022	0.787	AT5G17400	0.815418	0.808509	0.284145	0.677582	0.767782	0.652099	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	ATSG18910.1	Unstable_to_Unstable	Arabidopsis_thaliana	0.968	0.45	AT5G18910	0.733915	0.676044	0.229873	0.305408	0.814643	0.975402	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE		Protein kinase superfamily protein(source:Araport11)
	MELO3C021381.2.1	Unstable_to_Unstable	Cucumis_melo	0.608	0.328	AT5G18910	0.909959	0.77112	0.205291	0.601318	0.812727	0.74954	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
c.	Psat1g186120.1	Unstable_to_Stable	Pisum_sativum	0.454	0.621	AT5G18910	0.738785	0.653229	0.413545	0.837054	0.592568	0.502103	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	ESW19829	Unstable_to_Unstable	Phaseolus_vulgaris	0.92	0.153	AT5G18910	0.768756	0.815776	0.303914	0.80275	0.867472	0.588639	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Ca_00555	Unstable_to_Stable	Cicer_arietinum	0.378	0.789	AT5G18910	0.682383	0.585272	0.29804	0.778935	0.945034	0.640963	FALSE	TRUE	no UTR se	no UTR se	FALSE	FALSE	FALSE		
	ATSG20410.1	Unstable_to_Unstable	Arabidopsis_thaliana	0.962	0.796	AT5G20410	0.799587	0.72273	0.472284	0.562549	0.798556	1	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	MGD2	ARABIDOPSIS THALIANA MONOGALACTOSYLDIACYLGLYCEROL SYNTHASE 2
	MELO3C004474.2.1	Unstable_to_Unstable	Cucumis_melo	0.558	0.604	AT5G20410	0.776512	0.705964	0.237055	0.449288	0.870493	0.537182	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
d.	Psat4g035200.1	Unstable_to_Stable	Pisum_sativum	0.166	0.821	AT5G20410	0.747291	0.646937	0.742658	0.449359	0.933515	0.515881	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	ESW27071	Unstable_to_Unstable	Phaseolus_vulgaris	0.508	0.655	AT5G20410	0.803388	0.754597	0.06448	0.365228	0.703532	0.707894	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE		
	Ca_06724	Unstable_to_Unstable	Cicer_arietinum	0.844	0.665	AT5G20410	0.752133	0.587302	0.328422	0.555674	0.856248	0.547283	FALSE	FALSE	no UTR se	no UTR se	FALSE	FALSE	FALSE		

Supplementary Data S7. MultiQC outputs for the RNA-seq analysis pipeline. The MultiQC output for RNA-seq datasets from each species is stored as html files, and each file provides a visual summary of results from FastQC, Trimmomatic, and Kallisto. File available at https://doi.org/10.5061/dryad.9w0vt4bmk

Chapter3: Learning the rules that govern the strength of a single helix repression domain and its level of interaction with a small molecule inhibitor.

Introduction

As shown in Chapter1, repressor domains are powerful synthetic biology tools that can be used to control gene expression and build logic gates. One of the ways to improve the logic gates' performances of the logic gates is to strengthen the repressors. Having a powerful repressor can reduce transcriptional leakage, leading to a more robust circuit. These qualities become even more important when layering gates, an essential step in any large synthetic circuits (Gander et al. 2017). The current strategy in engineering repression strength usually involves recruiting additional copies of repressors to increase the local concentration of the repressor at chromatin (Brophy et al. 2022; Zhai et al. 2022). Here we attempt to improve repression by using machine learning models to engineer stronger repressors, which could be used either as monomers or multimers to engineer transcription.

TPL is a corepressor that is involved in multiple important biological processes including hormone signaling, embryogenesis, meristem maintenance, among others (Plant, Larrieu, and Causier 2021). Structurally, going from amino- to carboxy-termini, TPL contains a LIS1 homology (LisH) domain, C-terminal to LisH (CTLH) domain, CT11-RanBPM (CRA) domain and two WD40 beta-propeller domains (Martin-Arevalillo et al. 2017; Leydon et al. 2021). Work done in the lab dissecting the domains responsible for TPL's repressor activity has identified two main domains at the N-terminus that are independently capable of repression (Leydon et al. 2021). One of the two domains is the LisH domain, and surprisingly, it was found that just the first helix from the LisH (LisH H1, 18 amino acid) is capable of repression (Leydon, Ramos Báez, and Nemhauser 2022). LisH domains are generally thought of as a protein multimerization domain and are found across eukaryotes, in proteins spanning functions such as regulation of transcription , the cytoskeleton, and protein degradation (Leydon, Ramos Báez, and Nemhauser 2022). Curious whether the LisH domains found in some of the non-repression related proteins still showed repressive function, the lab assayed multiple LisH H1 sequences from a large collection of proteins and identified functional repressors in the majority of them (Leydon, Ramos Báez, and Nemhauser 2022). Leveraging this sequence to repression data available, we aim to use a machine learning approach to model LisH H1 repression to design stronger repressors.

