Developing and applying transient expression systems to identify and understand gene regulatory elements in plants

Jackson Tonnies

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Christine Queitsch

Takato Imaizumi

Jennifer Nemhauser

Program Authorized to Offer Degree:

Biology

University of Washington

Abstract

Developing and applying transient expression systems to identify and understand gene regulatory elements in plants

Jackson Tonnies

Chair of the Supervisory Committee:
Christine Queitsch
Genome Sciences

In this dissertation, I develop transient expression systems to enable massively parallel reporter assays (MPRAs) in crop models. I apply these transient assay systems to identify, characterize and design plant gene regulatory elements.

To allow for the scale of MPRAs, I develop a high-efficiency transformation protocol for maize mesophyll protoplasts. Using this protocol, I can transform millions of maize mesophyll protoplasts without losing viability.

I use transient transformation of maize protoplasts along with *Agrobacterium*-mediated transient transformation in tobacco to develop a plant-specific MPRA, plant STARR-seq, to

ascertain the activity of regulatory elements such as core promoters and enhancers. My contribution was essential in assuaging reviewer concerns about prior studies using a different, animal-specific assay design.

Using plant STARR-seq, I and my co-author Tobias Jores identify and characterize 75,000 core promoters from maize, sorghum and Arabidopsis. We identify features required for the function of core promoters in both maize and tobacco, and we find differences in the effect of GC content between maize and tobacco elements corresponding to the GC content of their respective genomes. We use machine learning and *in silico* evolution to design synthetic core promoters that rival the viral Cauliflower Mosaic Virus 35S core promoter in activity.

Finally, I use plant STARR-seq to dissect the activity of three known light-responsive enhancers. I perform deep mutational scans of all three enhancers and identify regions in which mutations affect their function in the dark and light. I combine these regions to create synthetic enhancers with a wide range of transcriptional responses including enhancers that show greater light response than any of the original enhancers. I show that most of the observed enhancer activity is explained by an additive model, albeit there are rare exceptions.

TABLE OF CONTENTS

List of Figures

List of Tables

ACKNOWLEDGEMENTS

Thank you to my partner Claudia who has supported me through my PhD, especially when projects didn't work out or I had a lot on my plate. You have been amazing throughout the difficult process of finishing a PhD and have reminded me to prioritize health during stressful times.

Thank you to my family for being there for me and listening to me vent about difficulties that occur along the way, even though we are separated by more distance than I'd like.

Thank you to my friends that I have made during my time in Seattle. My cohort has been one of the best friend groups anyone could hope to be a part of. Additionally, the friends I have made in my lab have been some of the brightest and funniest people I know. I am sad to move away for my upcoming job.

Thank you to my advisor, Christine Queitsch, for sharing my love of plants. Christine has consistently encouraged me to be excited about my research and to search for interesting questions to tackle. Additionally, thank you for providing the amazing group of people that are the Queitsch/Cuperus lab.

Thank you to all the members of the lab past and present. Especially, thank you to those who went out of their way to mentored me. Thank you Cris for showing me how to work with plants, Kerry and Ken who were pivotal to me developing computational skills, Beth for being an amazing example of a careful and thoughtful scientist and Tobias for collaborating with me on many projects (and getting plant STARR-seq to work).

Chapter 1. INTRODUCTION

1.1 PLANT CULTIVATION AND IMPROVEMENT

Plants form the basis of the agricultural system that feeds humanity. Our agricultural system and practices have developed over millennia with early evidence of domesticated crops dating back to 15,000 years ago[1]. During this same period, the human population has greatly increased. To match population growth, agricultural practices have been constantly improving and more land has been converted into farmland. It is estimated that 80% of deforestation is driven by agriculture[2]. However, the amount of land that can be converted into farm land is not infinite. As a society, we must strike a balance between increasing the amount of farmed land and protecting ever dwindling natural refuges. Meanwhile, yearly population growth has greatly increased in the past couple hundred years. This growth peaked in 1968 at a rate of 2.1% annual increase in population. This growth is much greater than during most of human history, with the average annual population increase being 0.04% between the years 10,000 BCE and 1700 CE[3]. As the population continues to increase in the coming century to a projected 11 billion people[4], it is imperative to improve the efficiency of plant-based agriculture. Improving plant-based agriculture is critical both to provide food security, but also to minimize the amount of protected natural land converted for farming. The challenge of improving agricultural systems and practices is even more pressing in the light of a changing climate which is predicted to negatively impact crop yields and the amount of farmable land[5,6].

Historically, there have been many improvements in agricultural practices to improve plant yields. These improvements center around changing the abiotic environment, mitigating unwanted biotic factors, and altering the crops we grow. Methods to improve the abiotic

environment in which plants grow include developing better irrigation techniques, improving fertilizer use and availability, and managing topsoil. Deterring unwanted pest animals, weeds, and pathogens have benefited from adoption of prey animals, techniques such as tilling under weeds, and the creation of novel chemicals (herbicides, pesticides, fungicides). Finally, we have directly altered the plants we grow by passively selecting for desirable traits, actively breeding plants, and utilizing biology principles like hybrid vigor. Modern-day farming results from a complex combination of the many improvements from a wide variety of fields and disciplines.

Some of the most interesting improvements in agriculture occur at the intersection of biology and farm management techniques. An example are modern apple orchards. Orchards have transitioned away from traditional apple trees towards growing dwarf apple trees. Growing dwarf apple trees provide two main benefits over their traditional counterparts. One benefit is that apple trees can be planted at much higher density on farmland increasing the apple yield per acre[7]. The second benefit is that the trunks and branches of dwarf trees can be coerced to grow into a planar structure that make maintenance of trees and collections of apples easier. As opposed to previous apple trees which have a cone of branches around a trunk, dwarf apple trees can be grown in a flat plane along a fence. This wall of apple trees is reminiscent of how grapes are grown in vineyards. This method simplifies spraying trees for pests and provides easier access to apples for picking, and allows the use of labor saving practices in commercial farms[8].

The dwarf phenotype that enabled the adoption of growing apples in a 2-D wall would have taken many years to breed into a wide range of apple varieties. This tedious work was circumvented by the discovery of a dwarfing root stock in apples. Researchers identified a variety of apple tree that would confer dwarfism to commercial tree varieties through traditional grafting onto the dwarfing variety's roots. This discovery of dwarfing root stock saved years of

intensive breeding of commercial lines and changed the way apples are grown. I am convinced that many future improvements in plant cultivation will also involve discoveries in fundamental plant biology that enable changes in farm management.

1.2 THE FUTURE OF PLANT-BASED AGRICULTURE

Breeding crops has been a staple method for improving plant-based agriculture for thousands of years. New technologies have enabled rapid advances in the last 40 years[9]. One such advance is the development of speed-breeding. The goal of speed-breeding is reducing the generation time of plants by using optimized temperature and light cycles in greenhouses. Additionally, comprehensive plant phenotyping combined with the use of dense single nucleotide polymorphism (SNP) marker sets have improved breeders' ability to identify genotypes associated with agriculturally valuable traits. Finally, breeders now use computational models to improve phenotype predictions of crosses and decide which plants should be crossed. These new models incorporate knowledge of previous generations of plants as well as other large datasets. As these models increase in complexity, they have begun to incorporate more types of data to inform breeding practices.

One tool that opens many new possibilities is CRISPR based genome editing to engineer plants. The discovery of the CRISPR-cas9 system sparked a revolution across almost every field of biology by simplifying the process of altering specific genomic locations. The field of crop improvement is no different. Particularly interesting examples include preventing disease by altering a set of crop genes that confer susceptibility to pathogens[10]. Additionally, CRISPR has been used to replicate the process of domestication in ground cherry, a relative of tomatoes[11]. Scientists achieved this by altering a set of genes in ground cherry that were homologous to those

that were known to be beneficial in tomatoes. In a few short years, this approach produced phenotypes in in ground cherry that replicated the results of breeding for thousands of years in tomato. Although there are many possible applications, there are significant regulatory hurdles before edited plants can be grown around the world. The European Union has strict rules requiring a lengthy approval process for gene-edited crops even if the crops contain only edits that do not introduce foreign DNA and are indistinguishable from natural variation[12]. In the United States, crop varieties that carry small deletions and do not contain foreign DNA can be grown similarly to varieties that obtained similar changes through traditional breeding.

An exciting new field with great potential to improve crops is synthetic biology. Research in the field of plant synthetic biology includes building plants biosensors, altering plant plastids, and creating new metabolic pathways[13]. These studies bring engineering principles to the field of biology to design and create new biological functionalities. The application of engineering principles requires reliable characterization of the function of biological elements. Synthetic biology approaches are becoming more powerful as biologists become better at characterizing the elements used to engineer biological systems.

Modern methods of improving and engineering future crops rely heavily on detailed knowledge of plant biology. As the ways in which we can improve plants become more nuanced they depend more on having a foundational understanding of how plants function. In my doctoral studies, I focus on improving understanding in one key area of plant biology. Specifically, I focus on how plants regulate their gene expression and how these mechanisms are encoded in regulatory DNA.

1.3 UNDERSTANDING PLANT REGULATION

A key aspect of understanding plant biology is understanding how plant genes are regulated. Plant genomes contain tens of thousands of genes that encode proteins which govern plant growth and development. Plant genomes also contain regulatory regions with complex instructions for when, where and to what level each gene will be expressed.

A key aspect of understanding plant biology is understanding how plant genes are regulated. Within plant genomes there are tens of thousands of genes that encode proteins which are instrumental in the creation and maintenance of a plant. Also encoded in plant DNA, are complex instructions for when and to what degree each gene should be expressed. However, finding which parts of the genome encodes the regulatory instructions is non-trivial. Plant genomes can consist of billions of base pairs, many of which do not have an expected function.

Beginning in the 70s, researchers found that transcribed regions of chromatin are hypersensitive to endonucleases[14–18]. Hypersensitivity to endonucleases – or DNA accessibility – was further localized to regions up- and downstream of genes undergoing active transcription[19]. DNA accessibility was shown to be associated with regions that contain cis-regulatory elements[20] and is now used to define regulatory regions. Accessibility upstream of a gene's transcription start site is associated with gene expression[21]. However, because accessibility can arise through suppressor binding or poised transcription factors, it is not a good predictor of gene expression *per se*. Even if regulatory sites show dynamic changes in accessibility, their respective target genes may not show corresponding changes in gene expression. Prior research in the Queitsch/Cuperus lab found that the vast majority of the about 45,000 accessible sites in the genome of the model Arabidopsis are static even though gene expression changed dramatically in some conditions[21]. How then, do we test the function of these many possible regulatory regions?

1.4 MASSIVELY PARALLEL REPORTER ASSAYS IN PLANTS

To meet the challenge of testing the function of many putative plant regulatory regions, we borrowed a method developed in the animal world. Specifically, we use massively parallel reporter assays (MPRAs). MPRAs are enabled by next-generation sequencing which allows cheap sequencing of short DNA fragments at massive scale. One of the early adaptations of MPRAs was STARR-seq (self-transcribing active regulatory region sequencing)[22]. This assay relies on cloning millions of DNA fragments of interest into the 3' UTR of a reporter gene, transforming this library into live cells and measuring the transcription of the reporter gene. DNA fragments that increase the expression of the reporter gene encode enhancers. Enhancers are regulatory elements that interact with core promoters to drive transcription in a condition –or tissue-specific manner. They can function in a distance and are orientation-independent. Testing DNA fragments in this way worked well in both fly and human cells, but we had to overcome hurdles before this method worked in plants.

One hurdle was that plants had few transformation systems that allowed for large-scale library transformations. One of the only transformation systems with the ability to transform millions of constructs in plant cells was transient transformation of tobacco using *Agrobacterium*. However, tobacco is only distantly related to important cereal crops such as maize and wheat. To enable the use of MPRAs in a cereal crop, I developed high-efficiency transient transformation of maize mesophyll protoplasts. Protoplasts are plant cells that have had their cell wall removed. The lack of cell walls allows for transformation of DNA into cells through the plasma membrane. Maize mesophyll protoplasts have been used before for studying a handful of constructs in transient transformation experiments but never with library of hundreds of thousands of

constructs[23]. Optimization of maize mesophyll protoplast preparation and transformation is covered in Chapter 2.

Another hurdle was the position of the tested DNA fragment in the 3' UTR of the reporter gene in the published assay. While this position worked well in animal systems and was used in a prior maize study[12], in our hands, this design failed to yield the expected high activity for a known viral regulatory element that we use as a positive control. Our lab found that testing DNA fragments upstream of 5' UTR resulted in a far larger dynamic range than the original placement in the 3' UTR. This observation held true both in the tobacco leaf and maize protoplast systems, as detailed in Chapter 3.

The optimized plant MPRA opened the way for us to probe the regulatory function of accessible DNA fragments throughout several plant genomes. The first regulatory element we pursued was the core-promoter. Core promoters reside near the transcription start and serve to assemble the core transcriptional machinery. On their own, they do not confer high levels of transcription; however, these are achieved through core promoter interaction with enhancers. We demonstrate differences in core promoter preference between the tobacco and maize system and used machine learning and *in silico* evolution to create new core promoters that lead to high-expression. Our insights into core-promoters in plants are covered in Chapter 4.

Finally, I combined the plant STARR-seq assay with a technique known as deep mutational scanning. This technique systematically tests all possible variants of a sequence for function to identify mutation-sensitive, presumably functional sites and motifs. I utilized mutational scanning to characterize the critical regions in a set of previously identified light-responsive enhancers. I then combined these regions to make synthetic light-responsive enhancers that showed a wider range of expression. This work is described in Chapter 5.

Chapter 2. OPTIMIZING MAIZE MESOPHYLL PROTOPLAST TRANSFORMATION

Abstract

Maize (Zea mays) is an important crop grown throughout the world. As such, it is critical to develop tools to transiently transform maize cells. Transforming maize mesophyll cells is often performed by protoplasting the cells, digesting plant cell walls, and subsequently inserting DNA through electroporation or application of polyethylene glycol (PEG). However, previous methods are optimized for testing relatively few constructs that require thousands to tens of thousands of transformants. Here, we describe a straight-forward method for isolating and transforming millions of maize leaf mesophyll protoplasts. We optimized protoplast isolation and PEG-mediated transformation to efficiently transform large libraries of plasmid constructs into many maize protoplasts. We streamlined our process by removing certain common protoplasting steps, such as washing the protoplasts in W5. Additionally, we improved steps such as centrifugation, transformation, and overnight incubation to work with a greater number of protoplasts. The scale of our protoplast isolation and transformation allows the use of genome-scale experiments such as massively parallel reporter assays in maize.

2.1 INTRODUCTION

Transient transformation of plant cells is a powerful tool for understanding plant biology. Plant cells are difficult to transform because they are encompassed by a cell wall. The plant cell wall comprises an additional barrier through which DNA needs to pass to transform a cell. This

barrier is not present in other commonly transformed cell types such as mammalian or human cells. Past research addressed this challenge by employing protoplasts which are plant cells that lack cell walls. Without cell walls, plant protoplasts are easy to work with in suspension and amenable to transformation techniques such as electroporation and polyethylene glycol (PEG) mediated transformation[24]. While the process of cell wall removal is stressful for plant cells, they retain most of their function. Thus, protoplasts are used in a variety of plants to study gene and protein function[25,26], understand signal transduction[27], and inform plant breeding[28].

We improved protoplasting techniques in maize (*Zea mays*). Maize is a major agricultural crop, with the total production of maize surpassing that of wheat and rice[29]. However, studies of maize biology are limited by its long generation times and our poor ability to transform it compared to other models. Model plants such as tobacco and Arabidopsis are readily transformed using agrobacterium-mediated transformation[30] and floral dip[31]. Unfortunately, these transformation techniques work poorly in maize. Instead of agrobacterium, protoplasting has been used to obtain and transform maize leaf mesophyll cells[23,32]. However, studies using this method rarely test more than a handful of different constructs in transient assays, limiting the experimental throughput.

Here, we detail a protocol to protoplast millions of maize mesophyll cells and transform them at high efficiency. The main steps are digesting the plant cell wall to release protoplasts and the subsequent transformation of the protoplasts using PEG. The critical aspect of this protocol is the large scale of the successful transformations while maintaining the cell viability required for functional assays. Transforming millions of maize protoplasts allows genome-scale experiments such as massively parallel reporter assays, that can test the function of hundreds of thousands of regions throughout the maize genome[33–35].

2.2 PROTOCOL

1 Growth of plant material

1.1 Rinse kernels twice with tap water and then soak kernels in tap water overnight at room temperature.

NOTE: For this work we used maize cultivar B73.

1.2 The next day, rinse the seeds again and plant the kernels tip cap down in 1 inch pots filled with wetted soil (1:1 vermiculite:peat moss).

1.3 Grow under long-day conditions (16 hours light, 8 hours dark) at 25° C for 3 days to germinate the seeds.

1.4 Transfer to 24 hours dark condition at 25° C and grow for 10-11 days.

2 Plasmid isolation

2.1 Transform *E. coli* with a plasmid or plasmid library using a high-efficiency transformation kit. Grow the transformed *E. coli* in LB with the appropriate antibiotics overnight at 37° C.

2.2 Harvest the *E. coli* by centrifuging for 10 minutes at 4000 g, room temperature. Extract the plasmid DNA using a commercial plasmid extraction kit.

2.3 If the plasmid concentration is less than 800 ng/µl use a vacuum concentrator to increase the concentration.

3 Protoplast isolation

3.1 Take the maize seedlings out of the dark and cut near the base. Put the stems of the seedlings into water and keep them in a dark place while not in use.

3.2 From each seedling take 10 cm sections of material from the second and third leaf making sure not to get damaged or browning parts of the leaf.

NOTE: For 20 ml of enzyme solution, cut 10-12 leaf sections of a length of 10 cm. Using too much material in the enzyme solution can decrease protoplast viability. If more protoplasts are needed, increase the volume of enzyme solution.

3.3 Stack the leaf sections on top of each other, bind them together using a binder clip, and cut the leaves perpendicular to the veins into 0.5-1 mm strips using fresh razor blades.

3.4 NOTE: We recommend cutting leaves on top of a white piece of paper and swapping razor blades whenever there is visible residue on the paper. Residue on the paper indicates mashing of the tissue due to a dull blade or poor cutting technique.

3.5 Quickly transfer the cut leaf strips into 20 ml of freshly prepared enzyme solution (Table 1) in a beaker covered in foil to block out light. Gently swirl the enzyme solution to wet the leaf strips.

NOTE: If the cut strips are not quickly put into solutions the edges will dry and reduce the number of protoplasts obtained.

3.6 Vacuum infiltrate the solution containing leaf strips at 15-20 in. Hg for 3 minutes, room temperature.

3.7 Incubate on a tabletop shaker at 40 RPM for 2.5 hours at room temperature.

NOTE: After incubation swirl the protoplasts and check to make sure that the solution is milky. If the protoplasts haven’t been released, wait up to an hour longer for digestion of the cell wall to occur. However, waiting too long can reduce the viability of the protoplasts.

3.8 Incubate on a tabletop shaker at 80 RPM for 10 minutes, room temperature.

NOTE: All solutions and protoplasts should be kept chilled on ice for the following steps until transformation. Additionally, when possible keep the protoplasts covered to prevent light exposure.

3.9 Add 20 ml of MMG (Table 1) to the enzyme solution and filter through a 40 micron cell strainer.

3.10 Split filtrate into equal volumes of no more than 10 ml each. Pour each volume into a round-bottom glass centrifuge vial and spin down for 4 minutes at 100 g, room temperature.

NOTE: When centrifuging solutions containing protoplasts, the pellets should be easy to resuspend. If you cannot resuspend such pellets by gentle swirling and flicking, then the pellet likely consists of cellular debris or damaged protoplasts.

3.11 Resuspend pellets in 1 ml of MMG each and combine samples into a single round-bottom vial. Add MMG to a total volume of 5 ml and spin down 3 minutes at 100 g, room temperature.

3.12 Wash with 5 ml of MMG and spin down 3 minutes at 100 g, room temperature. Repeat wash and spin down.

NOTE: Washing the protoplasts is a critical step to get rid of the leftover enzyme solution. After spinning down the second wash the supernatant should be almost completely clear. If it remains cloudy perform another wash. This is sometimes necessary when obtaining more than 40 million protoplasts.

3.13 Resuspend protoplasts in 1-2 ml of MMG. Take 4 µl of resuspended protoplasts and dilute 1:20 with MMG. Load the diluted protoplasts onto a hemocytometer and count the number of protoplasts obtained.

NOTE: The viability of cells at this step should be around 90% when assayed by staining with FDA (Fluorescein Diacetate). To check viability, mix FDA stock 1:1000 with solutions containing protoplasts. FDA stock is kept at 0.5% (wt/vol) FDA in acetone.

4 PEG mediated protoplast transformation

4.1 If protoplast concentration is low, centrifuge protoplasts again and resuspend in MMG to a concentration of ~ 10,000 protoplasts/µl.

4.2 For a small transformation, mix ~900,000 protoplasts with 15 µg of DNA in an Eppendorf tube. Top the mixture up with MMG up to 114.4 µl and incubate on ice for 30 minutes.

4.3 Gently resuspend the protoplasts by tapping the side of the tube. Add 105.6 µl of 25% PEG solution (Table 1) to the protoplasts to reach a final wt/vol of 12% PEG. Gently mix the PEG by inverting the Eppendorf tube 5-10 times and incubating for 10 minutes at room temperature in the dark.

NOTE: Mixing the protoplasts with the PEG solution roughly can lead to subsequent cell death.

NOTE: Transformations can be scaled up at least 20X. To transform 20 million protoplasts, use a 50 ml falcon tube, 300 µg of DNA, and scale the volumes of MMG and PEG 20X.

4.4 After incubation, dilute the transformation with five volumes of incubation solution (Table 1) and gently mix . Spin down for 4 minutes at 100 g, room temperature.

NOTE: If doing a large transformation make sure to split diluted protoplasts into aliquots of no more than 10 ml each to ensure proper pelleting. Combine protoplasts for the next wash step.

4.5 Wash with 5 ml of incubation solution and spin down for 3 minutes at 100 g, room temperature.

4.6 Resuspend transformed protoplasts in incubation solution to a concentration of 500-1000 cells/µl.

5 Protoplast incubation and transformation verification

5.1 Incubate the protoplasts in the dark overnight at room temperature to allow for transcription of the transformed plasmid or plasmid library.

5.2 In round-bottom glass vials, spin down protoplasts for 4 minutes at 100 g, room temperature, in a volume no more than 10 ml.

5.3 Resuspend and combine protoplasts into a single round-bottom vial. Wash with 5 ml of incubation solution, and spin down for 3 minutes at 100 g, room temperature.

5.4 Resuspend in 5 ml of incubation solution. Take an aliquot of the cells to check transformation efficiency under a microscope.

5.5 Spin down the protoplasts for 3 minutes at 100 g, room temperature.

5.6 If your library has a fluorescent reporter, count the percentage of transformed protoplasts using a hemocytometer and a fluorescence microscope.

5.7 If extracting RNA, resuspend cells in TRIzol.

2.3 REPRESENTATIVE RESULTS

The tissue best suited for protoplast transformation is the second and third leaves of 10-11 day old seedlings (Figure 2.1). We obtain roughly 10 million protoplasts from 16 leaf sections,

each of size 10 cm. The number of protoplasts isolated is dependent on the amount of leaf material used and the number of thin strips each section can be cut into before putting the material into the enzyme solution.

After isolating the protoplasts, we see many cells and little cell wall debris (Figure 2.2 A). Before proceeding to transforming the cells, we check the viability of our isolated protoplasts by staining with Fluorescein Diacetate (FDA). Protoplasts that glow under the GFP channel when stained with FDA are considered viable (Figure 2.2 B, C). Undamaged protoplasts tend to be round with chloroplasts speckling their surface. However, not all protoplasts that appear undamaged are still viable as seen in Figure 2.2 C. Additionally, some viable protoplasts may be in the process of dying. This can occur through evacuolation which is occurring in the protoplasts pointed to by the arrows in Figure 2.2 C. In these cells the vacuole of the protoplast is swelling and will eventually be ejected from the cell membrane.

Figure 2.3 shows protoplasts transformed with pJT01 derived from CD3-911 (ABRC), a maize transformation plasmid expressing sGFP [14]. In our derived plasmid, the GFP is tagged with a C-terminal nuclear localization signal from c-Myc. Localization of GFP to the nucleus can be seen in cells that have low and medium signals, but is drowned out by cells with particularly high fluorescence (Figure 2.3A). Using a hemocytometer, we counted fluorescent cells over multiple experiments. The percentage of cells that show fluorescence after incubation differs between biological replicates, ranging between 20-50% (Figure 2.4). If the plasmid or plasmid library doesn't have a fluorescent reporter, we recommend having a transformation control that consists of a plasmid with a fluorescent reporter to estimate transformation rates.