The platform used to assay these repressors is a yeast synthetic system that allows the measuring of a fluorescent reporter output via flow cytometry. The reporter is controlled by a repressible promoter that can be targeted by any repression domain of interest. If the repression domain is of clinical interest, any drug that interrupts the repressor's interaction with the basal transcriptional machinery will result in a relief of repression that can be quantified. We believe this allows the yeast synthetic system to also serve as a drug screening platform. One intriguing candidate that is well-positioned for building out the drug screening pipeline is the human protein Transducin-beta like 1 (TBL1).

TBL1 is also a LisH-containing protein that plays a critical role in the Wnt pathway that is implicated in a variety of cancers (Pray, Youssef, and Alinari 2022; Pai et al. 2017; Soldi et al. 2021). TBL1 serves as an exchange factor in the Wnt pathway that is part of the SMRT/NCoR co-repressor complex in the absence

of Wnt signaling, but switches to preferentially bind β-catenin, a co-activator, upon Wnt signaling and subsequent SUMOylation (Oberoi et al. 2011; Choi et al. 2011). The downstream genes of the Wnt pathway are essential in development and are involved in cell proliferation and cell fate decisions. Therefore, inappropriately active Wnt signaling can lead to multiple types of cancer.

TBL1's central role in the Wnt pathway makes it an attractive drug target, and currently there is only a single drug (Tegavivint) in Phase II clinical trial that targets TBL1 as a cancer therapeutic (Soldi et al. 2021). Crystal structure of the TBL1 homodimer reveals a hydrophobic pocket. Mutations in hydrophobic residues (Val60 and Ile39) within the pocket reduced TBL1 interaction with both SMRT and β-catenin (Soldi et al. 2021; Oberoi et al. 2011). This same pocket is predicted to be the binding site of Tegavivint based on a docking experiment (Soldi et al. 2021) (Figure1). By binding to this region of TBL1, Tegavivint would block interaction with β-catenin thereby reducing activation of Wnt target genes.

Using the yeast platform in the lab, we were able to measure relief of repression for TBL1's LisH H1 upon the addition of Tegavivint (Leydon, Ramos Báez, and Nemhauser 2022). I worked to further optimize the Tegavivint assay and to teste whether the screening platform would reveal differential effects on Tegavivint activity on different cancer-associated TBL1 mutations. If successful, this approach could be used to develop patient-specific chemotherapy recommendations.

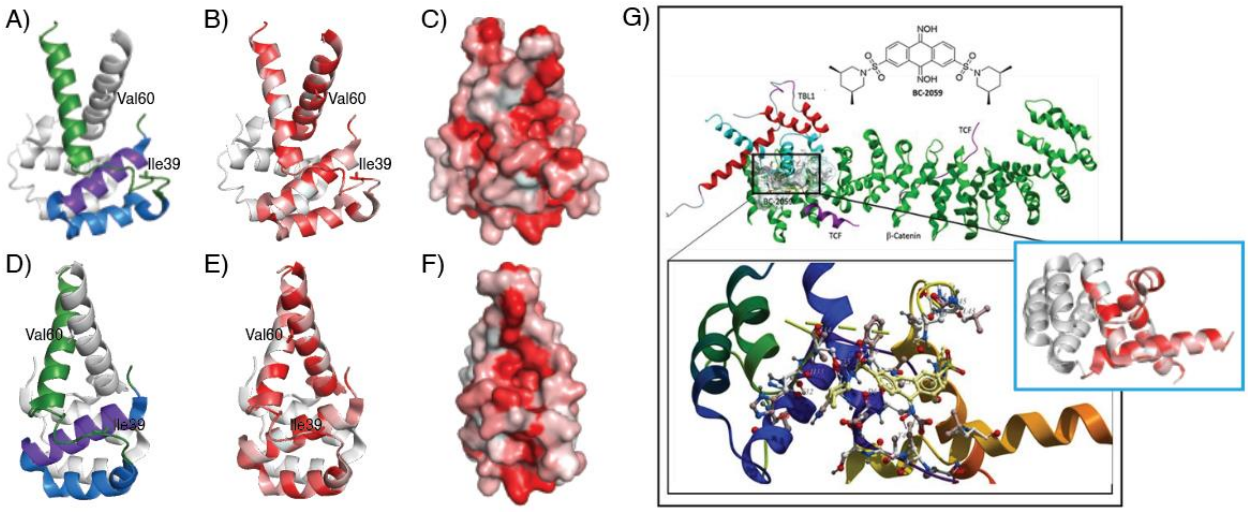


Figure1. TBL1 LisH H1 forms part of the hydrophobic pocket predicted to be important for Tegavivint binding. A, B, C and D, E, F are two different orientations of TBL1 (N terminus 52-141 amino acid) homodimer. A and D had one monomer of TBL1 colored while the other is in grey. The monomer is colored green and the LisH domain at the N terminus is colored blue, with the first helix used in synthetic yeast repression assays highlighted purple. B and E are the same structure colored by hydrophobicity with red meaning higher hydrophobicity. Residues further away from the front facing binding pockets are not colored for simplicity. C and F are surface representations colored by hydrophobicity. Residues Val60 and Ile39 are represented by sticks in A, B, D and E. The black box in G is the docking result for Tegavivint originally published in Soldi et al., 2021. The top portion shows the chemical structure of Tegavivint (BC2059) as well as its predicted binding with TBL1. The bottom potion shows a zoomed in view. The structure in the blue box is the same as B and E but oriented to match the orientation from Soldi et al. highlighting the hydrophobic binding pocket in red.

Experimental Results

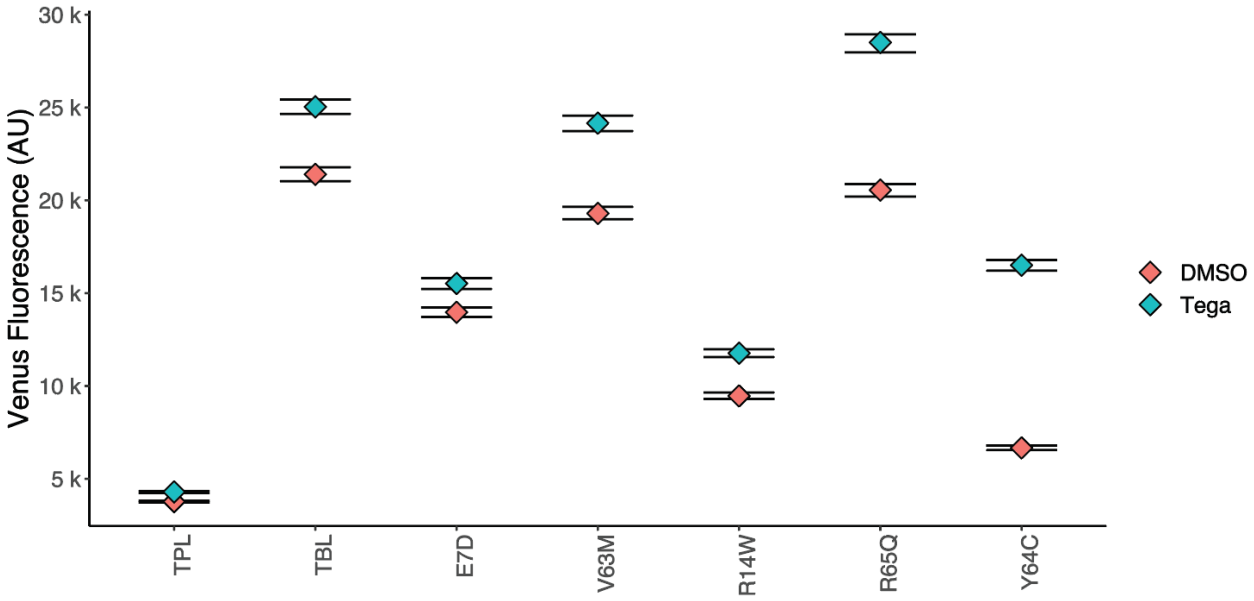


Figure2. Effect of Tegavivint on cancer associated TBL1 mutants. TPL serves as the negative control. Nonsynonymous TBL1 mutations are ordered to the right of TBL1. Diamonds represents the median of at least 10,000 individually measured yeast cells while the error bar represent 95% CI. The color corresponds to Tegavivint treatment or DMSO control.

We first optimized the Tegavivint assay and determined that treating with 100nM of Tegavivint after four hours gave the best resolution, and these parameters were kept for the rest of the experiment. To determine if the yeast platform has potential applications in precision medicine, we measured the effect of Tegavivint on multiple cancer associated TBL1 mutants (Figure2). While Tegavivint treatment resulted in measurable relief of repression in all the TBL1 strains tested, the degree of which varied among the strains. All the mutants are predicted to be located on the side of the LisH H1 helix that interacts with β-catenin which is also the predicted binding site of Tegavivint (Soldi et al. 2021), with the exception of V63M, which is located on the other face of the helix and shows little difference in response when compared to wildtype TBL1. E7D and R14W appear to be least responsive to Tegavivint, and those residues are closely located on the helix. R65Q, which is on the same side as E7D and R14W had surprisingly little effect on both repression and Tegavivint response, which might suggest the change from a positively charged R to a polar Q can be accommodated at the site that is likely important for both TBL1 and Tegavivint function.

Previous work measured repression activity for 65 LisH H1 sequences across eukaryotes as well as their protein stability (Leydon, Ramos Báez, and Nemhauser 2022). To predict the LisH H1 repressor function based on their amino acid sequence, we first attempted simple linear regression models on these 65 sequences. Simple one-hot encoding and NLF encoding (physiochemical encoding from Nanni and Lumini 2011) for the LisH H1 sequecnes as inputs resulted in poor predictability for both repressibility and protein stability. Likely because the training dataset is too small for these simple approaches (data not shown).