2.4 DISCUSSION

As seen in many other protocols, the quality of plant material used for protoplasting is essential to obtaining high quality protoplasts[36–38]. Thus, it is necessary to select healthy unblemished leaves. To obtain homogeneous mesophyll protoplasts, we prefer to not use the top 1 cm of leaves. We grow the maize seedlings used for this protocol in constant dark conditions because protoplasts isolated from etiolated leaves show better viability 16 hours after transformation compared to protoplasts from green leaves. For uses which don't require overnight incubation, using green leaf material is possible.

A step that is particularly critical to the successful digestion of cell walls is the proper cutting of leaf material into small strips before digestion. Leaf cutting must be done with sharp razor blades which are switched out immediately when they begin to dull. Cutting the leaves into many thin strips opens up more surface area for mesophyll cells to be exposed to the enzyme solution which increases the number of recovered protoplasts. Note that it is vital to make clean cuts and not crush the leaf edge to ensure high-quality viable protoplasts.

Washing the protoplasts is also a challenging step because the protoplasts are quite delicate after undergoing cell wall removal. After digestion, the protoplasts and all solutions are kept on ice and the dark-grown protoplasts are covered from light to reduce further stress. To prevent spinning down cellular debris and damaging the protoplasts, we centrifuge protoplasts at 100 g. Protoplasts centrifuged at 200 g show similar viability after isolation, but lower viability the following day and transform approximately half as well. Note that it is easier to pellet and wash protoplasts successfully when working with many protoplasts because one obtains a clearly visible pellet. For this reason, we transform 900,000 or more protoplasts at a time. At this number, we consistently obtain a visible pellet after transformation.

When scaling up the protocol, it is beneficial to centrifuge protoplasts in smaller than 10 ml aliquots to ensure that all protoplasts are pelleted and do not remain in the supernatant. For troubleshooting purposes, we recommend checking the viability of protoplasts by staining with FDA (Fluorescein diacetate) before using them for transformations. We regularly observe protoplast viability ranging between 70-90% immediately after transformation. Poor viability after isolations often leads to low transformation rates. Protoplasts with viability lower than 40% after isolation yield very few transformed cells!

Transforming and incubating the protoplasts is the last challenging step. As seen in other studies, the amount and quality of DNA are both critical factors for successful transformations[39]. We use 15 µg of DNA per million protoplasts for plasmids in the 4-6 kb range. Note that using poor quality DNA preparations can result in reduced viability of protoplasts after transformation, so we recommend using a commercial DNA extraction kit. Incubating the protoplasts overnight is done at room temperature in the dark to allow for transcription of the plasmid or plasmid library while minimizing stress of the dark-grown protoplasts. After overnight incubation, we recover approximately half the number of originally transformed protoplasts. Because of this loss, we dilute protoplasts with incubation buffer to a concentration of 500-1000 cells/µl before this step. Protoplasts left to incubate overnight at concentrations higher than 1000 cells/µl have lower recovery rates and lower viability. At higher concentrations, the debris from damaged and dying protoplasts may contribute to the death of neighboring protoplasts. Washing the protoplasts after overnight incubation serves both to remove this debris and remove excess transformed plasmid.

In summary, we have detailed a simple and scalable protocol to isolate and transform millions of high quality protoplasts from the crucial agricultural crop *Zea mays*. An important innovation of our protocol is the removal of the lengthy wash step in W5 solution that is listed in

many other protocols[35,39] that proved unnecessary in our hands. This protocol is optimized for the transformation of millions of protoplasts for use in large-scale assays such as massively-parallel reporter assays[33,34]. The ability to transform many protoplasts enables the use of genome-scale experiments in maize, allowing scientists to better understand maize gene regulation and gene expression in high-throughput.

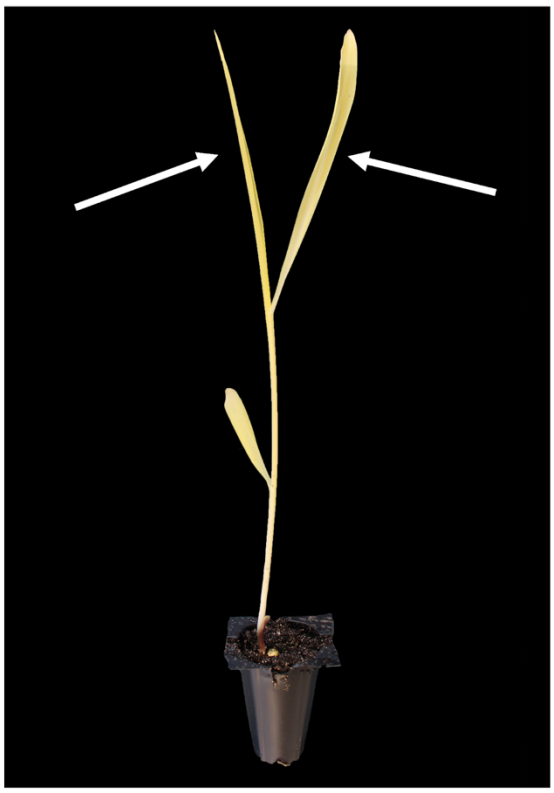


Figure 2.1.

Representative seedling of maize grown in constant dark for 10 days after germination.

Arrows point to the second and third leaves that are best suited for transformation.

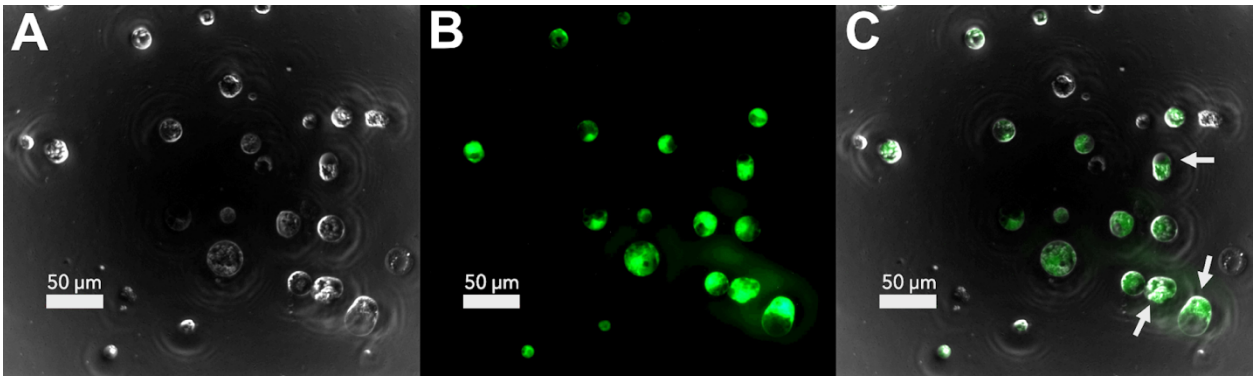


Figure 2.2.

Fluorescent microscopy images of protoplast viability after isolation.

(A) Bright field of protoplasts immediately after isolation. (B) The fluorescent signal of protoplasts stained with FDA in the GFP channel. (C) Overlap of fluorescence and bright field showing viable protoplasts. Arrows showing protoplasts undergoing evacuolation.

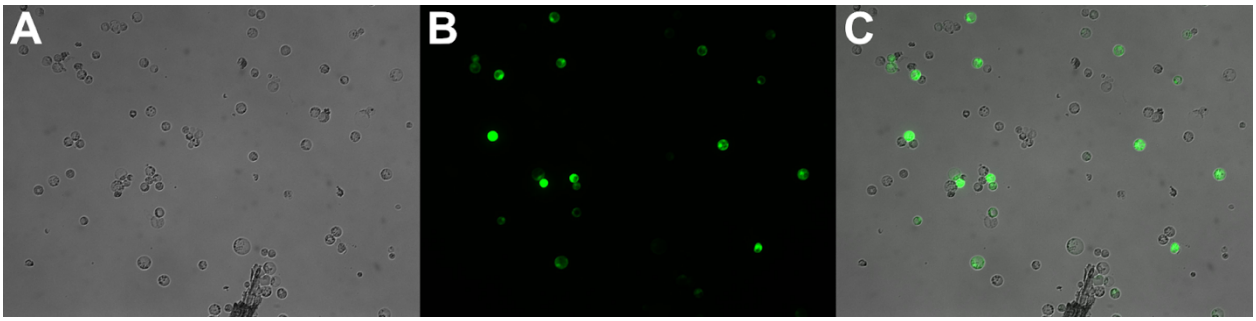


Figure 2.3.

Fluorescent microscopy images of transformed protoplasts with a nuclear localized sGFP.

(A) Bright field of the protoplasts without the fluorescence. (B) The fluorescent signal in the GFP channel of transformed cells. (C) Overlap of fluorescence and bright field image shows protoplasts that are transformed.

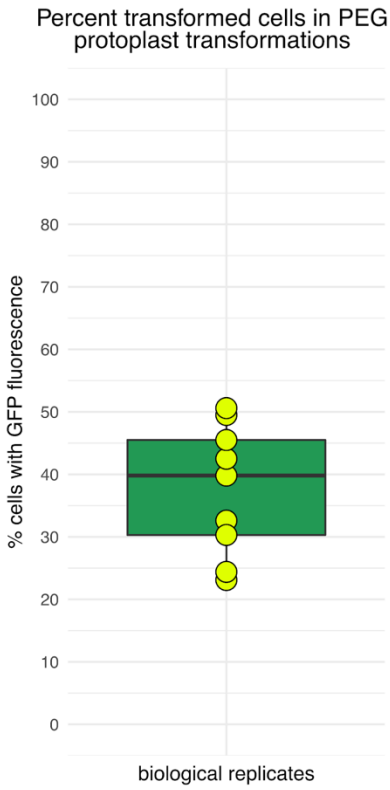


Figure 2.4.

Percentage of transformed protoplasts showing GFP fluorescence after 16 hours incubation.

Boxplot shows a median transformation rate of 40% of cells recovered after overnight incubation. Transformation rates vary between biological replicates with the lowest around 20% and the highest close to 50%. For each biological replicate, more than 200 cells were assessed for fluorescence using a fluorescence microscope.

Table 2.1. Solutions for protoplasting maize leaf mesophyll

Enzyme solution (freshly prepared)	
Mannitol	600 mM
MES (pH 5.7)	10 mM
cellulase R10	1.5% wt/vol
macerozyme	0.3% wt/vol
Heat at 55º C for 10 minutes. Then cool to room temperature before adding following components.	
CaCl2	1mM
Bovine Serum Albumin	0.1% wt/vol
Beta-mercaptoethanol	5 mM
MMg	
Mannitol	600 mM
MES (pH 5.7)	4 mM
MgCl	15 mM
Incubation solution	
Mannitol	600 mM
MES (pH 5.7)	4 mM
KCl	4 mM
25% PEG solution (freshly prepared)	
Mannitol	600 mM
CaCl2	100 mM
Poly-ethylene Glycol (MW 4000)	25% wt/vol
Put on 37º C shaker for ~30 minutes to dissolve PEG fully	
Use ultrapure water as the solvent for all solutions	

Chapter 3. IDENTIFICATION OF PLANT ENHANCERS AND THEIR CONSTITUENT ELEMENTS BY STARR-SEQ IN TOBACCO LEAVES

This work was spearheaded by Tobias Jores. I am second author on the paper for my contributions optimizing a protoplast system in crop species. I transformed maize leaf mesophyll protoplasts at high enough numbers to enable comparison between results from the maize and tobacco system. We used this comparison to show that conclusions were an unlikely to be an anomaly of the tobacco system. Additionally, we showed the ability to identify enhancers using crop species such as maize. This work is published in Plant Cell[33].

Abstract

Genetic engineering of *cis*-regulatory elements in crop plants is a promising strategy to ensure food security. However, such engineering is currently hindered by our limited knowledge of plant *cis*-regulatory elements. Here, we adapted self-transcribing active regulatory region sequencing (STARR-seq)—a technology for the high-throughput identification of enhancers—for its use in transiently transformed tobacco (*Nicotiana benthamiana*) leaves. We demonstrate that the optimal placement in the reporter construct of enhancer sequences from a plant virus, pea (*Pisum sativum*) and wheat (*Triticum aestivum*), was just upstream of a minimal promoter and that none of these four known enhancers was active in the 3′ untranslated region of the reporter gene. The optimized assay sensitively identified small DNA regions containing each of the four enhancers, including two whose activity was stimulated by light. Furthermore, we coupled the assay to saturation mutagenesis to pinpoint functional regions within an enhancer, which we recombined to create synthetic enhancers. Our results describe an approach to define enhancer

properties that can be performed in potentially any plant species or tissue transformable by *Agrobacterium* and that can use regulatory DNA derived from any plant genome.

3.1 INTRODUCTION

In a time of climate change and increasing human population, crop plants with higher yields and improved response to abiotic stresses will be required to ensure food security. As many of the beneficial traits in domesticated crops are caused by mutations in *cis*-regulatory elements, especially enhancers, genetic engineering of such elements is a promising strategy for improving crops[40,41]. However, this strategy is currently not feasible at large scale due to our limited knowledge of *cis*-regulatory elements in plants.

As in animals, plant gene expression is controlled by *cis*-regulatory elements such as minimal promoters and enhancers. A minimal promoter is the DNA sequence necessary and sufficient to define a transcription start site and recruit the basal transcription machinery. Such minimal promoters generally lead to low levels of expression[42]. Enhancers are DNA sequences that increase the basal transcription level established by minimal promoters. Enhancers serve as binding sites for transcription factors that interact with the basal transcription machinery to increase its rate of recruitment, transcription initiation, and/or elongation[42–44]. In contrast to promoters, enhancers function independently of their orientation. They can occur upstream or downstream of the minimal promoter and are active over a wide range of distances[45–47]. Enhancers can interact with minimal promoters that are several kilobases away, with such long-distance interactions assembled by chromatin loops that bring the enhancer and minimal promoter into close proximity[35,43,48,49].

Enhancers can be identified by self-transcribing active regulatory region sequencing (STARR-seq), a massively parallel reporter assay[22]. Here, candidate enhancer sequences are

inserted into the 3′ untranslated region (3′-UTR) of a reporter gene under the control of a minimal promoter. If an insert has enhancer activity, it can upregulate its own transcription. The resulting transcript can be detected by next-generation sequencing and linked to its corresponding enhancer element, which is incorporated within the mRNA. This method has been widely used in *Drosophila* and human cells[22,50–53]. In plants, STARR-seq has been described in only two studies that applied the method to the monocot species rice (*Oryza sativa*) and maize (*Zea mays*)[35,54]. These studies analyzed the enhancer activity of fragments in large genomic libraries obtained from sheared rice DNA[54] or from transposase-digested maize DNA[35]. The latter approach enriches the library for fragments from open chromatin regions of the genome where active enhancers reside[42,55]. Both these previous plant STARR-seq studies relied on species-specific protoplasts as recipient cells for the assay. However, efficient protoplasting and transformation protocols have been established for only a few species. Furthermore, protoplasts are often fragile and might not respond to external stimuli in the same way as intact plants.

Here, we established a STARR-seq assay that uses transient expression of STARR-seq libraries in tobacco (*Nicotiana benthamiana*) leaves. This assay bypasses the need for a species-specific protoplasting protocol and instead relies on efficient *Agrobacterium*-mediated transformation. Among species that are amenable to transformation with *Agrobacteria*, tobacco combines fast and robust growth with convenient transformation by syringe-infiltration of intact leaves. As transcription factors are highly conserved among plant species[56,57], the versatile tobacco system can serve as a proxy for many plant species, including crops. We optimized the placement of the enhancer candidates to provide an optimal dynamic range and performed proof-of-principle experiments to demonstrate that the assay can detect enhancers and characterize the underlying

functional elements. Furthermore, we show that our in planta assay is capable of detecting light-dependent changes of the transcriptional activity of known light-sensitive enhancers.

3.2 RESULTS

The Positioning of Enhancers Strongly Affects Their Activity in Tobacco STARR-seq

Transient expression in tobacco leaves is a well-established method for reporter assays. We tested whether STARR-seq, a massively parallel reporter assay to identify active *cis*-regulatory elements, could be performed by transient expression of libraries in tobacco. We created a reporter construct with a GFP gene under control of the cauliflower mosaic virus 35S minimal promoter and the 35S core enhancer (subdomains A1 and B1-3)[58,59]. *Agrobacterium tumefaciens* cells harboring this construct were used to transiently transform leaves of 3- to 4-week-old tobacco plants. After two d, the resulting mRNAs were extracted from the transformed leaves and analyzed by next generation sequencing (Figure 3. 1A).

To ensure a wide dynamic range of the assay, we systematically analyzed the position and orientation dependency of the 35S enhancer (Figure 3. 1A). We used a more generalized version of STARR-seq in which we placed a barcode in the GFP open reading frame. This barcode is linked to the corresponding enhancer variant by next-generation sequencing and serves as a readout for the activity of the variant. For each variant, we used five to 10 constructs with different barcodes. This barcode redundancy helps to mitigate potential effects that an individual barcode might have on the transcript level. As expected, the 35S enhancer was active in either orientation and both up- and downstream of the reporter gene (Figure 3. 1B). Similar to previous observations [58], the activity of the 35S enhancer was lower when present downstream of the gene as compared to upstream of the minimal promoter. In contrast to the mammalian system, when placed in the 3′-

UTR, the enhancer had almost no activity. Addition of a second copy of the enhancer in the "downstream" and "distal upstream" positions led on average to a 70% increase in transcript levels as compared to a single enhancer, while a second copy in the "upstream" position increased transcript levels by only 30% (Figure 3. 1B). These observations suggest that the transcriptional activation caused by a single 35S enhancer directly upstream of the minimal promoter is already close to the maximum level detectable in our assay.

We observed the strongest activation of transcription with the enhancer immediately upstream of the minimal 35S promoter and lower levels when the enhancer was placed ~1.5 kb away from the promoter as in the "downstream" and "distal upstream" constructs (Figures 3.1A and 3.1B). To characterize the distance-activity relationship, we inserted the 35S enhancer at different positions within a 2-kb spacer upstream of the minimal promoter (Figure 3.1C). Enhancer activity was strongest immediately upstream of the promoter. However, enhancer activity was greatly reduced by 500 bp or more of spacer between the enhancer and promoter (Figure 3.1C), consistent with a previously described distance-dependent decrease of 35S enhancer activity[60].

To test if the observed position dependency is unique to the 35S enhancer, we assayed three additional enhancers derived from the pea (*Pisum sativum*) *AB80* (chlorophyll a-b binding protein) and *rbcS-E9* (small subunit of ribulose-1,5-bisphosphate carboxylase) genes and the wheat (*Triticum aestivum*) *Cab-1* gene (chlorophyll a-b binding protein)[61–67]. Similar to the 35S enhancer, these enhancers were orientation independent and most active immediately upstream of the promoter, and they did not activate transcription when placed in the 3′-UTR (Figure 3. 1D).

The 35S Enhancer Is Not Active in the Transcribed Region

Although previous STARR-seq studies placed candidate enhancer fragments in the 3′-UTR[22,35,54], enhancers in this position were not active in our system. To test if the lack of enhancer

activity in the 3′-UTR is specific to our assay in transiently transformed *N. benthamiana* leaves or a more general feature of enhancers in plants, we performed STARR-seq in maize (*Z. mays* cv B73) protoplasts (Figure 3. 2A). The results with maize protoplasts were qualitatively similar to those from the assay in tobacco leaves. The 35S enhancer was most active upstream of the minimal promoter, and its activity was greatly reduced when placed in the 3′-UTR (Figure 3. 2B). Quantitatively, the activity of the 35S enhancer in the upstream position was lower in the maize protoplasts compared to that observed in tobacco leaves. However, the activity of the 35S enhancer in the 3′-UTR position was slightly higher in maize protoplasts than in tobacco leaves (compared with Figures 3.1B and 3.2B).

To explain the low activity of the 35S enhancer in the 3′-UTR, we hypothesized that such an mRNA could be degraded by nonsense-mediated decay, as long 3′-UTRs can subject mRNAs to this decay pathway[68]. To test whether the 35S enhancer in the 3′-UTR destabilizes the mRNA by promoting nonsense-mediated decay, we inserted the unstructured region from the Turnip crinkle virus 3′-UTR, shown to reduce nonsense-mediated decay[69], in between the stop codon and the enhancer. However, insertion of this region further reduced transcript levels when the 35S enhancer was placed in the 3′-UTR (Figures 3.3A and 3.3B). We next asked whether insertion of the 35S enhancer in an intron, which would also be transcribed, could confer transcriptional activation but found that it did not (Figures 3.3A and 3.3C). Furthermore, combining an upstream *AB80* enhancer with a 35S enhancer within the 3′-UTR transcribed region considerably reduced transcription compared to that from the *AB80* enhancer alone (Figure 3. 3D). Taken together, these findings demonstrate that the 35S enhancer residing within the transcribed region is not active in our system. Therefore, for subsequent experiments, we placed the enhancer fragments directly upstream of the minimal promoter, barcoding the reporter amplicons to enable detection by RNA-

seq. A similar approach with a barcode in the transcript was used in previous studies of enhancers in human cells[70,71].

The Tobacco STARR-seq Assay Can Detect Enhancer Fragments and Their Light Dependency

The *AB80*, *Cab-1*, and *rbcS-E9* enhancers are activated by light[61–63]. We tested the light dependency of these enhancers in our assay system by placing the transformed plants in the dark prior to mRNA extraction. The *AB80* and *Cab-1* enhancers demonstrated decreased activity in the dark. Although the activity of the *rbcS-E9* enhancer also showed a response to light, in this case the activity was higher in the dark (Figure 3. 4). A previous study found higher expression of Arabidopsis (*Arabidopsis thaliana*) *rbcS* genes in extracts from dark-grown plant cells compared to those from light-grown ones, with reversal of this tendency upon reconstitution of chromatin[72]. It is not clear if plant cells deposit nucleosomes onto the T-DNA harboring the reporter construct, but even if they do, nucleosome positioning and modifications might differ from those found at the endogenous loci of the enhancers.

Next, we tested if the assay could detect enhancer signatures among randomly fragmented DNA sequences from a plasmid containing embedded enhancers. We constructed a plasmid harboring the 35S, *AB80*, *Cab-1*, and *rbcS-E9* enhancers. We fragmented the plasmid using Tn5 transposase and inserted the fragments upstream of the 35S minimal promoter to generate a fragment library for use in the STARR-seq assay (Figure 3. 5A). This fragment library consisted of ~6200 fragments linked to a total of ~50,000 barcodes. About 40,000 (80%) of these barcodes were recovered with at least five counts from the extracted mRNAs. The STARR-seq assay identified the known enhancers as the regions with highest enrichment values (Figure 3. 5B). As expected, the orientation in which the fragments were cloned into the STARR-seq plasmid did not

affect their enrichment (Supplemental Figure 3.7A). This result confirms that the fragments act as enhancers instead of as autonomous promoters, whose activity would be orientation dependent. The assay was highly reproducible, with good correlation across replicates for the individual barcodes (Spearman's ρ = 0.79 to 0.82; Supplemental Figure 3.7B). The correlation further improved if the enrichment of all barcodes linked to the same fragment was aggregated (Spearman's ρ = 0.80 to 0.85; Supplemental Figure 3.7C). Replicate correlations were similar for all STARR-seq experiments in this study (Spearman's $\rho \geq 0.6$ for barcodes and ≥ 0.7 for fragments or variants).

We also used the fragment library in a STARR-seq experiment with plants kept in the dark prior to mRNA extraction to test for light dependency. We observed the expected changes in enrichment (Figure 3. 4), with the *AB80* and *Cab-1* enhancers less active and the *rbcS-E9* enhancer more active in the light-deprived plants (Figure 3.5B and 3.5C). We conclude that the STARR-seq assay established in this study can identify enhancers in a condition-specific manner.

The Tobacco STARR-seq Assay Can Pinpoint Functional Enhancer Elements

To further reveal individual elements of enhancers, we repeated the screen with a second library (5700 fragments with a total of 73,000 barcodes, more than 95% of which were recovered from the mRNA) that contained shorter fragments (median length 84 bp versus 191 bp in the initial library; Figure 3. 5D). As these shorter fragments were, on average, well below the size of the full-length enhancers, they are unlikely to contain all the elements required for maximum activity. The shorter fragments split the peaks of the *AB80* and *Cab-1* enhancers into two subpeaks, suggesting that these enhancers contain at least two independent functional elements. The sole functional element of the *rbcS-E9* enhancer resided in the 3′ half of the tested region (Figures 3.5D and 3.5E).