More sophisticated approaches were suggested by Dr. Alyssa La Fleur from Dr. Georg Seelig's lab who also helped with the interpretation of the results. We attempted the use EVE (evolutionary model of variant effect) model for prediction of repression. EVE is an unsupervised model that given a training set of sequences, learns constraints at each position and predicts how detrimental a mutation might be given how often it is found in the training dataset (Frazer et al. 2021). We trained an EVE model with the 65 sequences with experimental repression results as well as 1055 H1 sequences from UniProt. This modeling work was performed by Elizabeth S. Duan, a graduate rotation student in the lab.

After training the model, EVE takes as input a protein sequence and one of the main outputs from EVE is an EVE score that is a predicted score of "pathogenicity" with 0 being the most benign and 1 being the most pathogenic. This "pathogenicity" score predicts how likely the variations in the protein sequence are to be tolerated based on what is observed in the training set of sequences. We first attempted to correlate EVE score of the various LisH H1 sequences to repressibility, but both training sets gave poor predictability (data not shown). Instead of looking at whole H1 sequences, we next looked at EVE scores for all possible amino acid substitutions from the designated wildtype sequence in the training set (Figure3).

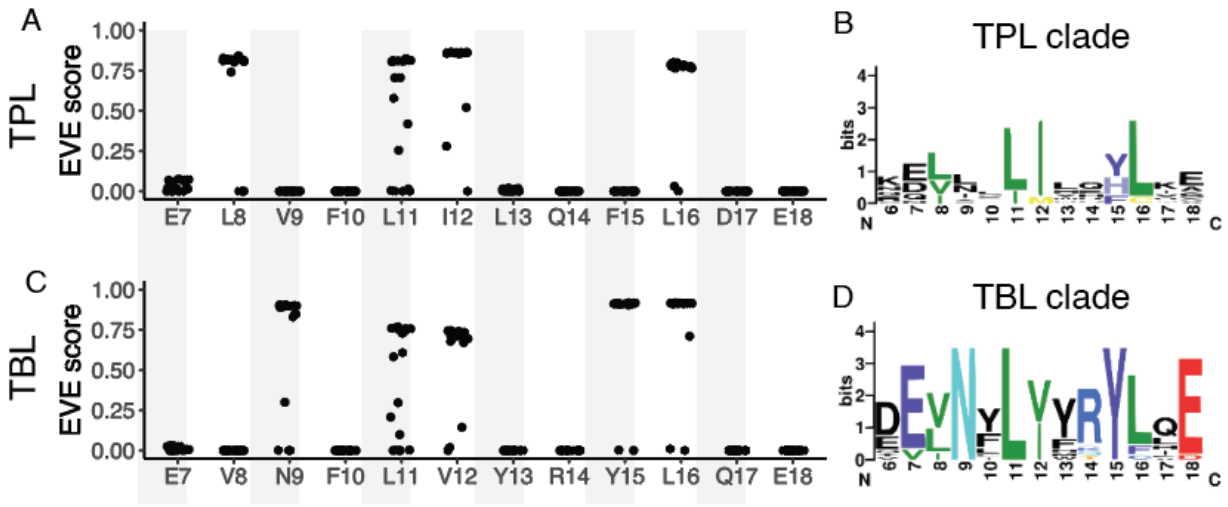


Figure3. EVE scores for models trained on 1055 H1 sequences using A) TPL or C) TBL1 LisH H1 as the designated wildtype sequence during training. Each dot is an amino acid substitution. B and D) Logo plots for clades I (TPL containing clade) and clade II (TBL1 containing clade) of LisH H1 sequences originally published in (Leydon, Ramos Báez, and Nemhauser 2022).

Training the model using TPL as the designated wildtype sequence predicted strong pathogenic effect for positions 8, 11, 12, and 16. Those positions correlated well with the logo plot for LisH H1's TPL containing clade where the three positions appears highly conserved (Leydon, Ramos Báez, and Nemhauser 2022). Using TBL1 as the designated wildtype sequence, however, resulted in the model predicting positions 9, 11, 12, 15, 16 as highly pathogenic when mutated. Given that TBL1 LisH H1 is a weaker repressor than TPL LisH H1, the loss of pathogenicity for position 8 for TBL1 might point to a position important for the strong repression observed in TPL LisH H1.

Conclusions

In this work we explored both the commercialization potential of the yeast system as a drug screening platform, and worked to build a model that could improve the repressibility a short LisH H1 repression domain. Testing Tegavivint treatment against different TBL1 LisH H1 mutants revealed the assay system was capable of measuring differences in the mutant's response to Tegavivint, and it revealed certain mutants may be less responsive to Tegavivint than others. In terms of prediction pression, our best performing model so far is the EVE model. Though it was not successful in quantitatively predicting repressibility based on sequence, the model appeared to have learned about amino acid conservation at each position. By comparing the EVE model built using TPL and TBL1 LisH H1 sequence pointed to a potentially important position for the strong repression observed in TPL.