Having established the capacity of the assay to distinguish enhancer subdomains, we tested its suitability for conducting saturation mutagenesis of *cis*-regulatory elements. To do so, we array-synthesized all possible single nucleotide substitution, deletion, and insertion variants of the minimal promoter and of the 35S enhancer as two separate variant pools and subjected the two pools to STARR-seq. Approximately 98% of all possible variants were linked to at least one barcode in the input library, and mRNAs corresponding to over 99% of these were recovered from the tobacco leaves. We first assayed the activity of variants of a 46-bp region containing the 35S minimal promoter, in constructs with and without an enhancer. The effects of the individual mutations were similar in both contexts (Supplemental Figure 3.8). As expected, mutations that disrupt the TATA box (positions 16 to 22) had a strong negative impact on promoter activity, while most others had a weak effect or no effect (Figures 3.6A and 3.6B).

In contrast to the minimal promoter, the 35S enhancer contained several mutation-sensitive regions (Figures 3.6C and 3.6D). These regions colocalize with predicted transcription factor binding sites[73,74]. Mutations in positions 116 to 135 were especially deleterious. This region, previously implicated in enhancer activity, can be bound by the tobacco activation sequence factor 1 (ASF-1), a complex containing the basic leucine zipper (bZIP) transcription factor TGA2.2[58,59,75,76]. Similarly, we observed mutational sensitivity of the 35S enhancer in positions 95 to 115, which contain a binding site for the basic helix-loop-helix (bHLH) transcription factor complex ASF-2[77]. A third mutation-sensitive region in positions 7 to 28 is predicted to be bound by ethylene responsive factor (ERF) and teosinte branched1/cincinnata/proliferating cell factor (TCP) transcription factors.

Enhancer Fragments Can Be Combined to Build Synthetic Enhancers

To demonstrate that these mutation-sensitive regions possess enhancer activity, we split the enhancer into three fragments that span positions 1 to 30 (A), 60 to 105 (B), and 106 to 140 (C; Figure 3. 6D). These fragments were cloned in one to four copies on average, in random order, and the enhancer activity of the resulting constructs was determined. We identified 100 different constructs linked to a total of 29,000 barcodes, 95% of which were present in the extracted mRNAs. Fragments A and C alone were sufficient to activate transcription, while fragment B was active only in the presence of a second fragment (Figure 3. 6E). In line with our observations from the enhancer mutagenesis, fragment C had the highest activity. The greater the number of fragments in a construct, the higher its activity. However, even four fragments combined did not reach the level of transcription achieved with the full-length enhancer, indicating that the sequences excluded from the A, B, and C fragments contribute to enhancer activity, either directly or by providing the correct spacing for the fragments (Figure 3. 6E). Although spacing may play a role in enhancer activity, the order of the fragments had only weak effects (Supplemental Figure 3.9). Taken together, we demonstrate that this assay can identify functional enhancer elements that can be recombined to create synthetic enhancers of varying strength.

3.3 DISCUSSION

In this study, we developed a massively parallel reporter assay in tobacco plants that can identify DNA regions with enhancer or promoter activity and can dissect these regions to characterize functional sequences with single-nucleotide resolution. The assay does not depend on efficient protoplasting and transformation protocols, which have been established only for a limited number of species and tissues. Furthermore, in contrast to protoplasts, the in planta system is more robust and can be exposed to a variety of environmental conditions to detect condition-specific *cis*-regulatory elements. Indeed, our tobacco STARR-seq assay can detect enhancer light

dependency. Such condition-specific *cis*-regulatory elements could play important roles in future genetic engineering efforts to help plants adapt to a rapidly changing environment.

We observed in our experiments that the tested enhancers were not active when placed in the transcribed region. Other studies have shown that plant genes can contain elements in their transcribed region, especially in the first intron, that drastically increase their expression[78–83]. However, the increased expression levels could have been the result of enhanced transcription or translation, improved mRNA processing, export, or stability, or a combination of these mechanisms. Few studies have dissected these potential mechanisms, and these have generally found that enhanced transcription played no role, or only a relatively small role, in the overall expression increase[79,81,82]. The apparent absence of strong transcriptional enhancers in the transcribed region of plant genes could be due to any of several reasons. The constraints placed on such regions to enable efficient mRNA processing and translation might not be compatible with the requirements for enhancers. Alternatively, strong binding of transcription factors within the transcribed region could inhibit transcription by physically blocking the RNA polymerase. Future studies will be required to address this issue in plants.

Comparing the activity of the 35S enhancer in transiently transformed tobacco leaves to its activity in maize protoplasts, a general trend of high activity upstream of the minimal promoter and low activity within the 3′-UTR was observed, but the levels differed between the two systems. Previous studies have reported that the 35S promoter constructs encompassing the 35S minimal promoter and enhancer are more active in dicots like tobacco than monocots such as maize and rice[84,85]. In agreement with these studies, we detected higher activity of the 35S enhancer upstream of the minimal promoter in tobacco compared to maize. By contrast, the maize system led to more 35S enhancer activity than the tobacco system when the enhancer was inserted into the 3′-UTR, a

possible effect of species-specific differences in the tolerance of an enhancer within the transcribed region. Consistent with these species differences, effects of intron-mediated enhancement of gene expression are stronger in monocots than dicots[81]. Alternatively, the physical state of the reporter construct-containing DNA could influence enhancer activity. In maize protoplasts, the reporter is expressed from a supercoiled plasmid, whereas it resides on linear T-DNA molecules in the transiently transformed tobacco cells. Linear and supercoiled DNA probably display different looping behavior that could influence medium- to long-range enhancer-promoter interactions.

In the two previous plant STARR-seq studies[35,54], the enhancer candidates were placed in the 3′-UTR of the reporter gene. While these studies successfully identified several strong enhancers, our results indicate that the dynamic range in these studies might have been improved by altering the placement of the enhancer. The reporter design we used here with enhancer candidates immediately upstream of the minimal promoter yields a high signal-to-noise ratio and enables confident discovery of intermediate and weak enhancers. This design requires an additional sequencing step to link the enhancer candidates to the corresponding barcodes. However, the barcodes are short and of the same length, whereas the length of enhancer candidates can vary considerably. Cloning these highly variable enhancer sequences into the 3′-UTR may have a profound impact on the stability of the resulting mRNAs and how readily they can be recovered and sequenced.

Apart from the enhancer placement, the plant species (tobacco, rice, or maize), and the recipient cells/tissue (protoplasts or intact leaves) differ between this study and the two previous plant STARR-seq studies[35,54]. As 35S enhancer activity in tobacco leaves and maize protoplasts showed similar trends, the choice of a model system for future STARR-seq studies will likely depend on what mode of transformation is most efficient in the target species. While several

species are amenable to transient transformation with *Agrobacteria*[86–89], others might be more easily transformed as protoplasts[90–92]. Transformation efficiency will pose a limit to the maximum library size that can be screened. The two previous plant STARR-seq studies transformed 15 to 30 million protoplasts[35,54]. The largest library in this study contained ~73,000 barcodes, of which we detected more than 95% in the extracted mRNAs. In experiments with a larger library, we could recover 250,000 fragments from a single tobacco leaf. As the extraction of mRNAs from 100 tobacco leaves can be performed in a single day, library sizes similar to the ones used in the other plant STARR-seq studies should be feasible.

Due to the widespread conservation of transcription factors in the plant lineage[56,57], the enhancer elements identified in tobacco leaves will likely be active in many other plant species. Furthermore, the STARR-seq assay described herein can potentially be performed in any species or tissue that can be transiently transformed by *Agrobacteria*. Apart from enhancers and promoters, the assay can likely be adapted to screen for silencers and insulators—*cis*-regulatory elements that are known from animals but have, so far, not been detected in plants.

Taken together, we describe a plant STARR-seq assay that is applicable to enhancer screens for any plant species to analyze plant gene regulation and to identify promising building blocks for future genetic engineering efforts. The data generated by these screens and subsequent saturation mutagenesis will enable deep learning approaches to identify defining characteristics of plant enhancers.

3.4 METHODS

Plasmid Construction and Library Creation

The STARR-seq plasmids used herein are based on the pGreen plasmid[93]. In their T-DNA region, they harbor a phosphinothricin resistance gene (BlpR) and the GFP reporter construct

terminated by the poly(A) site of the Arabidopsis (*Arabidopsis thaliana*) ribulose bisphosphate carboxylase small chain 1A gene. These plasmids were deposited at Addgene (Addgene no. 149416–149422, https://www.addgene.org/). The 35S minimal promoter followed by the synthetic 5′ UTR synJ[94] (ACACGCTGGAATTCTAGTATACTAAACC), an ATG start codon and a 15-bp random barcode (VNNVNNVNNVNNVNN) was cloned in front of the second codon of GFP by Golden Gate cloning[95]. Enhancers or DNA fragments were inserted by Golden Gate cloning into the indicated positions. The 2-kb spacer used in Figure 3. 1C was derived from enCas9 in pEvolvR-enCas9-PolI3M-TBD[96] (Addgene no. 113077; https://www.addgene.org/113077/). For constructs with full-length enhancers (Figures 3.1B to 3.1D, 3.2, 3.3, and 3.4), 5 to 10 uniquely barcoded variants were used. These enhancers were inserted into the SacI, XbaI, XhoI, and SfoI sites of the plasmid pZS*11-yfp0[97](Addgene no. 53241, https://www.addgene.org/53241/). The resulting plasmid (pZS*11_4enh, Addgene no. 149423, https://www.addgene.org/149423/) was fragmented with Nextera Tn5 transposase (Illumina), and the fragments were amplified with primers containing adapters suitable for Golden Gate cloning. The single-nucleotide variants of the 35S promoter and enhancer were ordered as an oligonucleotide array from Twist Bioscience (Figures 3.6A to 3.6D). The libraries were bottlenecked to ~10 barcodes per variant. Fragments A, B, and C (Figure 3.6D) were ordered as oligonucleotide with 5′-GTGATG overhangs, mixed with Golden Gate cloning adaptors and ligated with T4 DNA ligase. The IV2 intron was inserted into GFP at position 103. The unstructured region of the Turnip crinkle virus 3′-UTR (TACGGTAATAGT GTAGTCTTCTCATCTTAGTAGTTAGCTCTCTCTTATATT) was inserted after the GFP stop codon. Next-generation sequencing on an Illumina NextSeq platform was used to link the inserted fragments and barcodes. The STARR-seq plasmid libraries were introduced into *Agrobacterium tumefaciens* GV3101 strain harboring the helper plasmid pSoup[93] by electroporation.

Plant Cultivation and Transformation

Tobacco (*Nicotiana benthamiana*) was grown in soil (Sunshine Mix no. 4) at 25°C in a long-day photoperiod (16 h light and 8 h dark; cool-white fluorescent lights [Philips TL-D 58W/840]; intensity 300 μmol/m^2/s). Plants were transformed 3 to 4 weeks after germination. For transformation, an overnight culture of *A. tumefaciens* was diluted into 50 mL YEP medium (1% [w/v] yeast extract, 2% [w/v] peptone) and grown at 28°C to an OD of ~1. A 5-mL input sample of the cells was taken, and plasmids were isolated from it. The remaining cells were harvested and resuspended in 50 mL induction medium (M9 medium supplemented with 1% [w/v] Glc, 10 mM MES, pH 5.2, 100 μM $CaCl_2$, 2 mM $MgSO_4$, and 100 μM acetosyringone). After overnight growth, the bacteria were harvested, resuspended in infiltration solution (10 mM MES, pH 5.2, 10 mM $MgCl_2$, 150 μM acetosyringone, and 5 μM lipoic acid) to an OD of 1 and infiltrated into the first two mature leaves of two to four tobacco plants. The plants were further grown for 48 h under normal conditions or in the dark prior to mRNA extraction.

Maize Protoplast Generation and Transformation

We used a slightly modified version of a published protoplasting and electroporation protocol[90]. Maize (*Zea mays* cv B73) seeds were germinated for 4 d in the light, and the seedlings were grown in soil at 25°C in the dark for 9 d. The center 8 to 10 cm of the second leaf from 7 to 9 plants was cut into thin strips perpendicular to the veins and immediately submerged in 10 mL protoplasting solution (0.6 M mannitol, 10 mM MES, 15 mg/mL cellulase R-10 [GoldBio], 3 mg/mL Macerozyme R-10 [GoldBio], 1 mM $CaCl_2$, 5 mM β-mercaptoethanol, pH 5.7). The mixture was covered to keep out light, vacuum infiltrated for 30 min, and incubated with 40 rpm shaking for 2 h. Protoplasts were released with 80 rpm shaking for 5 min and filtered through a 40 μm filter. The protoplasts were harvested by centrifugation (3 min at 200*g*, room temperature) in

a round-bottom glass tube and washed with 3 mL electroporation solution (0.6 M mannitol, 4 mM MES, 20 mM KCl, pH 5.7). After centrifugation (2 min at 200*g*, room temperature), the cells were resuspended in 3 mL ice cold electroporation solution and counted. Approximately 1 million cells were mixed with 25 µg plasmid DNA in a total volume of 300 µL, transferred to a 4-mm electroporation cuvette and incubated for 5 min on ice. The cells were electroporated (300 V, 25 µFD, 400 Ω) and 900 µL incubation buffer (0.6 M mannitol, 4 mM MES, 4 mM KCL, pH 5.7) was added. After 10 min incubation on ice, the cells were further diluted with 1.2 mL incubation buffer and kept at 25°C in the dark for 16 h before mRNA collection.

STARR-seq Assay

For each STARR-seq experiment with tobacco plants, at least three independent biological replicates were performed. Different plants and fresh *Agrobacterium* cultures were used for each biological replicate, and the replicates were performed on different days. Depending on the library size, two samples of two to three leaves were collected from a total of two to four plants. They were frozen in liquid nitrogen, ground in a mortar, and immediately resuspended in 5 mL TRIzol (Thermo Fisher Scientific). The suspension was cleared by centrifugation (5 min at 4000*g*, 4°C), and the supernatant was thoroughly mixed with 2 mL chloroform. After centrifugation (15 min at 4000 × g, 4°C), the upper, aqueous phase was transferred to a new tube, mixed with 1 mL chloroform and centrifuged again (15 min at 4000*g*, 4°C). Then, 2.4 mL of the upper, aqueous phase was transferred to new tubes, and RNA was precipitated with 240 µL 8 M LiCl and 6 mL 100% (v/v) ethanol by incubation at −80°C for 15 min. The RNA was pelleted (30 min at 4000*g*, 4°C), washed with 2 mL 70% (v/v) ethanol, centrifuged again (5 min at 4000*g*, 4°C), and resuspended in 500 µL nuclease-free water. mRNAs were isolated from this solution using 100 µL magnetic Oligo$(dT)_{25}$ beads (Thermo Fisher Scientific) according to the manufacturer's protocol.

The mRNAs were eluted in 40 µL. The two samples per library were pooled and supplemented with DNase I buffer, 10 mM $MnCl_2$, 2 µL DNase I (Thermo Fisher Scientific), and 1 µL RNaseOUT (Thermo Fisher Scientific). After 1 h incubation at 37°C, 2 µL 20 mg/mL glycogen (Thermo Fisher Scientific), 10 µL 8 M LiCl, and 250 µL 100% (v/v) ethanol were added to the samples. Following precipitation at −80°C, centrifugation (30 min at 20,000*g*, 4°C), and washing with 200 µL 70% (v/v) ethanol (5 min at 20,000*g*, 4°C), the pellet was resuspended in 100 µL nuclease-free water. Eight reactions with 5 µL mRNA each and a GFP construct-specific primer (GAACTTGTGGCCGTTTACG) were prepared for cDNA synthesis using SuperScript IV reverse transcriptase (Thermo Fisher Scientific) according to the manufacturer's instructions. Half of the reactions were used as no reverse transcription control, in which the enzyme was replaced with water. After cDNA synthesis, the reactions were pooled and purified with DNA Clean and Concentrator-5 columns (Zymo Research). The barcode region was amplified with 10 to 20 cycles of PCR and read out by next-generation sequencing on an Illumina NextSeq platform.

For the STARR-seq assay in maize protoplasts, we performed three independent biological replicates on different days with different plants. Transformed protoplasts were harvested by centrifugation (3 min at 200*g*, 4°C) 16 h after electroporation. The protoplasts were washed three times with 1 mL incubation buffer and centrifuged for 2 min at 200*g* and 4°C. The cells were resuspended in 300 µL TRIzol (Thermo Fisher Scientific) and incubated for 5 min at room temperature. The suspension was thoroughly mixed with 60 µL chloroform and centrifuged (15 min at 20,000*g*, 4°C). The upper, aqueous phase was transferred to a new tube, mixed with 60 µL chloroform, and centrifuged again (15 min at 20,000*g*, 4°C). RNA was precipitated from 200 µL of the supernatant with 1 µL 20 mg/mL glycogen (Thermo Fisher Scientific), 20 µL 8 M LiCl, and 600 µL 100% (v/v) ethanol by incubation at −80°C for 15 min. After centrifugation (30 min at

20,000*g*, 4°C), the pellet was washed with 200 µL 70% (v/v) ethanol, centrifuged again (5 min at 20,000*g*, 4°C), and resuspended in 200 µL nuclease-free water. mRNAs were isolated from this solution using 50 µL magnetic Oligo$(dT)_{25}$ beads (Thermo Fisher Scientific) according to the manufacturer's protocol, and the mRNAs were eluted in 40 µL water. DNase I treatment and precipitation were performed as for the mRNAs obtained from tobacco plants but with half the volume. Reverse transcription, purification, PCR amplification, and sequencing were performed as for the tobacco samples. For the maize protoplast STARR-seq, the plasmid library used for electroporation was sequenced as the input sample.

Computational Methods

Binding site motifs for *N. benthamiana* transcription factors were obtained from the PlantTFDB[74], and Find Individual Motif Occurrences[73] (FIMO) was used to predict their occurrence in the 35S core enhancer. Fragments of pZS*11_4enh were aligned to the reference sequence using Bowtie2[98]. For analysis of the STARR-seq experiments, the reads for each barcode were counted in the input and cDNA samples. Barcode counts below 5 and barcodes present in only one of three replicates were discarded. Barcode enrichment was calculated by dividing the barcode frequency (barcode counts divided by all counts) in the cDNA sample by that in the input sample. For the pZS*11_4enh fragment library (Figure 3.5; Supplemental Figure 3.7) and the mutagenesis (Figures 3.6A to 3.6D) experiments, the enrichment of the fragments or variants was calculated as the median enrichment of all barcodes linked to them. Boxplots were created using all corresponding barcodes from all replicates performed and were normalized to the median enrichment of constructs without an enhancer. The enrichment coverage of pZS*11_4enh was calculated by summing up the enrichment of all fragments containing a given nucleotide and dividing this sum by the number of fragments. Nucleotides covered by fewer than five fragments

were excluded from analysis. Light dependency of enhancers or enhancer fragments was calculated as $\log_2$ of the enrichment in the light condition divided by the enrichment from the dark condition. Spearman correlations were calculated using the base R function. The code used for analyses is available at https://github.com/tobjores/tobacco-STARR-seq.

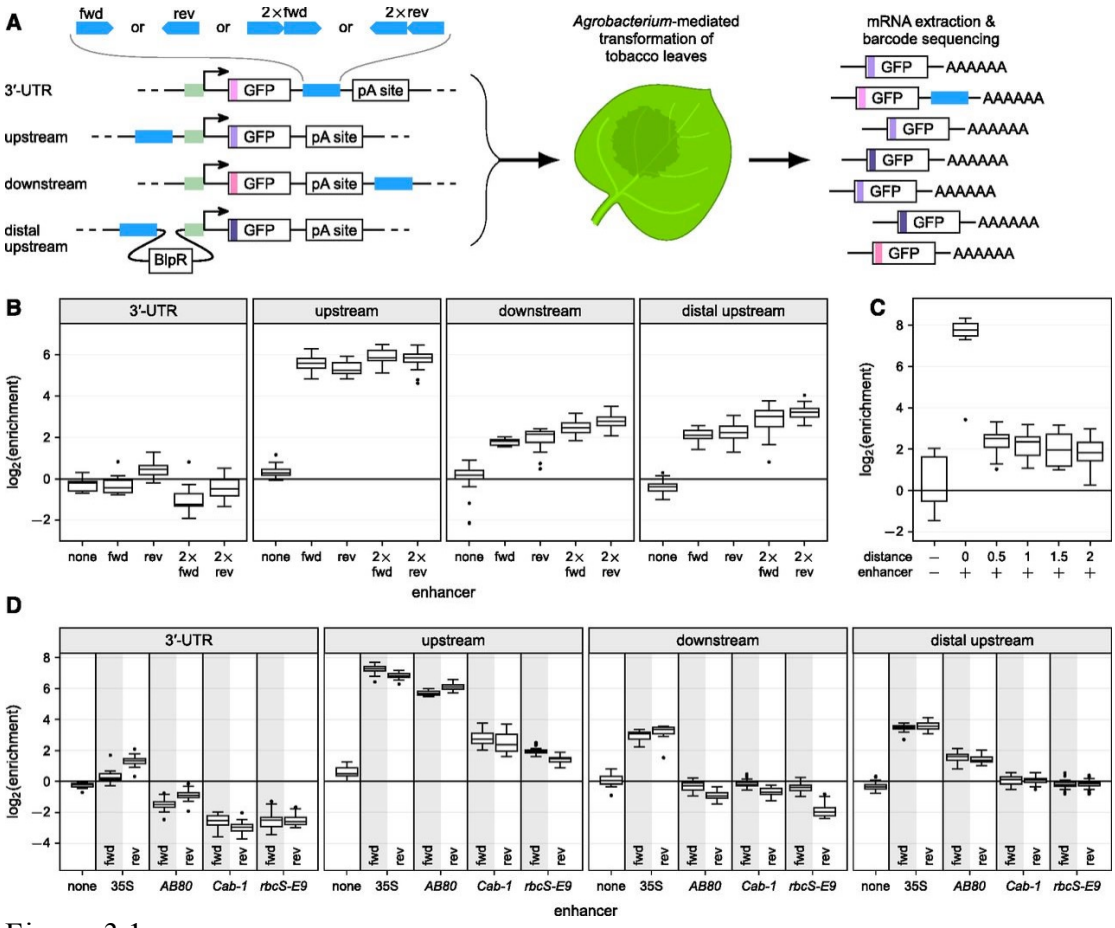


Figure 3.1.

The Positioning of Enhancers Has a Pronounced Impact on Their Activity.

(A) Scheme of the tobacco STARR-seq assay. All constructs are driven by the 35S minimal promoter (green). Enhancers (blue) are inserted in the indicated orientation and position. Barcodes (shades of purple) are inserted in the GFP open reading frame. BlpR, phosphinothricin resistance gene; pA site, poly-adenylation site. **(B)** STARR-seq was performed with constructs harboring a single or double (2×) 35S enhancer in the indicated positions and orientations. Plots show $\log_2$(enrichment) of recovered RNA barcodes compared to DNA input. Each boxplot (center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) represents all barcodes from three independent replicates combined, as do subsequent ones. **(C)** After introduction of a 2-kb spacer upstream of the minimal promoter, the 35S enhancer was

inserted at the indicated distance upstream of the minimal promoter, and STARR-seq was performed. **(D)** The 35S and three known plant enhancers were introduced in either the forward (fwd) or reverse (rev) orientation at the indicated positions and the STARR-seq assay was performed.