The fact that the screening platform was able to measure differences in responses to Tegavivint for the various mutants suggests it has potential application in precision medicine. Certain mutants (e.g. E7D and R14W) tested in this experiment may be less responsive to Tegavivint treatment. The platform can also potentially be used to screen for chemicals related to Tegavivint to identify better suited small molecule candidates for these mutants. It is important to note that since these results are all obtained from the yeast platform, their actual correlation to the mutant's pathogenicity and responsiveness to Tegavivint in the context of cancer has yet to be determined. Further study in mammalian cells are needed to verify the clinical validity of the findings presented here.

While none of the models presented here are successful in predicting repression strength based on sequence, EVE showed promise in identifying residues that are conserved for TPL and TBL1 LisH H1. This application was more successful likely because it is more closely in line with the original purpose of EVE. Here we only explored the EVE scores, which are related to predicted pathogenicity, but EVE also provides an evolutionary index, which is the log likelihood of a sequence compared to a given "wildtype". This metric could potentially be used as an additional parameter for an "augmented" approach of linear-regression analysis where our repression assay labeled data can be modeled by index combined with one-hot encoded amino acid sequences. This combination approach had been suggested to improve model performance (Hsu et al. 2022). Additional approaches utilizing transformer protein language models and deep learning can also be evaluated in the future (Lin et al. 2023; Biswas et al. 2021).

Methods

Flow Cytometry

Yeast strains were originally made and published in (Leydon, Ramos Báez, and Nemhauser 2022). All measurements recorded from a BD special order cytometer with a 514-nm laser with a cutoff at 525nm. Yeasts were streaked on the appropriate drop out media and grew for two days at 30C. A small amount of colony was picked for each strain and vortexed with 1mL of media. 100uL of this mixture was measured in the cytometer and the concentration was used to calculate the appropriate dilution to get to 1.5 events/uL in 6mL of the appropriate drop off media. The 5mL overnight was shaken overnight at 30C and 250rpm. Roughly sixteen hours later, 100uL of the overnight was measured in the cytometer and the concentration was used to adjust the overnight to 200 events/uL for all strains. 990uL of each strain was placed in a 96 deep well plate in triplicates for both Tegavivint treatment group and DMSO control. 10uL

of Tegavivint or DMSO was added to the appropriate wells. The working concentration of Tegavivint was 100nM. The deep-well plate was shook at 350rpm at 30C for four hours. The plate was read one row at a time using 100uL of the culture while the rest kept shaking in the shaker. Cytometer was set to record until 10,000 events or have reached 90uL under “fast fluidics”. The events were recorded and exported to be analyzed in R.

TBL1 structure
Protein structure files were downloaded from Uniprot (O60907) identifier 2XTE and visualized in PyMOL. Coloring scheme for hydrophobicity was from pymolwiki (https://pymolwiki.org/index.php/Color_h).

EVE modeling
EVE models were trained using methods outlined in (Frazer et al. 2021) and their GitHub page (https://github.com/OATML/EVE). The model was trained using 65 sequences with experimental repression results or 1055 H1 sequences from UniProt. The wildtype sequence was set to either TPL or TBL1 in both cases. The generated EVE score was used to generate the plots in R.

Contributions

Experimental design and analysis by Eric J.Y. Yang, Alexander R. Leydon, Elizabeth S. Duan, and Jennifer L. Nemhauser. EVE modeling and analysis performed by Elizabeth S. Duan. Figures by Eric J.Y. Yang, Alexander R. Leydon, and Elizabeth S. Duan.

References

Biswas, Surojit, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. 2021. "Low-N Protein Engineering with Data-Efficient Deep Learning." *Nature Methods* 18 (4): 389–96. https://doi.org/10.1038/s41592-021-01100-y.

Brophy, Jennifer A. N., Katie J. Magallon, Lina Duan, Vivian Zhong, Prashanth Ramachandran, Kiril Kniazev, and José R. Dinneny. 2022. "Synthetic Genetic Circuits as a Means of Reprogramming Plant Roots." *Science* 377 (6607): 747–51. https://doi.org/10.1126/science.abo4326.

Choi, Hyo-Kyoung, Kyung-Chul Choi, Jung-Yoon Yoo, Meiying Song, Suk Jin Ko, Chul Hoon Kim, Jin-Hyun Ahn, Kyung-Hee Chun, Jong In Yook, and Ho-Geun Yoon. 2011. "Reversible SUMOylation of TBL1-TBLR1 Regulates β-Catenin-Mediated Wnt Signaling." *Molecular Cell* 43 (2): 203–16. https://doi.org/10.1016/j.molcel.2011.05.027.

Frazer, Jonathan, Pascal Notin, Mafalda Dias, Aidan Gomez, Joseph K. Min, Kelly Brock, Yarin Gal, and Debora S. Marks. 2021. "Disease Variant Prediction with Deep Generative Models of Evolutionary Data." *Nature* 599 (7883): 91–95. https://doi.org/10.1038/s41586-021-04043-8.

Gander, Miles W., Justin D. Vrana, William E. Voje, James M. Carothers, and Eric Klavins. 2017. "Digital Logic Circuits in Yeast with CRISPR-DCas9 NOR Gates." *Nature Communications* 8 (1): 15459. https://doi.org/10.1038/ncomms15459.