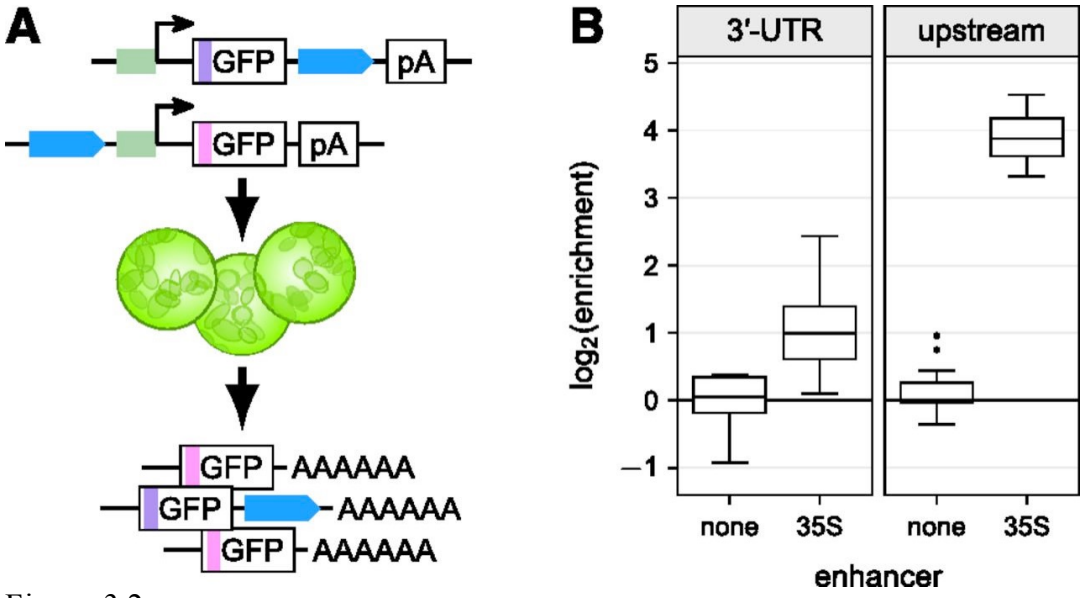


Figure 3.2.

The 35S Enhancer Is Most Active Upstream of the Minimal Promoter in Maize Protoplasts.

(A) Scheme of the maize protoplast STARR-seq assay. Colors and symbols are as in Figure 1A. **(B)** Maize protoplasts were transformed with constructs without an enhancer (none) or with the 35S enhancer (35S) in the indicated position and subjected to the STARR-seq assay. Boxplots were created as in Figure 1B.

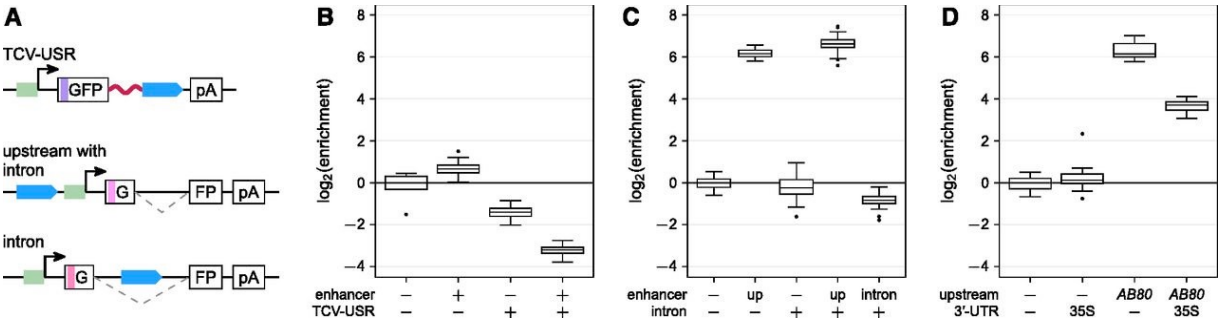


Figure 3.3.

The 35S Enhancer Is Not Active in the Transcribed Region.

(A) Scheme of the constructs used in this figure. All constructs contain a 35S minimal promoter (green) controlling expression of a GFP reporter gene with or without an intron. In some constructs, an unstructured region of the Turnip crinkle virus 3′-UTR (TCV-USR) was inserted after the stop codon. The 35S core enhancer (blue) was inserted into the indicated positions. **(B)** Constructs with the TCV-USR inserted after the stop codon were subjected to the STARR-seq assay. **(C)** STARR-seq with constructs harboring an intron in the GFP open reading frame. The 35S enhancer was inserted upstream (up) of the promoter or into the intron (intron). **(D)** STARR-seq using constructs with the *AB80* enhancer upstream of the minimal promoter and/or the 35S enhancer in the 3′-UTR. Boxplots in **(B)** to **(D)** were created as in Figure 1B.

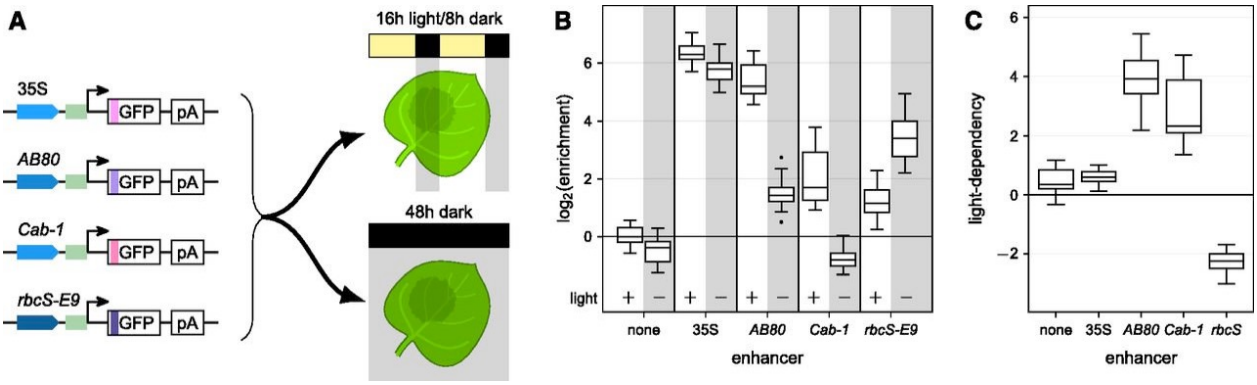


Figure 3.4.

STARR-seq Can Detect Light-Dependency of Plant Enhancers.

(A) Tobacco leaves were infiltrated with reporter constructs driven by the indicated enhancers. The plants were then grown for 2 d in normal light/dark cycles or completely in the dark prior to mRNA extraction. Colors and symbols are as in Figure 1A. **(B)** The activity of the indicated enhancers was determined after growth in normal light/dark cycles (+ light) or in the dark (− light). **(C)** Light-dependency ($\log_2$[enrichment^{light}/enrichment^{dark}]) was determined for the indicated enhancers. Boxplots in **(B)** and **(C)** were created as in Figure 1B.

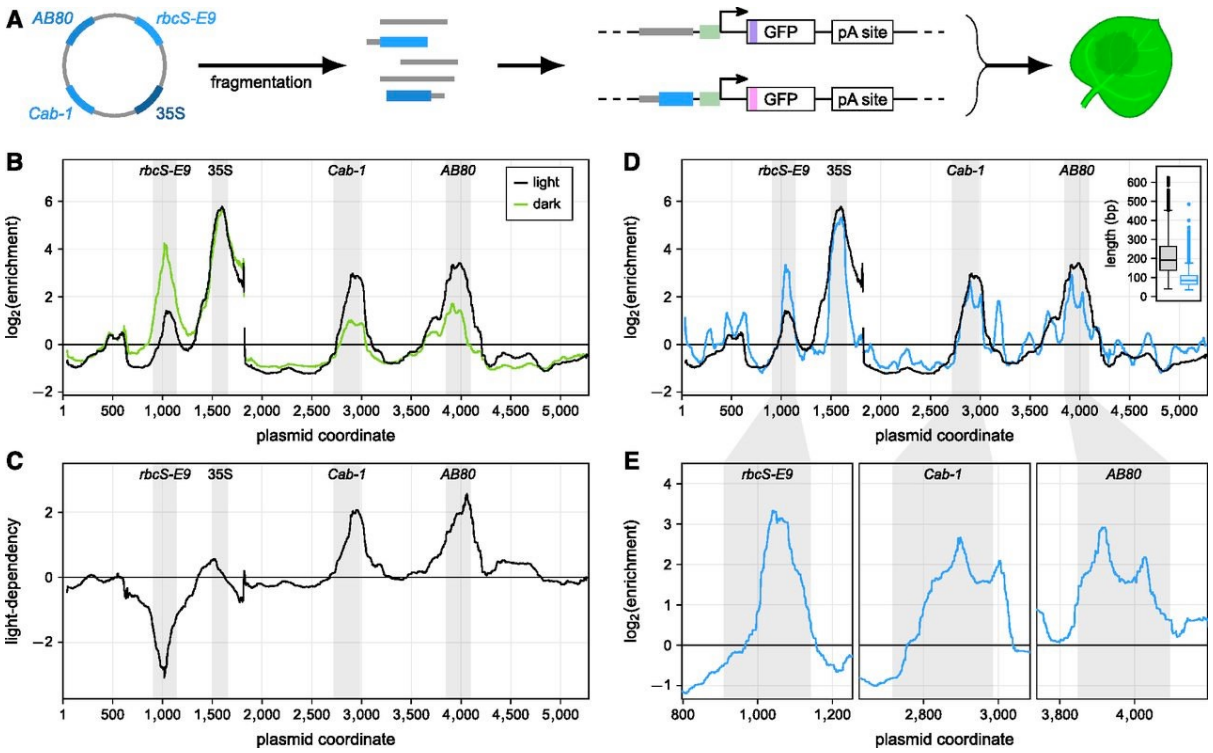


Figure 3.5.

The Tobacco STARR-seq Assay Identifies Condition-Specific Enhancer Fragments.

(A) A plasmid harboring the indicated enhancers was fragmented. The fragments were inserted in the upstream position of the STARR-seq construct and their activity was measured by the STARR-seq assay. Colors and symbols are as in Figure 1A. **(B)** Plants were grown for 2 d in normal light/dark cycles (light, black line) or completely in the dark (dark, blue line) prior to mRNA extraction. The log2(enrichment) of RNA expression over input of all fragments at each position was averaged. **(C)** Light-dependency ($\log_2[\text{enrichment}^{\text{light}}/\text{enrichment}^{\text{dark}}]$) was determined for each base of the original plasmid. **(D)** The STARR-seq assay was performed with plasmid fragment libraries with different fragment length distributions (see inset; boxplot—center line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers), and $\log_2$(enrichment) for each fragment library is shown across the whole plasmid. **(E)**

Log_2(enrichment) obtained from the library with shorter fragments is shown in more detail for regions of interest. Positions in the original plasmid that contain enhancers are shaded in gray.

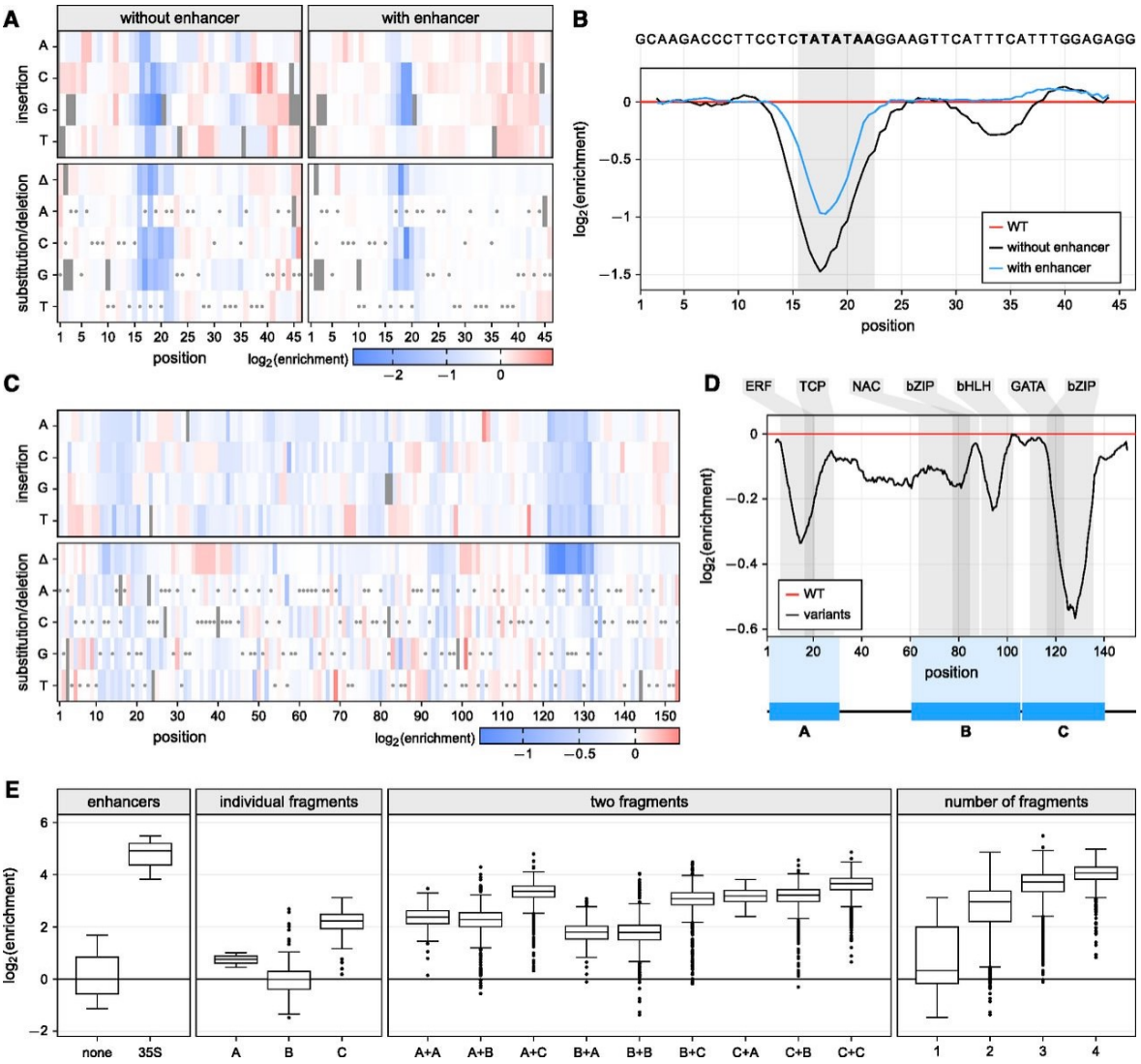
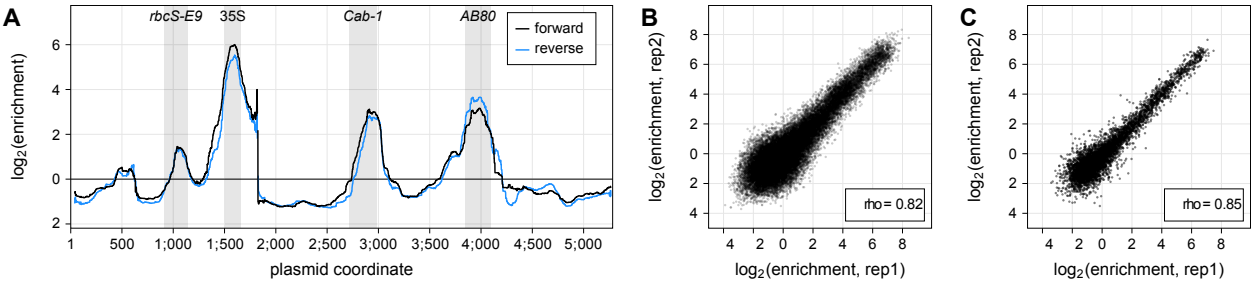


Figure 3.6.

Saturation Mutagenesis Identifies Functional Elements in the 35S Promoter and Enhancer.

(A) All possible single nucleotide variants of the 35S minimal promoter were inserted into constructs with or without the 35S enhancer. The enrichment of the individual promoter insertions, substitutions, and deletions was measured by the STARR-seq assay, normalized to the wild-type variant, and plotted as a heatmap. Missing values are shown in black and wild-type variants are marked with a gray dot. **(B)** A sliding average (window size = 4 bp) of the positional mean enrichment scores for all substitutions, insertions, and deletions was determined. The TATA box is highlighted in gray. **(C)** The enrichment of all possible single nucleotide insertions,

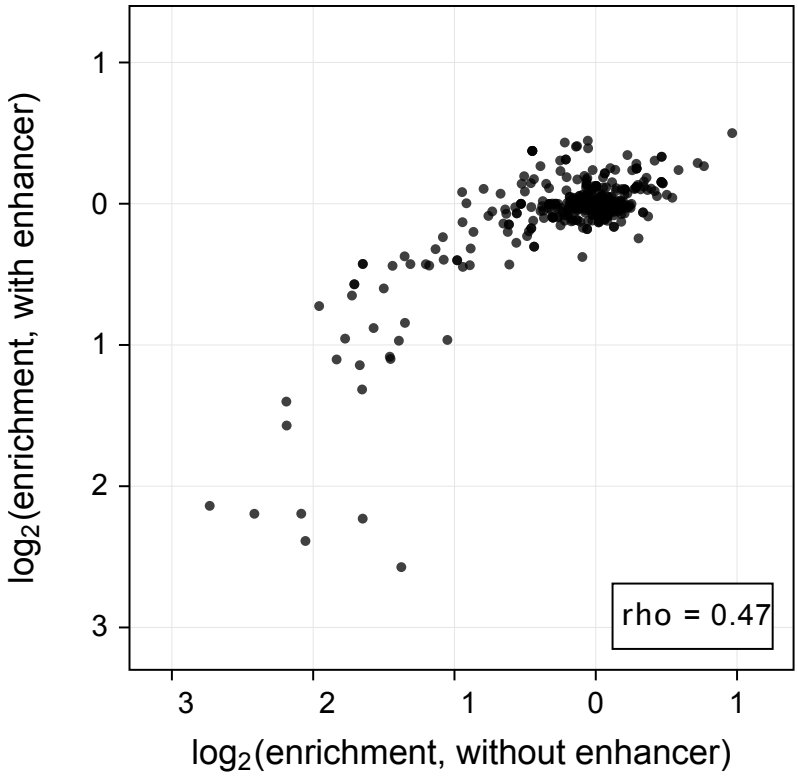
substitutions, and deletions of the 35S core enhancer was determined as in **(A)**. **(D)** A sliding average (window size = 8 bp) of the positional mean enrichment was determined via the STARR-seq assay. Predicted binding sites for the transcription factors from the indicated families are highlighted in gray. **(E)** Three fragments (A, B, C; see **[D]**) of the 35S enhancer were inserted into the STARR-seq plasmid in random number and order and assayed for their enhancer activity. Boxplots were created as in Figure 1B.



Supplemental Figure 3.7.

Activity in the STARR-seq assay is insensitive to orientation and reproducible.

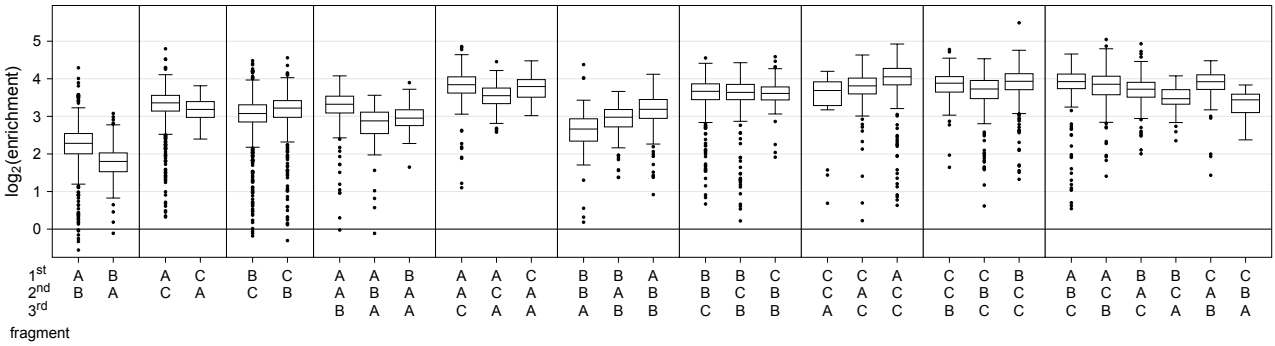
(Supports Figure 5) **(A)** The data from the STARR-seq assay (see Fig. 5B, light condition) was analyzed separately for fragments inserted in the forward or reverse orientation. **(B)** Correlation (Spearman's rho) between two (out of three) replicates for individual barcodes. **(C)** Correlation (Spearman's rho) between two (out of three) replicates for fragments (median enrichment of all linked barcodes).



Supplemental Figure 3.8.

Activity of the promoter variants is correlated with and without an enhancer in the construct.

(Supports Figure 6) The enrichment scores for 35S minimal promoter variants (see Fig. 6A) in constructs with or without an upstream 35S enhancer was compared. The correlation (Spearman's rho) is indicated.



Supplemental Figure 3.9.

The order of 35S enhancer fragments has a subtle influence on enhancer activity.

(Supports Figure 6) Three fragments (A, B, C) of the 35S enhancer (see Fig. 6D) were inserted into the STARR-seq plasmid in random number and order, and assayed for their enhancer activity. Each boxplot (center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers) represents all barcodes from three independent replicates combined.

Chapter 4. SYNTHETIC PROMOTER DESIGNS ENABLED BY A COMPREHENSIVE ANALYSIS OF PLANT CORE PROMOTERS

This work was spearheaded by Tobias Jores. I am second author on the paper for my contributions further improving the maize mesophyll protoplasts system to test hundreds of thousands of constructs for effect on transcription. We used this system to compare core promoter function in the C3 tobacco system and the C4 maize system. This work is published in Nature Plants[34].

Abstract

Targeted engineering of plant gene expression holds great promise for ensuring food security and for producing biopharmaceuticals in plants. However, this engineering requires thorough knowledge of *cis*-regulatory elements to precisely control either endogenous or introduced genes. To generate this knowledge, we used a massively parallel reporter assay to measure the activity of nearly complete sets of promoters from *Arabidopsis*, maize and sorghum. We demonstrate that core promoter elements—notably the TATA box—as well as promoter GC content and promoter-proximal transcription factor binding sites influence promoter strength. By performing the experiments in two assay systems, leaves of the dicot tobacco and protoplasts of the monocot maize, we detect species-specific differences in the contributions of GC content and transcription factors to promoter strength. Using these observations, we built computational models to predict promoter strength in both assay systems, allowing us to design highly active

promoters comparable in activity to the viral 35S minimal promoter. Our results establish a promising experimental approach to optimize native promoter elements and generate synthetic ones with desirable features.

4.1 INTRODUCTION

Precise control of gene expression is necessary to generate transgenic plants with new properties, such as growth in formerly incompatible environments or production of medically or nutritionally important products[13,99]. Much of this control occurs at the initiation of transcription, the first committed step in gene expression. Transcription initiation involves the recruitment of the basal transcription machinery, comprised of general transcription factors (TFs) and RNA polymerase, to core promoters. Core promoters define the transcription start site (TSS) but their activity typically leads to only low levels of expression[42,100]. This basal level of transcription is increased by the interaction of core promoters with enhancers, which can reside upstream or downstream of the TSS and over a wide range of distances from the promoter[35,45,46].

The first core promoter element identified was the TATA box. This motif, with the consensus sequence TATA(A/T)A(A/T), is recognized by the TATA-binding protein, a subunit of TFIID, and plays an important role in recruiting the basal transcription machinery and in determining the TSS location[100–102]. Since then, several other core promoter elements have been discovered in viral and animal promoters[101,103–110]. In plants, short motifs composed of pyrimidine bases, termed the TC motif or Y patch, have been described as potential plant-specific core promoter elements[111–113].

Apart from these elements, promoters also contain binding sites for TFs close to the TSS. In contrast to the core promoter elements, which often occur at specific distances from, and in a fixed orientation to, the TSS, the TF-binding sites can be functional in either orientation and their

activity is less constrained by their distance to the TSS. Promoter-proximal TF-binding sites can influence the transcriptional output from the nearby TSS and, in some cases, influence where transcription starts[114]. In this study, we refer to the region surrounding the TSS that harbours core promoter elements as the core promoter; the extended region that includes the core promoter and upstream TF-binding sites is referred to as the promoter.

To gain a better understanding of the regulatory principles governing promoter activity, several high-throughput studies have been performed in yeast, *Drosophila melanogaster* and human cells[115–122]. These studies validated the contribution of core promoter elements and promoter-proximal TF-binding sites to overall promoter activity and deduced rules governing the interaction among those elements. However, it is not clear whether these rules also apply to plant promoters. Although computational analyses have revealed that many of the core promoter elements identified in animals are enriched in plant promoters[111,112,123,124], only the TATA box and the Initiator (Inr) element have been functionally validated[33,125–127]. Some plant promoters do not harbour any of the known core promoter elements[123]. A recent study built synthetic plant promoters by combining TF-binding sites[128]. However, to date, large-scale functional studies have not been performed with plant core promoters.

A deeper understanding of the regulatory code of plant promoters and how it shapes transcription levels will further our knowledge of gene regulation, empower the controlled manipulation of gene expression for crop improvement and enable the rational design of promoters for use in genetic engineering. Here, we set out to comprehensively analyse the core promoters of the model plant *Arabidopsis thaliana* and the important crop maize (*Zea mays*) and its close relative sorghum (*Sorghum bicolor*). The genome of the crucifer *Arabidopsis* is compact (~135 megabases (Mb)) and AT-rich, while the genomes of the cereals maize and sorghum are

GC-rich and many times larger (~2.7 gigabases and ~730 Mb, respectively). We sought to determine how these differences in genome content and architecture would be reflected in features of their promoter elements. Here, we identified key determinants of core promoter strength and characterized similarities and differences in the regulatory code of monocotyledonous and dicotyledonous plants. Using this knowledge, we designed synthetic core promoters with activities reaching levels comparable to that of the 35S minimal promoter. Furthermore, we trained computational models that accurately predict promoter strength in our assays and help improve promoter activity.

4.2 RESULTS

Use of the STARR-seq assay to study plant core promoters

We used the self-transcribing active regulatory region sequencing (STARR-seq) assay, which we had established in plants[33], to measure the strength of nearly complete sets of core promoters from *Arabidopsis*, maize and sorghum. Specifically, for each species, we interrogated the sequences from −165 to +5 relative to the annotated TSS for protein-coding and microRNA (miRNA) genes. These 170-bp regions were tested for promoter strength by using them to drive expression of a barcoded green fluorescent protein (GFP) reporter gene (Fig. 4.1a). We included the first five bases after the TSS to cover core promoter elements that span the TSS, like the Inr, while avoiding substantial parts of the 5′ untranslated region (UTR). The 5′ UTRs affect messenger RNA levels posttranscriptionally and hence their inclusion could confound assessment of promoter strength[129]. Instead, we used the 5′ UTR of a sorghum histone H3 gene (SORBI_3010G047100) for all sorghum promoters and the 5′ UTR of a maize histone H3.2 gene (Zm00001d041672) for all maize and *Arabidopsis* promoters (the 5′ UTR of the *Arabidopsis* histone H3.1 gene

AT5G10390 had intrinsic promoter activity). We constructed three STARR-seq libraries that contained 18,329 *Arabidopsis*, 34,415 maize and 27,094 sorghum core promoters linked to ~400,000 unique barcodes per library. To test these promoters for their response to a strong enhancer, we also generated each library using a plasmid containing the cauliflower mosaic virus 35S enhancer[58,59] immediately upstream of the promoter insertion site[33]. The six libraries were assayed individually in transiently transformed tobacco leaves and maize protoplasts.

In each promoter library, we included two control constructs, one containing only the viral 35S minimal promoter (−46 to +5 relative to the TSS) and the other containing the 35S minimal promoter and enhancer (−199 to −47 relative to the TSS). The promoter strength for each tested plant promoter was normalized to the control construct containing only the 35S minimal promoter. The construct also containing the strong 35S enhancer upstream of the minimal promoter was used to test the dynamic range of the assay. Consistent with previous reports[33,84], the 35S enhancer was fourfold more active in the tobacco system than in maize protoplasts (Fig. 4.1b). We performed two biological replicates for each promoter library in each assay system. The replicates were highly correlated, especially for the libraries with the 35S enhancer, which reflected their generally higher promoter strength (Fig. 4.1c,d and Supplementary Fig. 4.19). Therefore, we used the average promoter strength from both replicates for all further analyses. We validated these results by retesting a subset of 166 and 173 promoters in two separate libraries, obtaining results that were highly correlated with the data from the comprehensive promoter libraries (Supplementary Fig. 4.20). Since the sorghum promoters were coupled to a sorghum 5′ UTR in the comprehensive library and to a maize 5′ UTR in the validation libraries, the high correlation between these datasets suggests that the two 5′ UTRs did not strongly affect promoter strength.

Promoter strengths as measured in the tobacco leaf system had a weak to intermediate (R^2 of 0.14–0.40) correlation with those obtained from maize protoplasts (Fig. 4.1e and Supplementary Fig. 4.19c,f), indicating that there are substantial differences in how the two systems interact with the core promoters. Irrespective of the assay system, the promoters spanned a wide range of activity, with >250-fold difference between the strongest and weakest promoters (Fig. 4.2a,b). Few promoters were stronger than the viral 35S minimal promoter, which is probably optimized for maximal activity. Overall, the promoters of the dicot *Arabidopsis* tended to perform better in the dicot tobacco system, while the promoters of the monocots maize and sorghum showed greater activity in protoplasts of the monocot maize (Fig. 4.2a,b).

Gene ontology (GO)-term enrichment analysis showed that the genes corresponding to the most active promoters in our assay were significantly (adjusted $P \leq 0.05$) enriched for components of nucleosomes, which are highly expressed housekeeping genes (Fig. 4.2c). In both systems, strong promoters often were also associated with genes annotated for response to stress and function in the extracellular region, including genes encoding defence and cell wall proteins. In the maize protoplast system, genes associated with strong promoters frequently encoded proteins with oxidoreductase activity or unfolded protein-binding functions. The latter is consistent with reports of wound-induced reactive oxygen species and a heatshock response in protoplasts[130]. Although these results show a qualitative agreement between core promoter strength and expression level for some genes, there was no substantial correlation overall between promoter strength and expression data[131–133] for the corresponding genes in planta (Supplementary Fig. 4.9). This lack of correlation is expected, as core promoters represent only a subset of all the regulatory elements that drive gene expression and other elements such as enhancers can drastically affect transcription rates in the genomic context.

Next, we asked if genes of different types use different promoters. The activity of miRNA promoters was indistinguishable from that of promoters of protein-coding genes (Fig. 4.2e,d). However, promoters from genes with an annotated 5′ UTR were generally stronger than those of genes without a 5′ UTR annotation. As the TSSs of the latter are probably not correctly annotated, these sequences are probably not true promoters, explaining their low activity.

Multiple sequence features influence promoter strength

Monocot genomes are more GC-rich than dicot genomes[123,134] and this bias holds true for their core promoter sequences (Fig. 4.3a). In the tobacco leaf system, GC content strongly affected promoter strength, with AT-rich promoters up to fourfold more active than GC-rich ones (Fig. 4.3b). A high GC content was especially detrimental close to the 5′ end of the promoters but was better tolerated towards the 3′ end (Fig. 4.3c). In contrast, in maize protoplasts, GC content was not predictive of promoter strength (Fig. 4.3d). Since the GC content of the *Arabidopsis* and tobacco genomes is similar[135], the transcriptional machinery in tobacco is probably tuned to AT-rich promoters and works less well with the GC-rich promoters of maize and sorghum. Conversely, the transcription machinery of maize commonly acts on GC-rich promoters and can effectively use them in protoplasts. The correlation between promoter strength and GC content is, therefore, a characteristic of the assay system and not an intrinsic feature of the promoters.

We next tested how known core promoter elements affect promoter strength. Considering first the location of TATA box motifs, we noticed marked differences among the promoters of *Arabidopsis*, maize and sorghum. In *Arabidopsis* promoters, the distribution of TATA boxes had a peak ~30 bp upstream of the TSS (Fig. 4.4a). Although this location also is common for maize promoters, the maize promoters showed two additional peaks for the TATA box at: ~55 and ~70 bp

upstream of the TSS. In sorghum promoters, the TATA box distribution peaked at ~40 bp upstream of the TSS, with a shoulder ~30 bp upstream of the TSS.

Core promoters harbouring a TATA box were up to fourfold stronger than TATA-less ones, especially when the TATA box is located within the region from 23 to 59 bp upstream of the TSS, where most TATA boxes in the promoters of *Arabidopsis*, maize and sorghum reside (Fig. 4.4a–c). The location of the TATA box in maize promoters affected their strength only in maize protoplasts. In this assay system, maize promoters with a TATA box in one of the three peaks of the TATA box distribution were stronger than those with a TATA box elsewhere. Furthermore, maize promoters with a TATA box in the peak closest to the TSS were strongest and they became successively weaker in the other two peaks as the TATA box is located increasingly more TSS-distal (Supplementary Fig. 4.10). The effect of the TATA box on promoter strength was not a consequence of an increased AT-content in the promoters containing a TATA box. (Supplementary Fig. 4.21). To directly measure the effect of the TATA box, we mutated this motif in native promoters. Replacement of one or both T nucleotides in the core TATA motif with a G resulted in decreased transcriptional activity (Fig. 4.4d,e). Similarly, promoter strength was increased when a canonical TATA box was inserted into a TATA-less promoter; a mutated version of the TATA box did not have this effect (Fig. 4.4f,g).

In animal promoters, the TATA box is often surrounded by the upstream (BRE^u) and/or downstream (BRE^d) TFIIB recognition element. Mutational studies have demonstrated that these elements can modulate promoter strength[106,109]. In tobacco leaves, neither of the two elements had a strong effect on promoter activity; however, in maize protoplasts, BRE^u was associated with 25% increased, and BRE^d with 10% decreased, promoter strength (Supplementary Fig. 4.11a–d). Consistent with these results, mutations that inactivate BRE^u decreased promoter strength in maize

protoplasts but not in tobacco leaves. Inserting a canonical BRE^u led to increased promoter activity, especially in maize protoplasts. In contrast, mutating or inserting BRE^d had only modest effects on promoter activity in both assay systems (Supplementary Fig. 4.11e–h). A valine residue in the helix-turn-helix motif of the general transcription factor TFIIB is crucial for the recognition of BRE^u in animals[106,136]. Although this residue is not conserved in any plant TFIIB protein, the maize genome encodes an additional TFIIB-related protein with a valine at the corresponding position (Supplementary Fig. 4.22). The presence of this maize-specific TFIIB-related protein may explain the increased activity of BRE^u in the maize protoplast system.

Computational analyses of plant promoters[111–113] have detected an enrichment of short, pyrimidine-rich motifs upstream of the TSS (Supplementary Fig. 4.12a). Because such an enrichment was not detected in animal promoters, these motifs, termed Y patches, were proposed to be plant-specific core promoter elements. Our data support this hypothesis, as Y patch-containing promoters showed 10–15% greater strength compared to those without the element (Supplementary Fig. 4.12b,c).

Consistent with previous studies[125,127], we observed that promoters with an Inr at the TSS were generally stronger than those without it. In contrast, the polypyrimidine initiator TCT, previously described in animals[110], was less effective (Supplementary Fig. 4.13).

Finally, we asked whether promoter-proximal TF-binding sites affect promoter strength. We first clustered TFs by similarity of their binding site motifs and created a consensus motif for each of the 72 clusters. We then compared the strength of promoters with a predicted binding site to that of promoters lacking it. About 67% of the TF clusters did not have a significant impact on promoter strength. However, 23 TF motifs were significantly ($P \leq 0.0005$) associated with altered promoter strength in at least one assay system. For example, the TCP TF motif tends to reside in

promoters that were strong in tobacco leaves, while this effect was not observed in maize protoplasts (Supplementary Fig. 4.14a,b). On the other hand, promoters with a motif for heatshock factors (HSFs) were stronger than those without it in maize protoplasts but not in tobacco leaves (Supplementary Fig. 4.14c,d).

We asked whether core promoter elements and TF-binding sites are spatially constrained in relation to one another. In contrast to core promoter elements, most TF-binding sites did not show a preferential position relative to the TSS. However, we observed that TF-binding sites upstream of the TATA box were generally associated with a higher promoter strength compared to those downstream of the TATA box (Supplementary Fig. 4.15). Since RNA polymerase is recruited to the region downstream of the TATA box, this enzyme may displace TFs bound here and thereby prevent them from activating transcription.

Promoters show varying degrees of enhancer responsiveness

In animals, promoters can interact differentially with enhancers[118,137]. Similarly, the 35S enhancer activated some plant core promoters more than others. However, the presence of the 35S enhancer resulted in increased transcription from almost all core promoters, up to 60-fold for the most responsive promoters in the tobacco leaf system and up to 15-fold in maize protoplasts; the 35S enhancer is less active in maize protoplasts[33,84]. Consistent with the notion that enhancers are the drivers of tissue- and condition-specific transcription[42,59], promoters of genes with high tissue specificity (top third of the genes as ranked by the tissue-specificity index τ; ref. [138]) showed on average 33% increased enhancer responsiveness compared to promoters of genes with low tissue specificity (bottom third of the τ distribution) (Fig. 4.5a,b). Similarly, promoters of miRNA genes, which are often differentially expressed in response to environmental or developmental cues, were

33% more responsive to the 35S enhancer than promoters of protein-coding genes (Supplementary Fig. 4.23).

To understand which promoter features influence enhancer responsiveness, we analysed the elements that affect promoter strength. Promoters with a TATA box were up to 67% more responsive to the 35S enhancer than TATA-less promoters; however, the location of the TATA box did not have a consistent impact on enhancer responsiveness (Fig. 4.5c,d). Furthermore, promoter GC content influenced enhancer responsiveness in the tobacco leaf system but not in maize protoplasts (Fig. 4.5e,f). While the GC content and TATA box had a similar effect on enhancer responsiveness as on promoter strength, the same was not true for TFs. Instead, TFs that increased promoter strength often reduced enhancer responsiveness (Supplementary Fig. 4.16a–d), potentially due to competition for a limited pool of TFs or because of incompatibilities between recruited downstream factors. In contrast, some TFs that did not influence promoter strength affected enhancer responsiveness (Supplementary Fig. 4.16e,f). The effects on enhancer responsiveness possibly reflect synergistic effects, whereby the core transcriptional machinery and the TFs at promoters and enhancers interact with one another.

Core promoter strength can be modulated by light

The plant STARR-seq assay can identify light-responsive enhancers[33]. To test whether core promoters that respond to light can also be identified, we subjected the promoter libraries to STARR-seq experiments in tobacco leaves that were kept in the light (16 h light, 8 h dark) for 2 d after transformation (Fig. 4.6a). We did not perform the same experiment with maize protoplasts, as known light-responsive enhancers were not active in this system (Supplementary Fig. 4.24). As expected, most promoters did not respond to the light. However, about 2,400 promoters were at

least four times more active in the light or in the dark (Fig. 4.6b). The genes associated with the most highly light-dependent promoters were enriched for those encoding plastid proteins, especially for proteins in thylakoids, the membrane-bound chloroplast compartments that are the site of the light-dependent reactions of photosynthesis (Fig. 4.6c).

While promoters that are AT-rich were more light-dependent than GC-rich ones (Fig. 4.6d), the effects of GC content on light-dependency were much less pronounced than on promoter strength and enhancer responsiveness. Similarly, the presence of a TATA box showed weaker and even inconsistent effects on light-dependency compared to TATA box effects on promoter strength and enhancer responsiveness (Fig. 4.6d). We found that the light-dependency of a promoter was mainly determined by the TF-binding sites it contains. The presence of the TCP-binding site, for example, led to increased expression in the light (Fig. 4.6e) and, consistent with previous studies[139], the presence of the WRKY-binding site led to repressed expression in the light (Fig. 4.6f). These trends were confirmed by mutational analysis. Mutations that disrupt a binding site for WRKY TFs increased the light-dependency of the promoter, while mutations that disrupt a binding site for TCP TFs led to a noticeable, albeit not significant, decrease in light-dependency (Supplementary Fig. 4.17).

Design of synthetic plant promoters

After identifying key features of native plant promoters, we sought to use these features in the design of synthetic promoters. We started by generating random sequences with nucleotide frequencies resembling either an average *Arabidopsis* or average maize promoter (Fig. 4.7a). We designed ten sequences each for the two nucleotide frequencies; however, due to their AT-rich nature, the synthesis of approximately half of the sequences with an *Arabidopsis* promoter-like

base composition failed. Consistent with the findings for native promoters, the synthetic promoters with low GC content, similar to that of *Arabidopsis* promoters, were 30% more active in tobacco leaves than those with GC content similar to that of maize promoters (Fig. 4.7b,c). However, as expected, these random synthetic promoters were weak. To increase their activity, we modified them by adding an Inr, Y patch element or TATA box (Fig. 4.7a). Although all three of these core promoter elements, both alone and in combination, increased promoter strength, the TATA box showed the strongest effect and the Inr the weakest (Fig. 4.7b,c). The relative activity of these three elements was similar across synthetic promoters with initial nucleotide frequencies similar to either *Arabidopsis* or maize and across the two assay systems. However, in tobacco leaves, the absolute change in promoter strength was different for synthetic promoters of different GC content, indicating that the elements tested in this assay system require a favourable sequence environment to achieve full activity (Fig. 4.7b). Taken together, the results demonstrate that it is possible to rationally design synthetic core promoters of varying strength by choosing an appropriate background nucleotide frequency and adding canonical core promoter elements. The strongest synthetic promoters reached activities comparable to the viral 35S minimal promoter.

We also used the synthetic promoters to further analyse the effect of promoter-proximal TF-binding sites. We focused on four different binding sites: two sites for TCP TFs and one each for HSF TFs and NAC TFs. The TF-binding sites were introduced at three positions in the synthetic promoters in which a TATA box had been added (Fig. 4.7d). Because we did not observe position-dependent differences for any of the three TF-binding sites, we grouped their respective data to perform the subsequent analyses. Consistent with our observations for native promoters, the TCP-binding sites had the strongest effect in tobacco leaves, the HSF sites were most active in maize protoplasts and the NAC sites had a weak but consistent effect across both assay systems (Fig.

4.7e). When more than one TF-binding site was introduced into the synthetic promoters, their activities were additive and the relative strengths of the promoters were conserved in combinations. The more binding sites that were present, the higher the promoter strength (Fig. 4.7f, Supplementary Fig 4.25).

Finally, to test whether the TFs show position-dependent activity with regard to the TATA box, the binding sites for TCP, HSF and NAC TFs were inserted at several positions upstream and downstream of the TATA box. While these TF-binding sites at all tested positions upstream of the TATA box led to similar increases in promoter strength, they did not increase promoter strength when inserted downstream of the TATA box (Fig. 4.7g,h, Supplementary Fig. 4.26). These results probably reflect competition with the core transcriptional machinery that binds to this region.

Computational models predict and improve promoter strength

Computational models have been used to optimize synthetic gene-regulatory sequences[122,140]. Therefore, we set out to develop predictive models for core promoter strength using the data from the libraries with the 35S enhancer to train the models, as they had a better replicate correlation. For each assay system, we trained a separate model using 90% of the promoters, with the remaining 10% used to validate the model. We initially used a linear regression model for this task. The GC content and the maximum score for a match to the position weight matrices for the core promoter elements and TF clusters of each sequence were used as input features. The linear models explained 51% and 45% of the variability in promoter strength in tobacco leaves and maize protoplasts, respectively (Fig. 4.8a). In both systems, the TATA box score was the most important feature for promoter strength, followed by GC content.

To obtain models with increased predictive power, we turned to a machine learning approach using a convolutional neural network (CNN). The models used the DNA sequence of the core promoters as input and predicted the strength of the promoters in the test set, resulting in an R^2 of 0.71 and 0.67 for the tobacco and the maize systems, respectively (Fig. 4.8b).

We used these models for in silico evolution of 150 native promoters with weak, intermediate or strong activity in our assay. Additionally, we subjected the synthetic promoters with or without various core promoter elements to evolution. For each promoter, we generated every possible single nucleotide substitution variant and scored these variants with the CNN models. The best variant was retained and subjected to another round of evolution. We synthesized the starting sequences and those obtained after three and ten rounds of evolution and experimentally determined their activity. As predicted, we observed a large increase in promoter strength after three rounds of evolution and another, albeit less pronounced, increase after ten rounds (Fig. 4.8c,d and Supplementary Fig. 4.18). We obtained the best results when the evolution was performed with the CNN model trained on data from the same assay system. However, when we used a combination of both models to score the promoter variants, we could generate promoters with high activities in both systems that were on par with those evolved with the CNN model that was trained on data from the system in which the evolved sequences were tested (Fig. 4.8c–f). The models used for the in silico evolution were trained on data from libraries with an upstream 35S enhancer; however, when we tested the evolved promoters without the 35S enhancer, their activities followed the same trend, with a large increase in activity after three rounds and an additional increase after ten (Fig. 4.8e,f). These results suggest that the increased promoter strength generated by the evolution process was not enhancer-dependent and that these promoters might similarly work well with other enhancers.

4.3 DISCUSSION

The use of plants to synthesize medical and nutritional products requires precise control of foreign genes; similarly, precise control of endogenous genes is required to generate plants that can better withstand stresses. This precision can be realized through the design of synthetic promoters with optimal sequences, spacings and orientations of regulatory elements. Here, we used the STARR-seq assay to characterize plant core promoters in depth. We demonstrate that the most critical element of a strong plant core promoter is the presence of a TATA box ~30–40 bp upstream of the TSS. The next most critical element is a nucleotide composition appropriate for the plant that is being engineered. A promoter can further be improved with an Inr motif at the TSS and a pyrimidine-rich region between the TATA box and the Inr. Such rationally designed promoters can reach activities comparable to the highly active viral 35S minimal promoter.

While it might be optimal to conduct these experiments within the genomic context in planta, current technologies make such large-scale studies feasible only with transient expression of reporter constructs. However, the lack of genomic context may be less important for promoter strength than is commonly assumed. Studies in human and *Drosophila* cells found that results from plasmid-based regulatory elements are highly correlated with those from genome-integrated ones in massively parallel reporter assays[22,141]. Moreover, human core promoters retain their relative strength regardless of where they are inserted in the genome or if they drive expression of a plasmid-encoded reporter; the genomic context appears merely to scale their activity but does so independently of promoter identity[142]. Furthermore, we and others have previously demonstrated that transient STARR-seq assays in plants recapitulate the relative strength and the condition-specificity of known regulatory elements[33,35]. Our findings about the relative strength of promoters should, therefore, apply to promoters integrated in the genome, with the caveat that nearby

enhancers may modulate the absolute expression level in addition to tissue- and condition-specificity.

Promoter activity and conditional response can be further modified by the addition of TF-binding sites upstream of the TATA box. Such binding sites affected promoter strength in an additive manner. The choice of binding site, however, will depend on the assay system and on the TFs that are present and active in it. TF presence and activity cannot simply be inferred from TF motifs because plant TF families are large and often encode both activating and repressing factors with highly similar binding preferences. However, single-cell genomics can determine which TFs are expressed in specific cell types and associated with chromatin accessibility of regulatory elements[143–145]. This knowledge offers a promising avenue to explore the activity of cell type-specific regulatory elements. In the absence of an assay system derived from a cognate cell type, cell type-specific TFs can be co-expressed in the assay systems used here. Alternatively, a large array of promoters can be designed with an assortment of TF-binding sites, followed by an assay like the one described here to identify the most active ones.

Nevertheless, the design of strong core promoters appears feasible without such cell type-specific or even species-specific data. Our CNN models accurately predicted promoter strength and could be used for in silico evolution to yield native and synthetic promoters with increased activity. Moreover, a combination of CNN models trained on data from the tobacco and maize assay systems yielded promoters active in both systems. Such promoters are robust candidates to use across a broad range of tissues and species and in conjunction with multiple enhancers.

In animals, enhancer–promoter interactions are fine-tuned to execute distinct regulatory programmes, like expression of housekeeping or developmental genes[118,137]. Here, we studied the

effect of only the viral 35S enhancer on plant promoters. However, this assay could be applied to study interactions between promoters and native plant enhancers; such experiments might reveal specific interactions between distinct types of promoters and enhancers. Combining the potent core promoters characterized here with equally well-characterized enhancers will add the desired condition-specific and cell type-specific regulation needed for applications in plant engineering and biotechnology.

4.4 METHODS

Library design and construction

For this study, we used the sequence from −165 to +5 relative to the annotated TSS as core promoters. We used the Araport11 annotation[146] for *A. thaliana* Col-0 and the NCBI_v3.43 annotation[147] for *S. bicolor* BTx623. For *Z. mays* L. cultivar B73 promoters, we used experimentally determined TSSs[148] and supplemented this set with the B73_RefGen_v4.42 annotation[149] for genes without an experimentally confirmed TSS. The core promoter sequences were ordered as an oligo pool from Twist Biosciences.

The STARR-seq plasmids used herein are based on the plasmid pPSup[33] (https://www.addgene.org/149416/). It harbours a phosphinothricin resistance gene (BlpR) and a GFP reporter construct terminated by the polyA site of the *A. thaliana* ribulose bisphosphate carboxylase small-chain 1A gene in the transfer DNA region. The plant core promoters followed by a 5′ UTR from maize (Zm00001d041672; used for the *Arabidopsis*, maize and validation promoter libraries) or sorghum (SORBI_3010G047100; used the sorghum promoter library) histone H3 gene, an ATG start codon and a 12-bp random barcode (VNNVNNVNNVNN; V = A, C or G) was cloned in front of the second codon of GFP by Golden Gate cloning[95]. For control constructs, the 35S minimal promoter was used instead of the plant core promoters. Each library

was bottlenecked to contain, on average, 10–20 barcodes per promoter. The 35S core was inserted upstream of the core promoters by Golden Gate cloning. The STARR-seq plasmid libraries were introduced into *Agrobacterium tumefaciens* GV3101 strain harbouring the helper plasmid pSoup[93] by electroporation.