Hsu, Chloe, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. 2022. "Learning Protein Fitness Models from Evolutionary and Assay-Labeled Data." *Nature Biotechnology* 40 (7): 1114–22. https://doi.org/10.1038/s41587-021-01146-5.

Leydon, Alexander R., Román Ramos Báez, and Jennifer L. Nemhauser. 2022. "A Single Helix Repression Domain Is Functional across Diverse Eukaryotes." *Proceedings of the National Academy of Sciences* 119 (41): e2206986119. https://doi.org/10.1073/pnas.2206986119.

Leydon, Alexander R, Wei Wang, Hardik P Gala, Sabrina Gilmour, Samuel Juarez-Solis, Mollye L Zahler, Joseph E Zemke, Ning Zheng, and Jennifer L Nemhauser. 2021. "Repression by the Arabidopsis TOPLESS Corepressor Requires Association with the Core Mediator Complex." Edited by Irwin Davidson, James L Manley, Irwin Davidson, and Lucia Strader. *ELife* 10 (June): e66739. https://doi.org/10.7554/eLife.66739.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. 2023. "Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model." *Science* 379 (6637): 1123–30. https://doi.org/10.1126/science.ade2574.

Martin-Arevalillo, Raquel, Max H. Nanao, Antoine Larrieu, Thomas Vinos-Poyo, David Mast, Carlos Galvan-Ampudia, Géraldine Brunoud, Teva Vernoux, Renaud Dumas, and François Parcy. 2017. "Structure of the Arabidopsis TOPLESS Corepressor Provides Insight into the Evolution of Transcriptional Repression." *Proceedings of the National Academy of Sciences of the United States of America* 114 (30): 8107–12. https://doi.org/10.1073/pnas.1703054114.

Nanni, Loris, and Alessandra Lumini. 2011. "A New Encoding Technique for Peptide Classification." *Expert Systems with Applications* 38 (4): 3185–91. https://doi.org/10.1016/j.eswa.2010.09.005.

Oberoi, Jasmeen, Louise Fairall, Peter J. Watson, Ji-Chun Yang, Zsolt Czimmerer, Thorsten Kampmann, Benjamin T. Goult, et al. 2011. "Structural Basis for the Assembly of the SMRT/NCoR Core Transcriptional Repression Machinery." *Nature Structural & Molecular Biology* 18 (2): 177–84. https://doi.org/10.1038/nsmb.1983.

Pai, Sachin Gopalkrishna, Benedito A. Carneiro, Jose Mauricio Mota, Ricardo Costa, Caio Abner Leite, Romualdo Barroso-Sousa, Jason Benjamin Kaplan, Young Kwang Chae, and Francis Joseph Giles. 2017. "Wnt/Beta-Catenin Pathway: Modulating Anticancer Immune Response." *Journal of Hematology & Oncology* 10 (1): 101. https://doi.org/10.1186/s13045-017-0471-6.

Plant, Alastair Robert, Antoine Larrieu, and Barry Causier. 2021. “Repressor for Hire! The Vital Roles of TOPLESS-Mediated Transcriptional Repression in Plants.” *New Phytologist* 231 (3): 963–73. https://doi.org/10.1111/nph.17428.

Pray, Betsy A., Youssef Youssef, and Lapo Alinari. 2022. “TBL1X: At the Crossroads of Transcriptional and Posttranscriptional Regulation.” *Experimental Hematology* 116 (December): 18–25. https://doi.org/10.1016/j.exphem.2022.09.006.

Soldi, Raffaella, Tithi Ghosh Halder, Samuel Sampson, Hariprasad Vankayalapati, Alexis Weston, Trason Thode, Kapil N. Bhalla, et al. 2021. “The Small Molecule BC-2059 Inhibits Wingless/Integrated (Wnt)-Dependent Gene Transcription in Cancer through Disruption of the Transducin β-Like 1-β-Catenin Protein Complex.” *Journal of Pharmacology and Experimental Therapeutics* 378 (2): 77–86. https://doi.org/10.1124/jpet.121.000634.

Zhai, Haotian, Li Cui, Zhen Xiong, Qingsheng Qi, and Jin Hou. 2022. “CRISPR-Mediated Protein-Tagging Signal Amplification Systems for Efficient Transcriptional Activation and Repression in Saccharomyces Cerevisiae.” *Nucleic Acids Research* 50 (10): 5988–6000. https://doi.org/10.1093/nar/gkac463.