Tobacco cultivation and transformation

Tobacco (*Nicotiana benthamiana*) was grown in soil (Sunshine Mix no. 4) at 25 °C in a long-day photoperiod (16 h light and 8 h dark; cool-white fluorescent lights (Philips TL-D 58 W/840; intensity 300 μmol m^{-2} s^{-1}). Plants were transformed 3–4 weeks after germination. For transformation, an overnight culture of *A. tumefaciens* was diluted into 100 ml of YEP medium (1% (w/v) yeast extract, 2% (w/v) peptone) and grown at 28 °C to an optical density (OD) of ~1. A 5-ml input sample of the cells was taken and plasmids were isolated from it. The remaining cells were harvested and resuspended in 100 ml of induction medium (M9 medium supplemented with 1% (w/v) glucose, 10 mM MES, pH 5.2, 100 μM $CaCl_2$, 2 mM $MgSO_4$ and 100 μM acetosyringone). After overnight growth, the bacteria were harvested, resuspended in infiltration solution (10 mM MES, pH 5.2, 10 mM $MgCl_2$, 150 μM acetosyringone and 5 μM lipoic acid) to an OD of 1 and infiltrated into the first two mature leaves of three to six tobacco plants. The plants were further grown for 48 h under normal conditions or in the dark before mRNA extraction.

Maize protoplast generation and transformation

We used a slightly modified version of a published protoplasting and electroporation protocol[90]. Maize (*Z. mays* L. cultivar B73) seeds were germinated for 4 d in the light and the seedlings were grown in soil at 25 °C in the dark for 9 d. The centre 8–10 cm of the second leaf

from ten to 12 plants were cut into thin strips perpendicular to the veins and immediately submerged in 10 ml of protoplasting solution (0.6 M mannitol, 10 mM MES, 15 mg ml^{-1} cellulase R-10 (GoldBio), 3 mg ml^{-1} Macerozyme R-10 (GoldBio), 1 mM $CaCl_2$, 5 mM β-mercaptoethanol, 0.1% (w/v) BSA, pH 5.7). The mixture was covered to keep out light, vacuum infiltrated for 30 min and incubated with 40 r.p.m. shaking for 2 h. Protoplasts were released with 80 r.p.m. shaking for 5 min and filtered through a 40 μm filter. The protoplasts were harvested by centrifugation (3 min at 200*g*, room temperature) in a round-bottom glass tube and washed with 3 ml of ice-cold electroporation solution (0.6 M mannitol, 4 mM MES, 20 mM KCl, pH 5.7). After centrifugation (2 min at 200*g*, room temperature), the cells were resuspended in 3 ml of ice-cold electroporation solution and counted. Approximately one million cells were mixed with 25 μg of plasmid DNA in a total volume of 300 μl, transferred to a 4-mm electroporation cuvette and incubated for 5 min on ice. The cells were electroporated (300 V, 25 μFD, 400 Ω) and 900 μl of ice-cold incubation buffer (0.6 M mannitol, 4 mM MES, 4 mM KCL, pH 5.7) was added. After 10 min of incubation on ice, the cells were further diluted with 1.2 ml of incubation buffer and kept at 25 °C in the dark for 16 h before mRNA collection. To cover each library, four electroporation reactions were performed, except for the smaller validation libraries in which two electroporation reactions were performed. For the maize protoplast STARR-seq, the plasmid library used for electroporation was sequenced as the input sample.

STARR-seq assay

For each STARR-seq experiment, two independent biological replicates were performed. Different plants and fresh *Agrobacterium* cultures were used for each biological replicate and the replicates were performed on different days. For experiments in tobacco, 12 transformed leaves

were collected from six plants. They were frozen in liquid nitrogen, ground in a mortar and immediately resuspended in 25 ml of TRIzol (Thermo Fisher Scientific). The suspension was cleared by centrifugation (5 min at 4,000*g*, 4 °C) and the supernatant was thoroughly mixed with 5 ml of chloroform. After centrifugation (15 min at 4,000*g*, 4 °C), the upper, aqueous phase was transferred to a new tube, mixed with 5 ml of chloroform and centrifuged again (15 min at 4,000*g*, 4 °C). Then 13 ml of the upper, aqueous phase was transferred to new tubes and RNA was precipitated with 1.3 ml of 8 M LiCl and 32.5 ml of 100% (v/v) ethanol by incubation at –80 °C for 15 min. The RNA was pelleted (30 min at 4,000*g*, 4 °C), washed with 10 ml of 70% (v/v) ethanol, centrifuged again (5 min at 4,000*g*, 4 °C) and resuspended in 1.5 ml of nuclease-free water. The solution was split into two halves and mRNAs were isolated from each using 150 µl of magnetic Oligo$(dT)_{25}$ beads (NEB) according to the manufacturer's protocol. The mRNAs were eluted in 40 µl. The two samples per library were pooled and supplemented with 10 µl of DNase I buffer, 10 µl of 100 mM $MnCl_2$, 2 µl of DNase I (Thermo Fisher Scientific) and 1 µl of RNaseOUT (Thermo Fisher Scientific). After 1 h incubation at 37 °C, 2 µl of 20 mg ml^{-1} glycogen (Thermo Fisher Scientific), 10 µl of 8 M LiCl and 250 µl of 100% (v/v) ethanol were added to the samples. Following precipitation at –80 °C, centrifugation (30 min at 20,000*g*, 4 °C) and washing with 200 µl of 70% (v/v) ethanol (5 min at 20,000*g*, 4 °C), the pellet was resuspended in 100 µl of nuclease-free water. Eight reactions with 5 µl of mRNA each and a GFP construct-specific primer were prepared for complementary DNA synthesis using SuperScript IV reverse transcriptase (Thermo Fisher Scientific) according to the manufacturer's instructions. Half of the reactions were used as no reverse transcription control, in which the enzyme was replaced with water. After cDNA synthesis, the reactions were pooled and purified with DNA Clean & Concentrator-5 columns (Zymo Research). The barcode region was amplified with 10–20 cycles of polymerase chain

reaction (PCR) and read out by next generation sequencing. For the smaller validation libraries, only six leaves were used and all volumes except the reverse transcription were halved.

For the STARR-seq assay in maize protoplasts, transformed protoplasts were harvested by centrifugation (3 min at 200*g*, 4 °C) 16 h after electroporation. The protoplasts were washed three times with 1 ml of incubation buffer and centrifuged for 2 min at 200*g* and 4 °C. The cells were resuspended in 600 µl of TRIzol (Thermo Fisher Scientific) and incubated for 5 min at room temperature. The suspension was thoroughly mixed with 120 µl of chloroform and centrifuged (15 min at 20,000*g*, 4 °C). The upper, aqueous phase was transferred to a new tube, mixed with 120 µl of chloroform and centrifuged again (15 min at 20,000*g*, 4 °C). RNA was precipitated from 400 µl of the supernatant with 1 µl of 20 mg ml^{-1} glycogen (Thermo Fisher Scientific), 40 µl of 8 M LiCl and 1 ml of 100% (v/v) ethanol by incubation at –80 °C for 15 min. After centrifugation (30 min at 20,000*g*, 4 °C), the pellet was washed with 200 µl of 70% (v/v) ethanol, centrifuged again (5 min at 20,000*g*, 4 °C) and resuspended in 200 µl of nuclease-free water. The mRNAs were isolated from this solution using 50 µl of magnetic $Oligo(dT)_{25}$ beads (NEB) according to the manufacturer's protocol and the mRNAs were eluted in 40 µl of water. DNase I treatment and precipitation were performed as for the mRNAs obtained from tobacco plants but with half the volume. Reverse transcription, purification, PCR amplification and sequencing were performed as for the tobacco samples.

Subassembly and barcode sequencing

Paired-end sequencing on an Illumnia NextSeq 550 system was used for the subassembly of promoters with their corresponding barcodes. The promoter region was sequenced using partially overlapping, paired 144-bp reads and two 15-bp indexing reads were used to sequence

the barcodes. The promoter and barcode reads were assembled using PANDAseq[150] and the promoters were aligned to the designed core promoter sequences. Promoter-barcode pairs with less than five reads and promoters with a mutation or truncation were discarded. Barcode sequencing was performed using paired-end reads on a Illumnia NextSeq 550 platform. The reads were trimmed to only the barcode portion assembled with PANDAseq. All sequencing results were deposited in the NCBI Sequence Read Archive under the BioProject accession PRJNA714258. The scripts used for processing the raw reads are available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters.

Computational methods

For analysis of the STARR-seq experiments, the reads for each barcode were counted in the input and cDNA samples. Barcode counts below five were discarded. Barcode enrichment was calculated by dividing the barcode frequency (barcode counts divided by all counts) in the cDNA sample by that in the input sample. The enrichment of the promoters was calculated as the median enrichment of all barcodes linked to them. We calculated the promoter strength as the $\log_2$ of the promoter enrichment normalized to the enrichment of 35S minimal promoter. We used the average promoter strength from both replicates for all analyses. Spearman and Pearson correlations were calculated using the base R function. Significance was determined using the two-sided Wilcoxon rank-sum test as implemented in base R. GO-term enrichment analysis was performed using the ggprofiler2[151] (v.0.1.9) library for R and a custom gmt file with GOslim terms. Gene expression data was obtained from the EMBL-EBI Expression Atlas (https://www.ebi.ac.uk/gxa/about.html) using experiments E-MTAB-7978[133], E-GEOD-50191[131] and E-MTAB-5956[132] for *Arabidopsis*,

maize and sorghum, respectively. The tissue-specificity index τ was calculated as previously published[138]. Sequences for TFIIB proteins were obtained from Uniprot (https://www.uniprot.org/) and aligned using Clustal Omegan[152]. The code used for analyses is available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters.

Prediction of core promoter elements and TF-binding sites

The TATA box and Inr motifs were obtained from the plant promoter database[153] and for each a consensus motif was created by merging the motifs from dicot and monocot promoters using the universalmotif (v.1.6.3) library for R. Motifs for BREu and BREd were obtained from JASPAR[154]. The motifs for the polypyrimidine initiator TCT and the Y patch were created from published sequences of these elements[110,112]. Binding site motifs for *Arabidopsis* TFs were obtained from the PlantTFDB[74]. TF motifs were clustered by similarity using the compare_motifs() function from the R library universalmotif. The original clusters were improved by manual inspection and reannotation. Consensus motifs for the final TF motifs were created using the merge_motifs() function from universalmotif. Meme files with the motifs used in this study are available at https://github.com/tobjores/Synthetic-Promoter-Designs-Enabled-by-a-Comprehensive-Analysis-of-Plant-Core-Promoters. Promoter sequences were analysed with the universalmotif library assuming a neutral background nucleotide frequency. For the initiator elements, only the last ten (Inr) or the last six (TCT) bases were scanned. For BREu and BREd, the sequences immediately upstream and downstream of the highest scoring TATA box were analysed. For each sequence, the maximum motif score was calculated and normalized to the minimum (set to 0) and maximum (set to 1) scores possible. Sequences with a score of at least

0.85 were considered positive. For testing the effect of the BREu and BREd motifs (Supplementary Fig. 4.11), only sequences with a TATA box score of at least 0.7 were considered.

Design of validation sequences

To directly validate the importance of the TATA box, BRE^u and BRE^d elements, we picked 30 promoters (ten each from *Arabidopsis*, maize and sorghum if possible) according to the following criteria: for mutations of a canonical TATA box, we selected promoters with a TATA box motif score >0.9 in the −59 to −23 region. The two conserved T nucleotides in the core TATA motif were replaced individually or together with Gs. We also selected 30 promoters with a maximum TATA box motif score of 0.7 to 0.75. This weak TATA box was replaced with either a canonical TATA box motif (TATAAAT) or a mutated version of it (TAGAAAT). For the BRE elements, we first filtered for promoters with a TATA box motif score of at least 0.85 in the −59 to −23 region. From these, we picked promoters with a BRE motif score >0.85. For the BRE^u element, we mutated bases 3, 6 and 7 to T, A and A respectively. For the BRE^d element, we mutated bases 2,4 and 6 to A. We also selected promoters where both the BRE^u and the BRE^d motif scores are <0.5 to insert either a canonical BRE^u (AGCGCGCC) or BRE^d (GTTTGTT) element.

Synthetic promoter design

Synthetic promoters were designed by generating 170-bp long random sequences with a nucleotide composition similar to an average *Arabidopsis* (35.2% A, 16.6% C, 15.3% G, 32.8% T) or maize (24.5% A, 29.0% C, 22.5% G, 23.9% T) promoter. We filtered out any random sequence with motif scores higher than 0.75 for a TATA box, Inr or Y patch element or for TF-

binding site of clusters 1, 15, 16 or 22. Promoters containing recognition sites for the restriction enzymes used for cloning (BsaI and BbsI) were also removed. From each set of promoters (*Arabidopsis* or maize nucleotide composition) that passed the filters, we randomly selected ten variants for further modification. The promoters were kept as is or modified with a TATA box (TATAAATA) at positions 133–140, a Y patch (A and G nucleotides of the promoter were changed to C) at positions 147–154 and/or an Inr element (yyyyTCAyyy, where y indicates a change of A to T or G to C) at positions 160–169. To study the effect of TFs, the synthetic promoters with the TATA box were chosen as backgrounds. Binding sites for NAC (cluster 1, TTACGTGnnnnACAAG, where n represents bases of the promoter background), TCP (cluster 15, TGGGGCCCAC and cluster 22, GGGACCAC) or HSF/S1Fa-like (cluster 16, GAAGCTTCTAGAA) TFs were inserted at various positions of these promoters.

Computational modelling of promoter strength

To predict promoter strength, we built separate models for the tobacco leaf and the maize protoplast system. We used the results from the libraries with the 35S enhancer in the dark for training and validation. The models were trained on a set of 90% of all measured promoters and tested against the held-out set of the remaining 10% of the promoters.

We used the base R function lm() to build a linear model for predicting promoter strength on the basis of the promoter's GC content and its maximum motif score for six core promoter elements (TATA box, Inr, TCT, BRE^u, BRE^d and Y patch) and 72 consensus TF-binding motifs.

To build a direct sequence to promoter strength model we built a CNN using the tensorflow (v.2.2) package in python. The model consists of two forward- and reverse-sequence scan layers adapted from DeepGMAP[155] with 128 filters and a kernel width of 13 that feed into a regular

convolutional layer (128 filters, kernel width 13, ReLU activation). Each convolutional layer is followed by a dropout layer with a 0.15 dropout rate. The output of the convolutional layers is fed into a dense layer with 64 filters with batch-normalization and ReLU activation that is followed by a final dense layer generating the single output. We initialized the first convolutional layer kernel with the clustered TF motifs. The source code and the models are available on GITHUB.

In silico evolution of promoter sequences

We used the CNNs to improve promoter performance in an iterative fashion. In each round, we generated all possible single nucleotide variants of a given promoter, scored them with the CNN models and kept the variant with the highest predicted activity for the next round. The sequences were scored with either just one of the models trained on the tobacco leaf or the maize protoplast data or with both models in which case the mean of both predictions was used to select the best-performing variant. We experimentally tested these sequences after three and ten rounds of this process. For the evolution, we selected native promoters showing either weak, intermediate or strong activity in both assay systems or were strong in one system and weak in the other one. Additionally, we also performed the in silico evolution with the synthetic promoters described above.

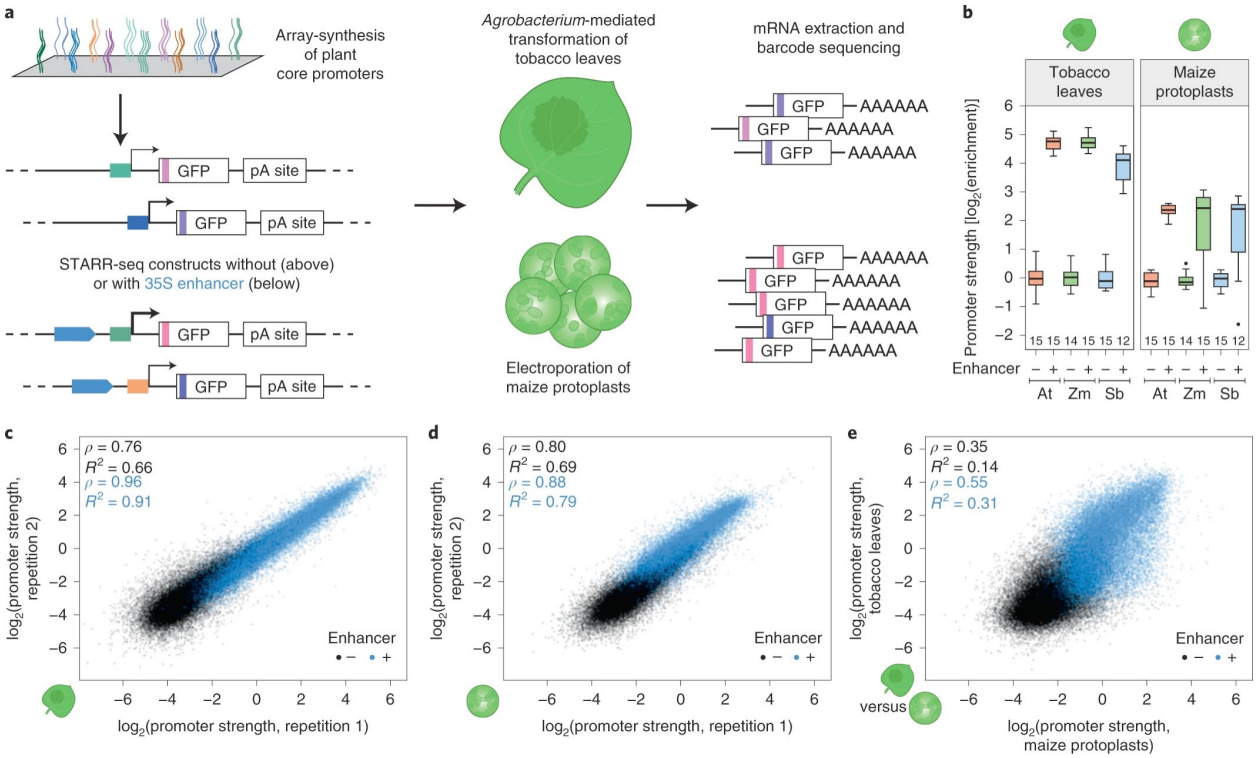


Figure 4.1.

STARR-seq measures core promoter strength in tobacco leaves and maize protoplasts.

a, Assay scheme. The core promoters (bases −165 to +5 relative to the TSS) of all genes of *Arabidopsis*, maize and sorghum were array-synthesized and cloned into STARR-seq constructs to drive the expression of a barcoded GFP reporter gene. For each species, two libraries, one without and one with a 35S enhancer upstream of the promoter, were created. The libraries were subjected to STARR-seq in transiently transformed tobacco leaves and maize protoplasts. **b**, Each promoter library (At, *Arabidopsis*; Zm, maize; Sb, sorghum) contained two internal control constructs driven by the 35S minimal promoter without (−) or with (+) an upstream 35S enhancer. The enrichment ($\log_2$) of recovered mRNA barcodes compared to DNA input was calculated with the enrichment of the enhancer-less control set to 0. In all following figures, this metric is indicated as promoter strength. Each boxplot (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) represents the enrichment of all barcodes linked

to the corresponding internal control construct. The number of barcodes is indicated at the bottom of the plot. **c**,**d**, Correlation (Pearson's R^2 and Spearman's ρ) of two biological replicates of STARR-seq using the maize promoter libraries in tobacco leaves (**c**) or in maize protoplasts (**d**). **e**, Comparison of the strength of maize promoters in tobacco leaves and maize protoplasts. Pearson's R^2 and Spearman's ρ are indicated.

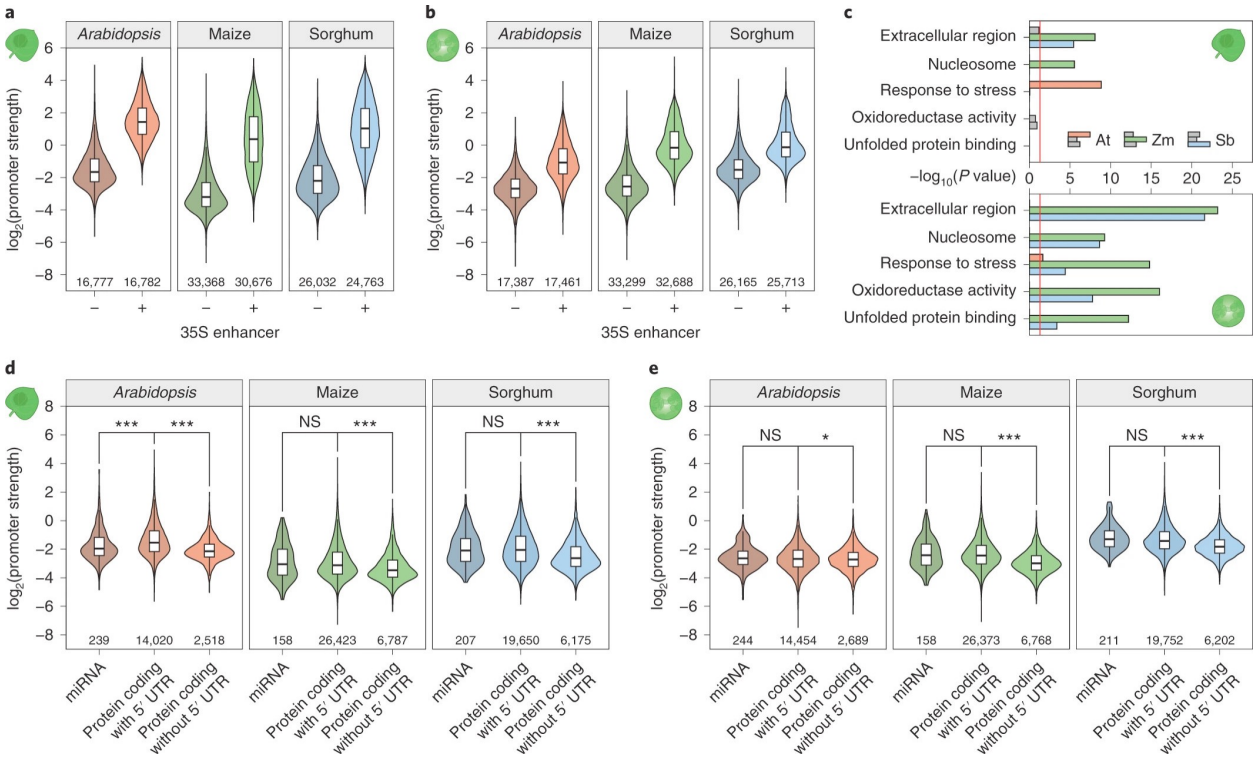


Figure 4.2.

Plant core promoters span a wide range of activity.

a,**b**, Violin plots of the strength of plant promoters from the indicated species as measured by STARR-seq in tobacco leaves (**a**) or maize protoplasts (**b**) for libraries without (−) or with (+) the 35S enhancer upstream of the promoter. **c**, Enrichment of selected GO terms for genes associated with the 1,000 strongest promoters in the *Arabidopsis* (At), maize (Zm) and sorghum (Sb) promoter libraries without enhancer in tobacco leaves (top panel) and maize protoplasts (bottom panel). The red line marks the significance threshold (adjusted $P \leq 0.05$). Non-significant bars are grey. The *P* values were determined using the gprofiler2 library in R with gSCS correction for multiple testing. **d**,**e**, Violin plots of promoter strength (libraries without 35S enhancer) in tobacco leaves (**d**) or maize protoplasts (**e**). Promoters were grouped by gene type. In **a**,**b**,**d** and **e**, violin plots represent the kernel density distribution and the boxplots within represent the median (centre line), upper and lower quartiles (box limits) and 1.5× the interquartile range (whiskers) for

all corresponding promoters. Numbers at the bottom of the plot indicate the number of tested promoters. Significant differences between two samples were determined using the two-sided Wilcoxon rank-sum test and are indicated: $^{*}P \leq 0.01$; $^{**}P \leq 0.001$; $^{***}P \leq 0.0001$; NS, not significant.

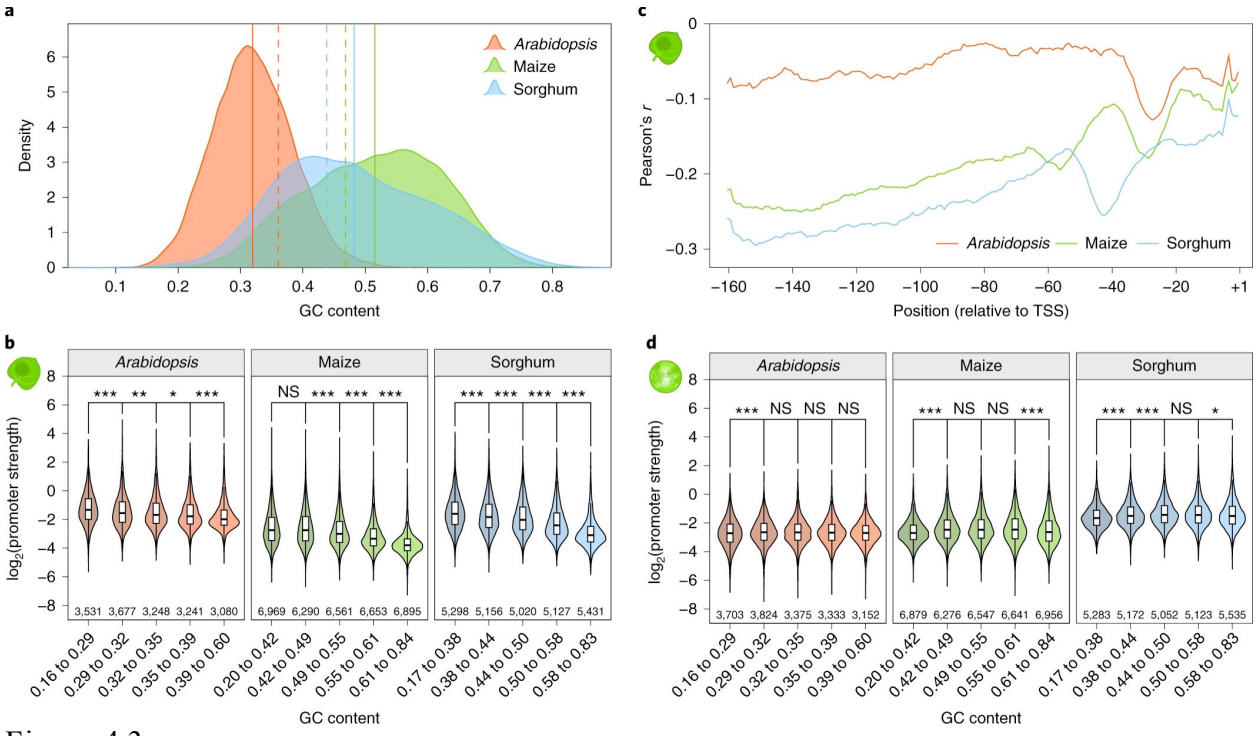


Figure 4.3.

GC content affects promoter strength in tobacco leaves.

a, Distribution of GC content for all promoters of the indicated species. Lines denote the mean GC content of promoters (solid line) and the whole genome (dashed line). **b**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for libraries without enhancer in tobacco leaves. Promoters are grouped by GC content to yield groups of approximately similar size. **c**, Correlation (Pearson's *r*) between promoter strength and the GC content of a ten-base window around the indicated position in the plant promoters. **d**, As **b** but for promoter strength in maize protoplasts.

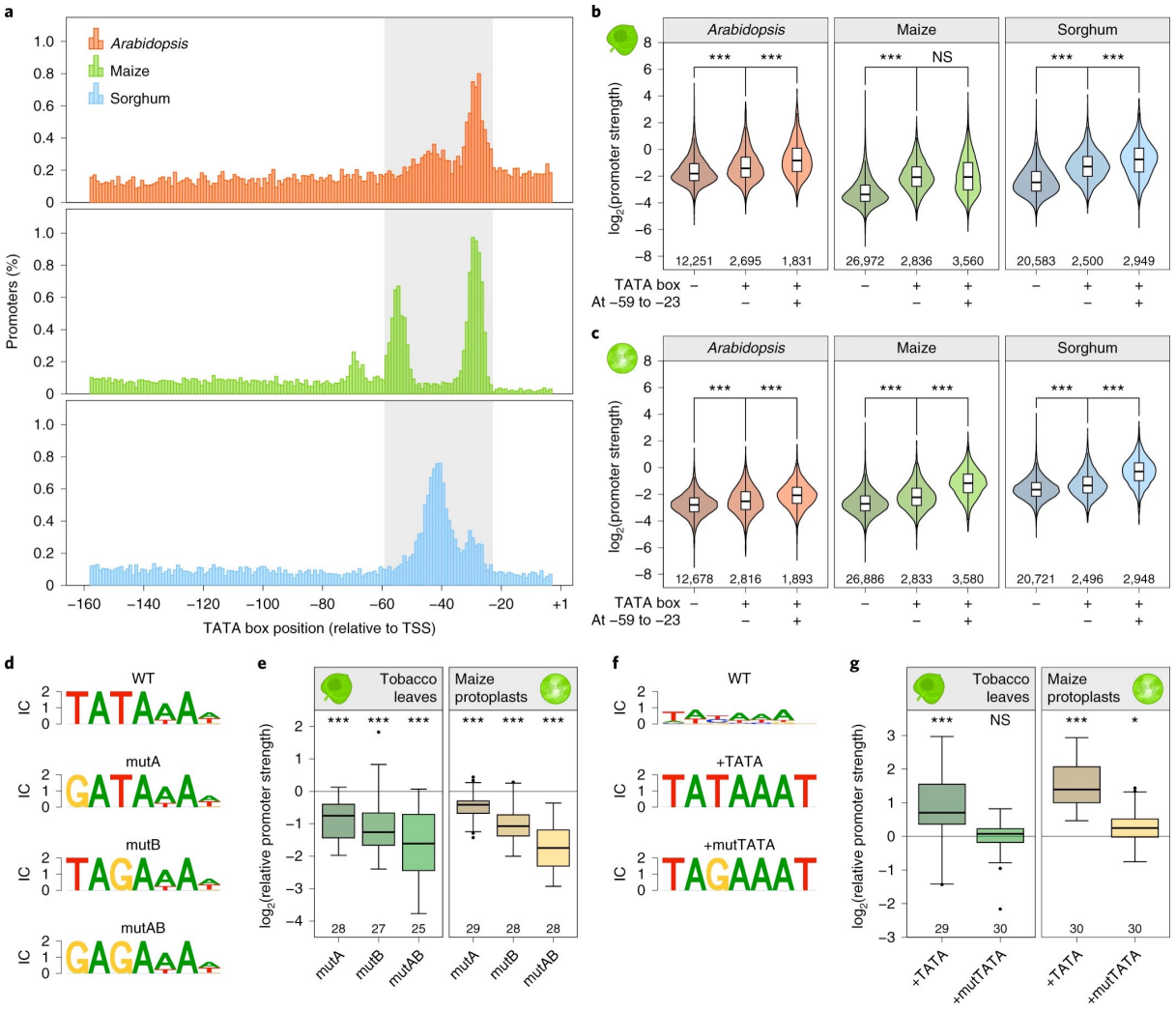


Figure 4.4.

The TATA box is a key determinant of promoter strength.

a, Histograms showing the percentage of promoters with a TATA box at the indicated position. The region between positions −59 and −23 in which most TATA boxes reside is highlighted in grey. **b**,**c**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for libraries without enhancer in tobacco leaves (**b**) or maize protoplasts (**c**). Promoters without a TATA box (−) were compared to those with a TATA box outside (+/−) or within (+/+) the −59 to −23 region. **d**–**g**, Thirty plant promoters with a strong (**d**,**e**) or weak (**f**,**g**) TATA box (wild type, WT) were tested. One (mutA and mutB) or two (mutAB) T > G mutations

were inserted into promoters with a strong TATA box (**d**,**e**). A canonical TATA box (+TATA) or one with a T > G mutation (+mutTATA) was used to replace the weak TATA box (**f**,**g**). Logoplots (**f**,**d**) of the TATA box regions of these promoters and their strength (**g**,**e**) relative to the WT promoter (set to 0, horizontal black line) are shown. Boxplots (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points, outliers) denote the strength of the indicated promoter variants. Numbers at the bottom of the plot indicate the number of tested promoter elements. Significant differences from a null distribution were determined using the two-sided Wilcoxon signed rank test and are indicated: *$P \leq 0.01$; **$P \leq 0.001$; ***$P \leq 0.0001$; NS, not significant. IC, information content.

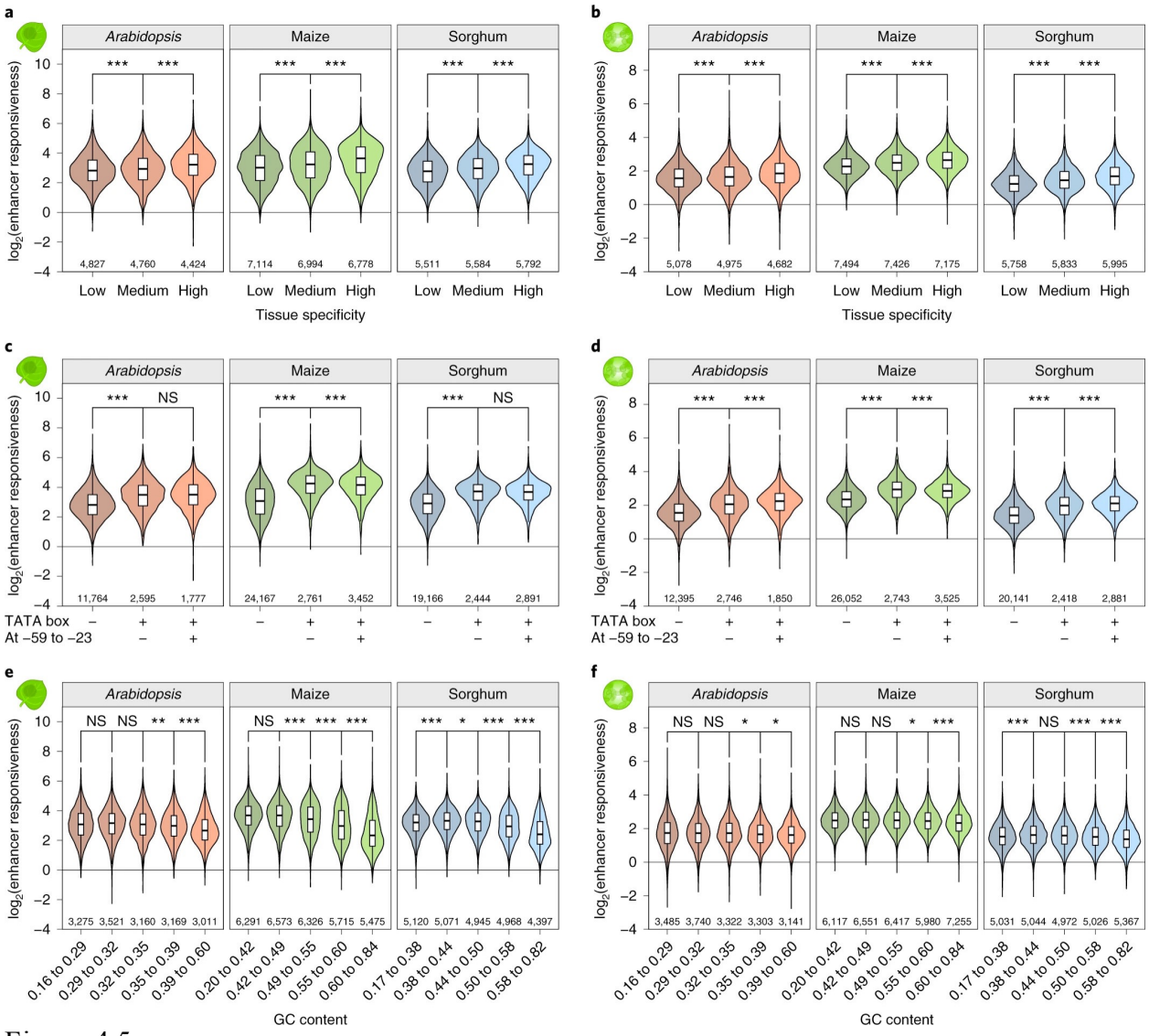


Figure 4.5.

Enhancer responsiveness of promoters depends on the TATA box and GC content.

a,**b**, Violin plots of enhancer responsiveness (promoter strength$^{\text{with enhancer}}$ divided by promoter strength$^{\text{without enhancer}}$) in tobacco leaves (**a**) or maize protoplasts (**b**). Promoters were grouped into three bins of approximately similar size according to the tissue-specificity τ of the expression of the associated gene. **c**,**d**, Violin plots of enhancer responsiveness in tobacco leaves (**c**) or maize protoplasts (**d**). Promoters without a TATA box (−) were compared to those with a TATA box outside (+/−) or within (+/+) the −59 to −23 region. **e**,**f**, Violin plots of enhancer

responsiveness in tobacco leaves (**e**) or maize protoplasts (**f**) for promoters grouped by GC content. Violin plots, boxplots and significance levels in (**a**–**f**) are as defined in Fig. 2.

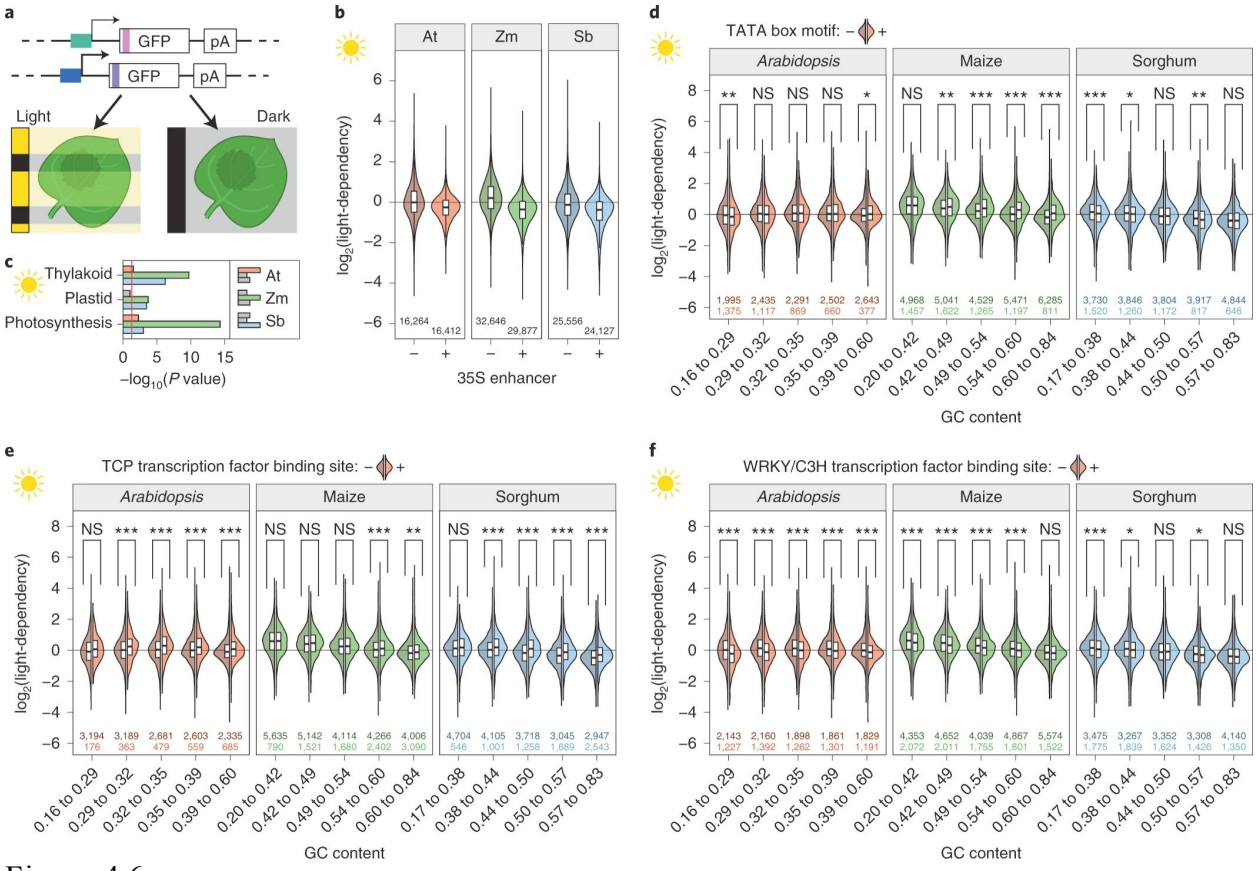


Figure 4.6.

Promoter strength can be modulated by light.

a, Tobacco leaves were transiently transformed with STARR-seq promoter libraries and the plants were kept for 2 d in 16 h light/8 h dark cycles (light) or completely in the dark (dark) before mRNA extraction. **b**, Violin plots of light-dependency (promoter strength^{light} divided by promoter strength^{dark}) for promoters in the libraries with (+) or without (−) the 35S enhancer. **c**, Enrichment of selected GO terms for genes associated with the 1,000 most light-dependent promoters. The red line marks the significance threshold (adjusted $P \leq 0.05$). Non-significant bars are grey. The *P* values were determined using the gprofiler2 library in R with gSCS correction for multiple testing. **d**–**f**, Violin plots of light-dependency. Promoters are grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a TATA

box (**d**) or a binding site for TCP (**e**) or WRKY (**f**) TFs. Violin plots, boxplots and significance levels in **b** and **d**–**f** are as defined in Fig. 2. Only one half is shown for violin plots in **d**–**f**.

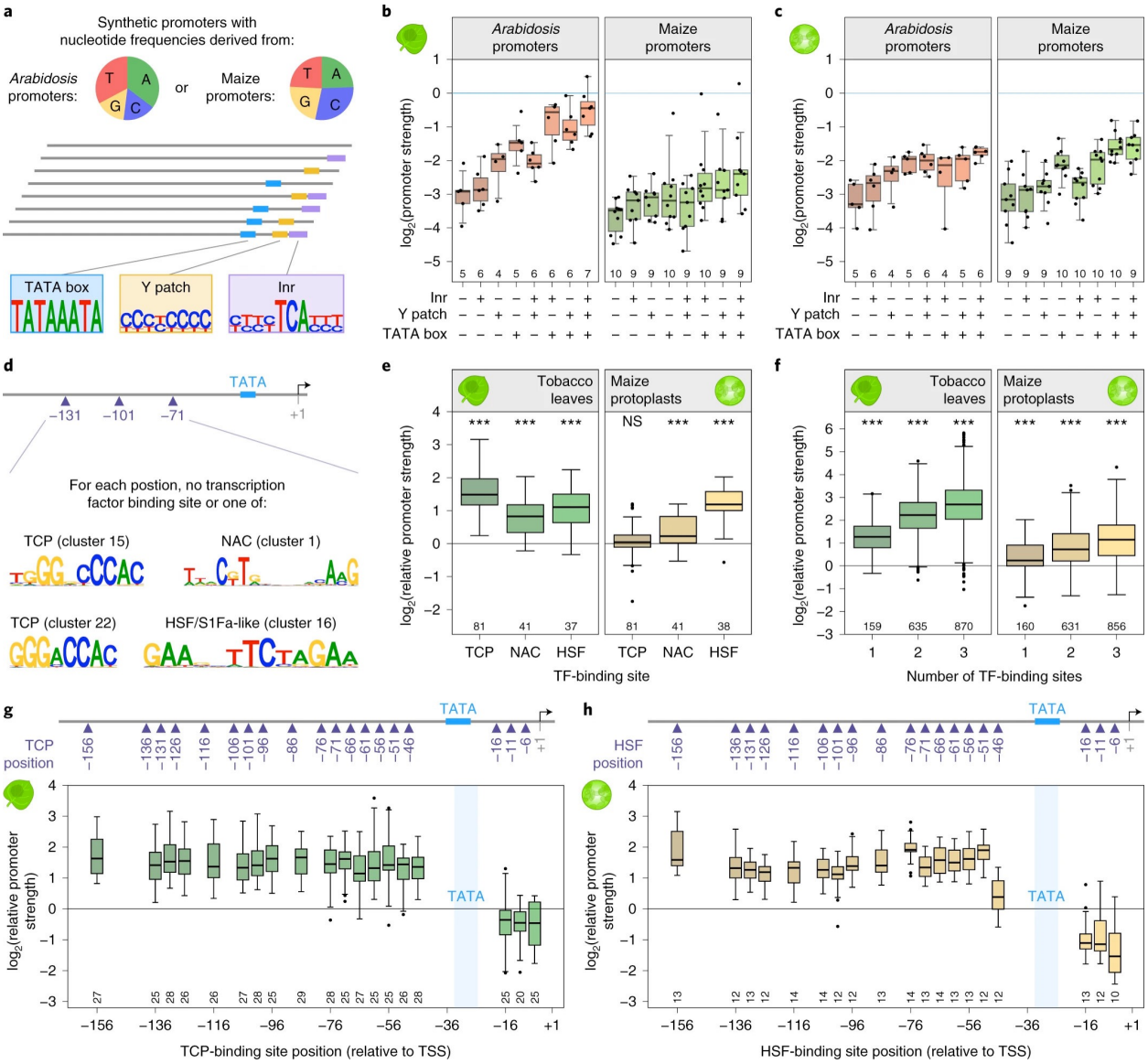


Figure 4.7.

Design and validation of synthetic promoters.

a–**c**, Synthetic promoters with nucleotide frequencies similar to an average *Arabidopsis* (35.2% A, 16.6% C, 15.3% G and 32.8% T) or maize (24.5% A, 29.0% C, 22.5% G and 23.9% T) promoter were created and modified by adding a TATA box, Y patch and/or Inr element (**a**); promoter strength was determined by STARR-seq in tobacco leaves (**b**) and maize protoplasts (**c**). Promoters with an *Arabidopsis*-like nucleotide composition are shown on the left, those with maize-like base frequencies on the right. The strength of the 35S minimal promoter is indicated by

a horizontal blue line. Individual data points are shown. **d**–**f**, TF-binding sites for TCP, NAC and HSF transcription factors were inserted at positions 35, 65 and/or 95 of the synthetic promoters with a TATA box (**d**) and the activity of promoters with a single binding site for the indicated TF (**e**) or multiple binding sites (**f**) was determined in tobacco leaves (left panel) or maize protoplasts (right panel). **g**,**h**, A single TCP (**g**) or HSF (**h**) TF-binding site was inserted at the indicated position in the synthetic promoters containing a TATA box. The strength of these promoters was measured in tobacco leaves (**g**) or maize protoplasts (**h**). Boxplots and significance levels in **b**,**c** and **e**–**h** are as defined in Fig. 4. In **e**–**h**, the corresponding promoter without any TF-binding site was set to 0 (horizontal black line).