Reflection

Synthetic biology sets itself apart from the larger field of biotechnology with its emphasis on engineering principles, parts characterization, and modularity (National Academies of Sciences et al., 2018; Patron, 2020). Many pioneering breakthroughs in synthetic biology happened in prokaryotic systems due to their relative ease in manipulation and faster turnaround for the design-build-test cycle (Elowitz & Leibler, 2000; Gander et al., 2017; Zúñiga et al., 2020). However, many of the technologies developed and lessons learned from bacterial systems have been successfully adapted to the field of plant synthetic biology: efforts in engineering synthetic promoters, building logic circuits, and cell lineage tracing just to name a few (Brophy et al., 2022; Brückner et al., 2015; Cai et al., 2020; Guiziou et al., 2023; Yang & Nemhauser, 2022). Certain tools specific to plants had also been developed such as the plant hormone-based HACR (hormone activated Cas9-based repressor) system, while other tools borrowed from plants proved to be useful in bacterial systems, such as the Auxin-inducible degron technology (Khakhar et al., 2018; Nishimura et al., 2009). However, the current state of genetic parts characterization and modularity is still a far cry from the equivalent electrical engineering system that we hope to emulate. To truly realize the goal of having standardized genetic parts that can be mixed and matched with predictable behavior, there are certain gaps that I believe still need to be bridged.

I believe a standardized way of quantifying promoter activity is necessary for the plant synthetic biology community. One major effort that had been a major driving force behind these innovations for synthetic biology is the standardization of genetic parts to facilitate sharing among labs and speed up the construction process. This is evident in the emphasis in assembly of documented parts in iGEM competition, as well as the various cloning standards proposed in the plant synthetic biology community (Engler et al., 2014; Sarrion-Perdigones et al., 2011). Work done in Chapter1 contributes to this overall effort in characterizing parts for dissemination by adhering to the MoClo standard. Recent publications on promoter activity relies on ratiometric reporting where the output of a reporter driven by the promoter in question is normalized by a second florescent output driven by a constitutive promoter. This approach attempts to reduce the variability found between samples, experiments, and the randomness of DNA insertion events. However, the parts characterized, while comparable and consistent within the manuscript, often are not directly comparable to other published promoter parts. A survey of recent publications on promoter engineering reveals variable choices in primary reporter, secondary reporter, and secondary promoter (Brophy et al., 2022; Brückner et al., 2015; Cai et al., 2020; Moreno-Giménez et al., 2022). These differences make it difficult to compare the activity of parts across experiments. There are on-going efforts in tackling this problem. The Golden Braid system championed by Dr. Deigo Orzaez, for example, proposed a Luciferase/Renilla transient assay for standardization where the promoter in question

will be driving Luciferase while a 35S promoter drives Renilla (Vazquez-Vilar et al., 2017). More recent publications have also suggested alternative luciferases for reporting (González-Grandío et al., 2021). Apart from the choice of reporter, the choice of the ratiometric promoter also requires consensus. As Zhou and colleagues have pointed out, more work is needed to determine the ideal normalization construct and a well-chosen constitutive promoter will be crucial (Zhou et al., 2023). The constitutive promoters screened in Chapter1 provided additional candidates for this normalizing promoter. It would also be interesting to see whether promoters taken from genes with orthologs across species, like those identified in Chapter2, maintain stable expression across species. If so, that would make them even more attractive candidates for ratiometric reporting. It is also likely that depending on the activity of the promoter of interest, a strong ratiometric promoter may not always be ideal for normalization, and so a series of standardized constitutive promoters of various strengths may be necessary.

Additionally, I believe genetic parts should be characterized in ways that can be used with predictive models to truly realize the ideal of plug-and-play synthetic biology. In an ideal scenario, the output of each transcriptional unit in a construct can be predicted just by knowing the properties of its components. Electrical circuits behave extremely predictably and if the resistance of each of the components is known, the change in voltage and current at each point in the circuit can be calculated. The interaction between genetic components is complex and we are learning that individual parts are not as modular as we hope them to be. For example, it was found that the choice of terminator effects promoter activity in ways we are only just beginning to explore (Andreou et al., 2021). For me, in an ideal scenario, each genetic component will have a coefficient associated with it, and plugging these numbers into a model will allow estimation of transcriptional activity when the parts are assembled. The current method of characterizing genetic parts does not really allow these kinds of predictions of transcription level after assembly. Given the complex and noisy environment of a biological system, focusing on parts that have stable behaviors might be a great starting point, such as those borrowed from constitutively expressed genes. While the various promoters and terminator parts borrowed from stably expressed genes likely still interact to affect the final transcription rate, the behavior of these parts are at least stable across external conditions and development, making it more likely to capture their behavior and interaction using simple models.

One piece that I believe is still missing from the plant synthetic biology toolbox is a titratable promoter. Having such a tool will not only speed up the design-build-test cycle, but also help characterize logic gates. A problem that remains even with well characterized promoters is that a lot of the actual construction of gene circuits still relies on trial-and-error since we often do not know what is the exact level of output required for a construct which really slows down the engineering cycle (Guiziou et al.,

2023; Moreno-Giménez et al., 2022). Having a tunable promoter would allow the testing of a wide range of promoter activities with a single construct. Another powerful application of a titratable promoter is in characterizing logic gates. Using the NOR gates built in Chapter1 as an example, the input and outputs from the NOR gates are gRNAs. As the concentration of one of the input gRNA increases, the output from the gate should decrease, and we could define regions of low and high concentrations of inputs that result in output levels that is considered either ON or OFF, a graphical representation of which is called a response function of the logic gate. To allow stringing together of logic gates, the output level of an upstream gate needs to match input level required for the downstream gate to turn the downstream gate OFF. However, currently there are no easy ways to determine the response function of logic gates in plants, and as a result the assembly of logic gates into more complex circuits would necessarily require trial-and-error. One approach to solve this problem in bacterial systems leverages titratable promoters to map out response functions of logic gates which allows computer assisted circuit design automation (Nielsen et al., 2016). Such a tool would significantly speed up the process of combining individual logic gates into more complex circuits in plants.

Synthetic biology has been making tremendous progress over the years, and the subfield of plant synthetic biology is no exception. While this thesis contributed to providing promoter parts and logic gates adhering to the current best practices and standards, I believe moving forward with more emphasis on consensus in parts characterization among the community with a specific focus on integrating these experimental measurements into predictive models will be crucial if genetic constructs are to be designed with more predictable outputs. I believe these will be some of the important next steps for the field of plant synthetic biology.

Andreou, A. I., Nirkko, J., Ochoa-Villarreal, M., & Nakayama, N. (2021). *Mobius Assembly for Plant Systems highlights promoter-terminator interaction in gene regulation* (p. 2021.03.31.437819). bioRxiv. https://doi.org/10.1101/2021.03.31.437819

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

Elowitz, M. B., & Leibler, S. (2000). A synthetic oscillatory network of transcriptional regulators. *Nature*, *403*(6767), Article 6767. https://doi.org/10.1038/35002125

Engler, C., Youles, M., Gruetzner, R., Ehnert, T.-M., Werner, S., Jones, J. D. G., Patron, N. J., & Marillonnet, S. (2014). A Golden Gate Modular Cloning Toolbox for Plants. *ACS Synthetic Biology*, *3*(11), 839–843. https://doi.org/10.1021/sb4001504

Gander, M. W., Vrana, J. D., Voje, W. E., Carothers, J. M., & Klavins, E. (2017). Digital logic circuits in yeast with CRISPR-dCas9 NOR gates. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15459

González-Grandío, E., Demirer, G. S., Ma, W., Brady, S., & Landry, M. P. (2021). A Ratiometric Dual Color Luciferase Reporter for Fast Characterization of Transcriptional Regulatory Elements in Plants. *ACS Synthetic Biology*, *10*(10), 2763–2766. https://doi.org/10.1021/acssynbio.1c00248

Guiziou, S., Maranas, C. J., Chu, J. C., & Nemhauser, J. L. (2023). An integrase toolbox to record gene-expression during plant development. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-37607-5

Khakhar, A., Leydon, A. R., Lemmex, A. C., Klavins, E., & Nemhauser, J. L. (2018). Synthetic hormone-responsive transcription factors can monitor and re-program plant development. *ELife*, *7*, e34702. https://doi.org/10.7554/eLife.34702

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. *ACS Synthetic Biology*, *11*(9), 3037–3048. https://doi.org/10.1021/acssynbio.2c00238

National Academies of Sciences, E., Studies, D. on E. and L., Sciences, B. on L., Technology, B. on C. S. and, & Biology, C. on S. for I. and A. P. B. V. P. by S. (2018). Biotechnology in the Age of Synthetic Biology. In *Biodefense in the Age of Synthetic Biology*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK535871/

Nielsen, A. A. K., Der, B. S., Shin, J., Vaidyanathan, P., Paralanov, V., Strychalski, E. A., Ross, D., Densmore, D., & Voigt, C. A. (2016). Genetic circuit design automation. *Science*, *352*(6281), aac7341. https://doi.org/10.1126/science.aac7341

Nishimura, K., Fukagawa, T., Takisawa, H., Kakimoto, T., & Kanemaki, M. (2009). An auxin-based degron system for the rapid depletion of proteins in nonplant cells. *Nature Methods*, *6*(12), Article 12. https://doi.org/10.1038/nmeth.1401

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

Sarrion-Perdigones, A., Falconi, E. E., Zandalinas, S. I., Juárez, P., Fernández-del-Carmen, A., Granell, A., & Orzaez, D. (2011). GoldenBraid: An Iterative Cloning System for Standardized Assembly of Reusable Genetic Modules. *PLOS ONE*, *6*(7), e21622. https://doi.org/10.1371/journal.pone.0021622

Vazquez-Vilar, M., Quijano-Rubio, A., Fernandez-del-Carmen, A., Sarrion-Perdigones, A., Ochoa-Fernandez, R., Ziarsolo, P., Blanca, J., Granell, A., & Orzaez, D. (2017). GB3.0: A platform for plant bio-design that connects functional DNA elements with associated biological data. *Nucleic Acids Research*, *45*(4), 2196–2209. https://doi.org/10.1093/nar/gkw1326

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

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

Zúñiga, A., Guiziou, S., Mayonove, P., Meriem, Z. B., Camacho, M., Moreau, V., Ciandrini, L., Hersen, P., & Bonnet, J. (2020). Rational programming of history-dependent logic in cellular populations. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-18455-z