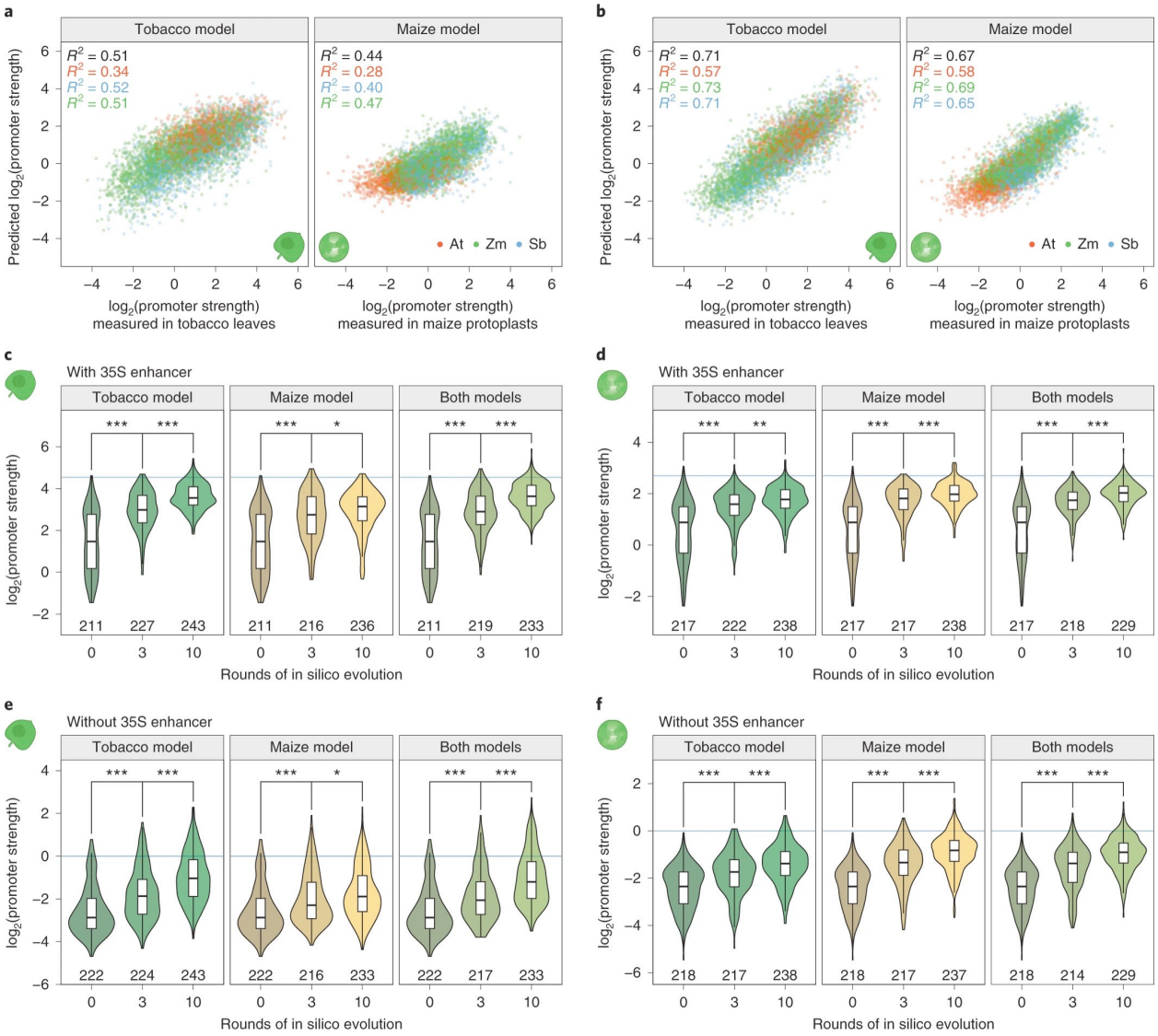
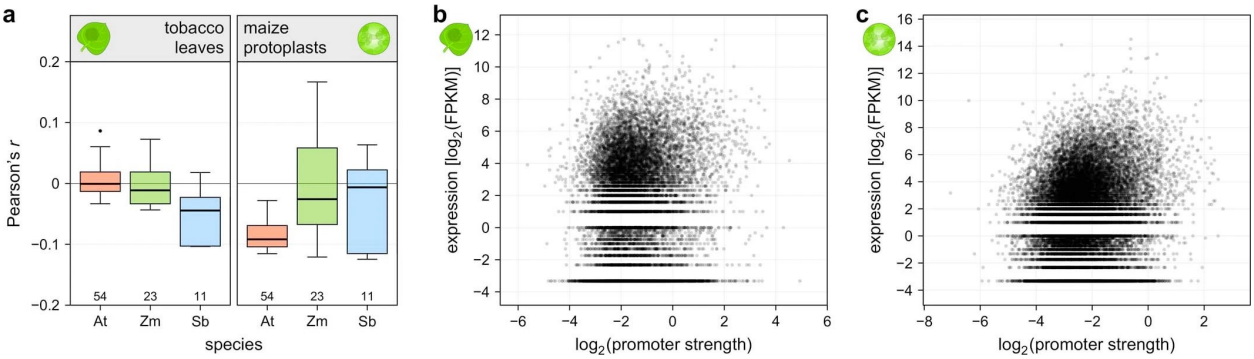


Figure 4.8.

Computational models can predict promoter strength and enable in silico evolution of plant promoters.

a, Correlation between the promoter strength as determined by STARR-seq using promoter libraries with the 35S enhancer and predictions from a linear model based on the GC content and motif scores for core promoter elements and TFs. The models were trained on data from the tobacco leaf system (tobacco model) or the maize protoplasts (maize model). The overall

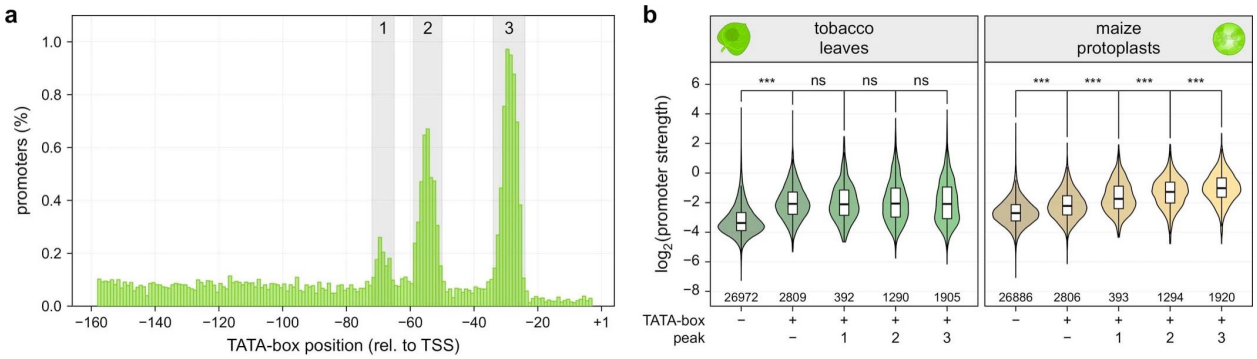
correlation is indicated in black and correlations for each species are coloured as indicated (inset). Correlations (Pearson's R^2) are shown for a test set of 10% of all promoters. **b**, Similar to **a** but the prediction is based on a CNN trained on promoter sequences. **c**–**f**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength of the unmodified promoters (0 rounds of evolution) or after they were subjected to three or ten rounds of in silico evolution as determined in tobacco leaves (**c**,**e**) or maize protoplasts (**d**,**f**). The promoters were tested in a library with (**c**,**d**) or without (**e**,**f**) an upstream 35S enhancer. The models used for the in silico evolution are indicated on each plot. The promoter strength of the 35S promoter is indicated by a horizontal blue line.



Supplemental Figure 4.9.

Promoter strength and in vivo expression levels of corresponding genes are not correlated.

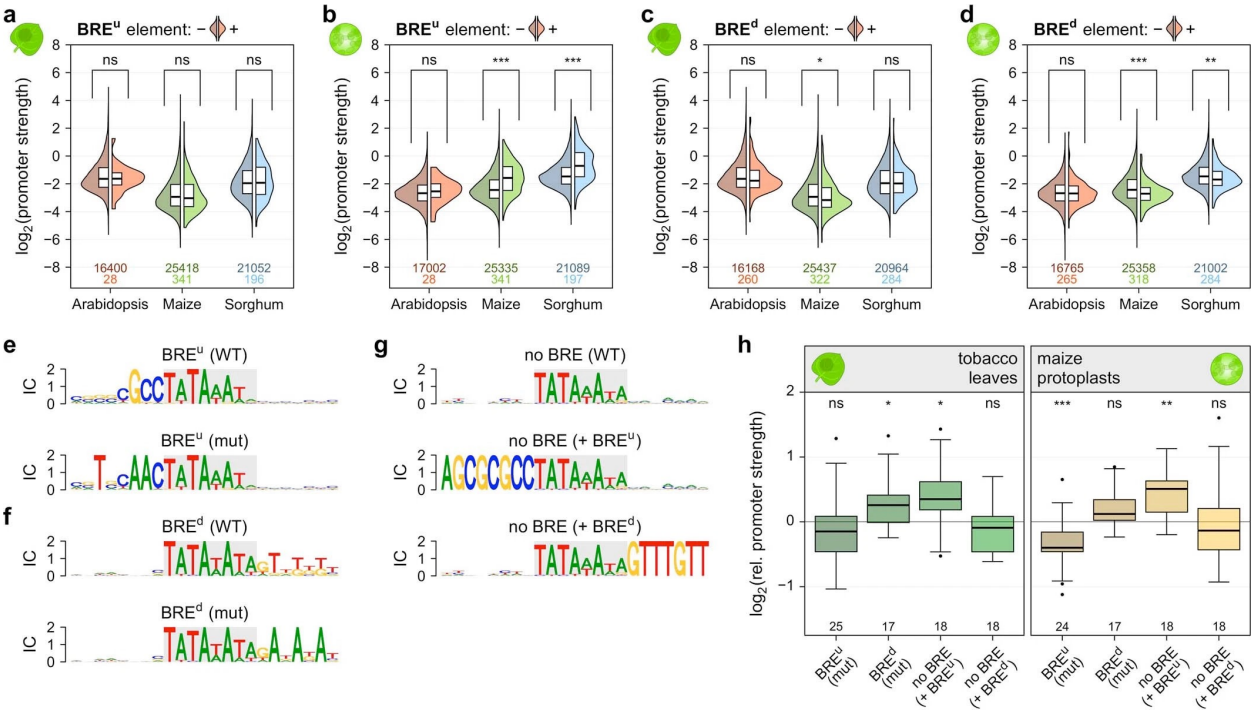
a, Correlation (Pearson's *r*) between the promoter strength and expression levels of the corresponding genes in the indicated species. Each boxplot (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range; points, outliers) represents the correlation for all individual tissue samples in the RNA-seq dataset (see Methods). The number of samples in the RNA-seq dataset is indicated at the bottom of the plot. **b**,**c**, Examples of the correlation between gene expression (Arabidopsis adult cotyledon (**b**) or maize root cortex (**c**) samples) and promoter strength as determined in tobacco leaves (**b**) or maize protoplasts (**c**). These examples correspond to the highest correlations in (**a**).



Supplemental Figure 4.10.

Strength of maize promoters depends on the TATA box location in maize protoplasts.

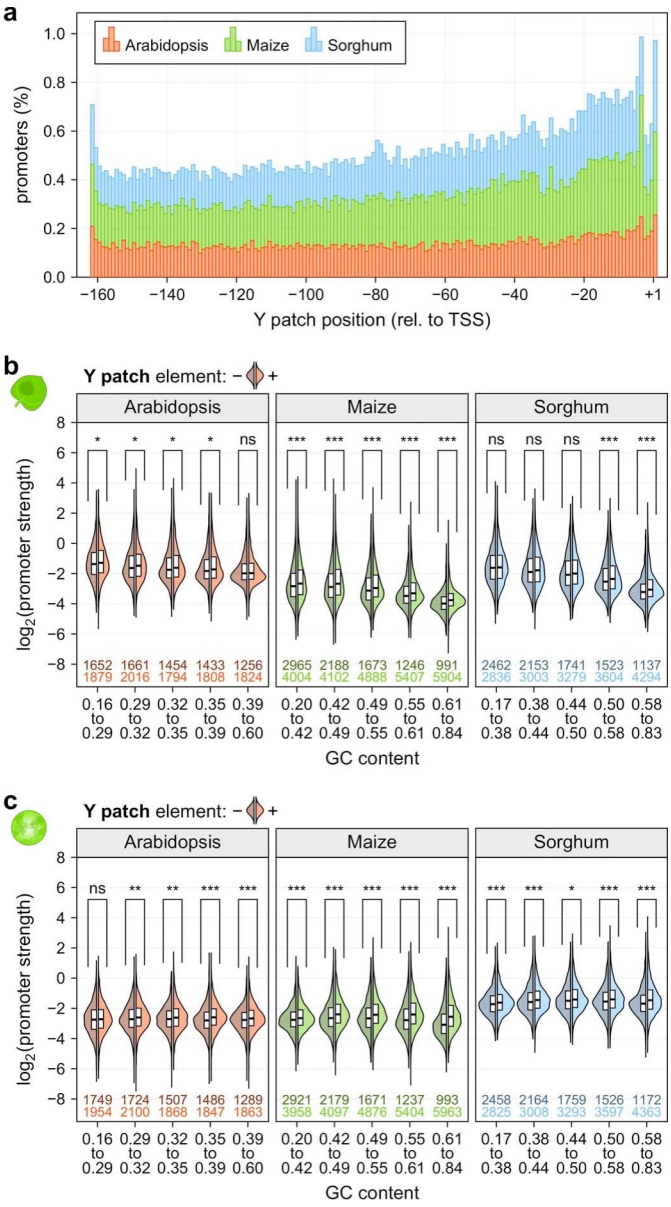
a, Histogram showing the percentage of maize promoters with a TATA box at the indicated position (reproduced from Fig. 4). Three peaks in the distribution of TATA boxes are highlighted in grey. Peak 1 spans bases −72 to −65, peak 2 spans bases −59 to −50, and peak 3 spans bases −34 to −24. **b**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for maize promoters without enhancer in the indicated assay system. Promoters without a TATA box (−) were compared to those with a TATA box outside (+/−) or within one of the three peaks highlighted in (**a**).



Supplemental Figure 4.11.

The BREu element is most active in maize protoplasts.

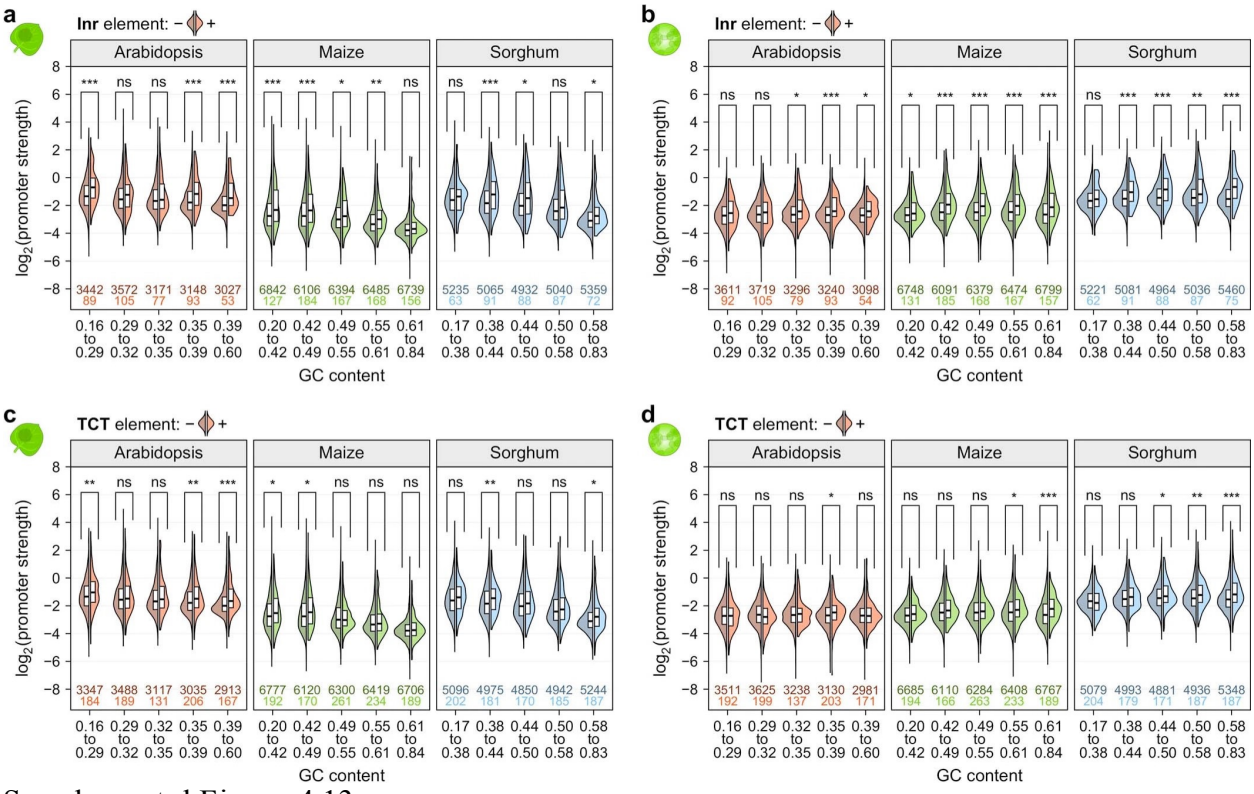
a-d, Violin plots of promoter strength in tobacco leaves (**a**,**c**) or maize protoplasts (**b**,**d**). Promoters with a strong or intermediate TATA box (motif score ≥ 0.7; see Methods) were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a BRE^{u} (**a**,**b**), or BRE^{d} (c,d) element. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots. **e**,**f**, Logoplots for promoters with a BRE^{u} (**e**) or BRE^{d} (**f**) before (WT) and after (mut) introducing mutations that disrupt the elements. **g**, Logoplots for promoters without a BRE (WT) and with an inserted BRE^{u} (+ BRE^{u}) or BRE^{d} (+ BRE^{d}) element. **h**, Boxplots and significance levels (as defined in Fig. 4) for the relative strength of the promoter variants shown in (**e-g**). The corresponding WT promoter was set to 0 (horizontal black line).



Supplemental Figure 4.12.

The Y patch is a plant-specific core promoter element.

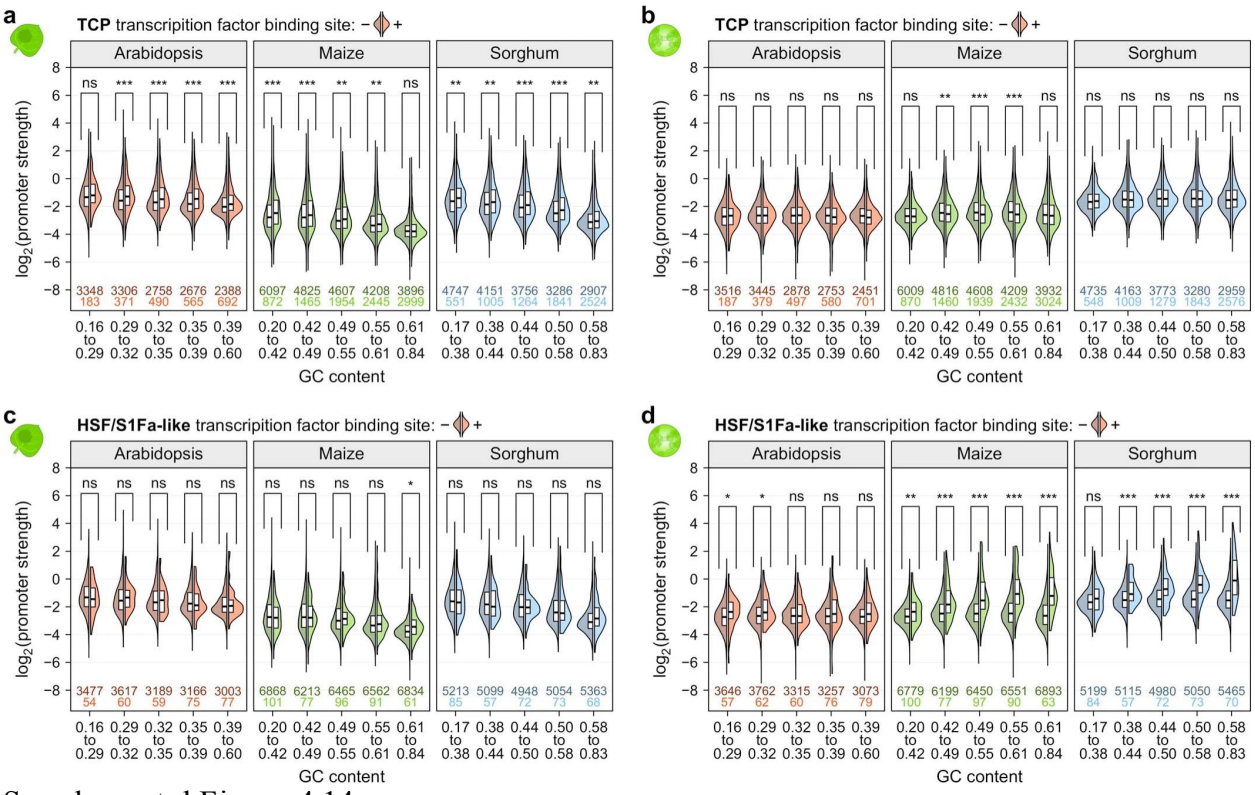
a, Histogram showing the percentage of promoters with a TATA box at the indicated position. **b**,**c**, Violin plots of promoter strength in tobacco leaves (**b**) or maize protoplasts (**c**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a Y patch. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



Supplemental Figure 4.13.

Core promoter elements at the TSS influence promoter strength.

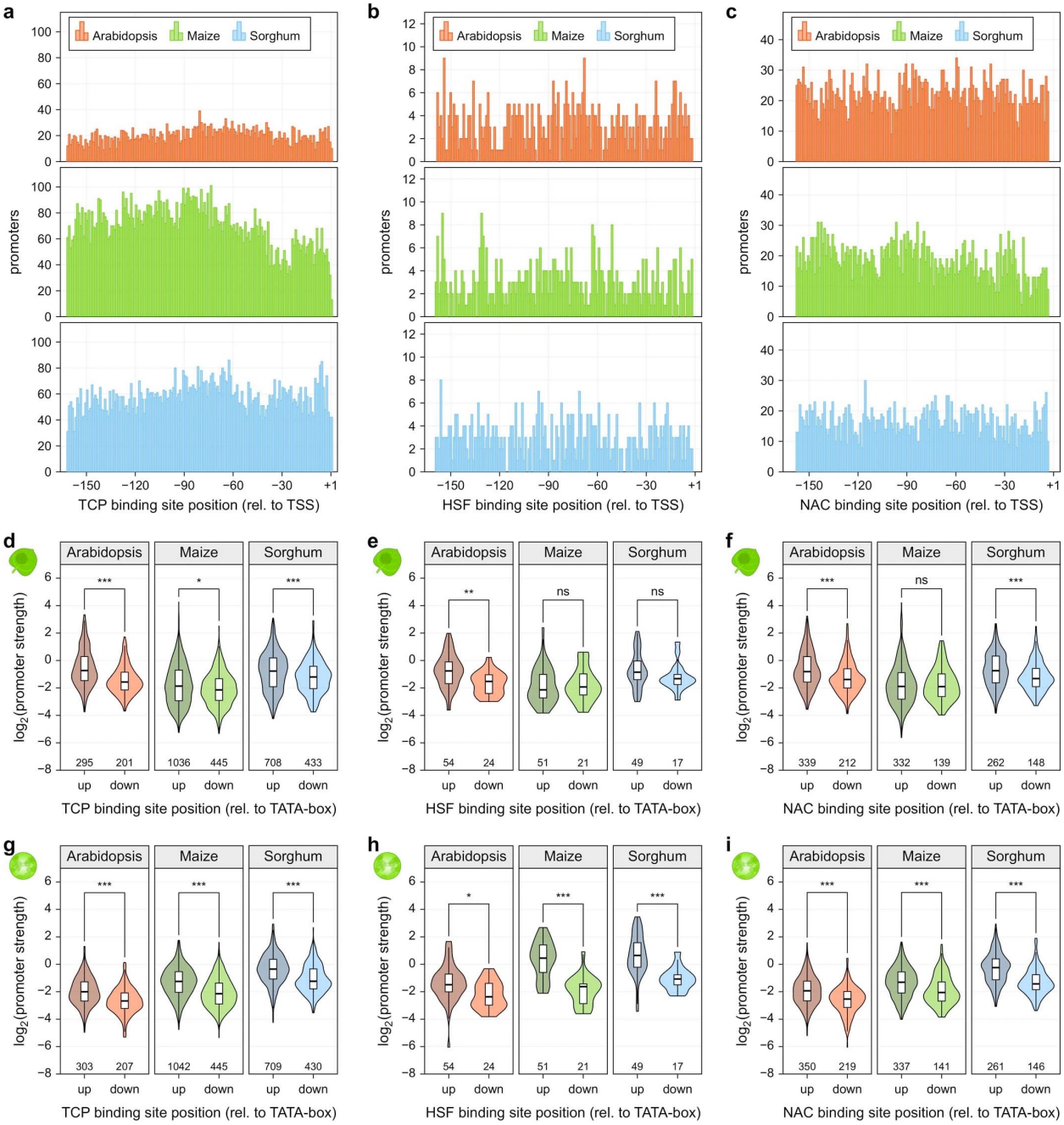
a-d, Violin plots of promoter strength in tobacco leaves (**a**,**c**) or maize protoplasts (**b**,**d**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) an Inr (**a**,**b**), or TCT (**c**,**d**) element at the TSS. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



Supplemental Figure 4.14.

Transcription factor binding sites contribute to promoter strength in an assay system-dependent manner.

a-d, Violin plots of promoter strength for libraries without enhancer in tobacco leaves (**a**,**c**) or maize protoplasts (**b**,**d**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a binding site for TCP (**a**,**b**) or HSF (**c**,**d**) transcription factors. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.

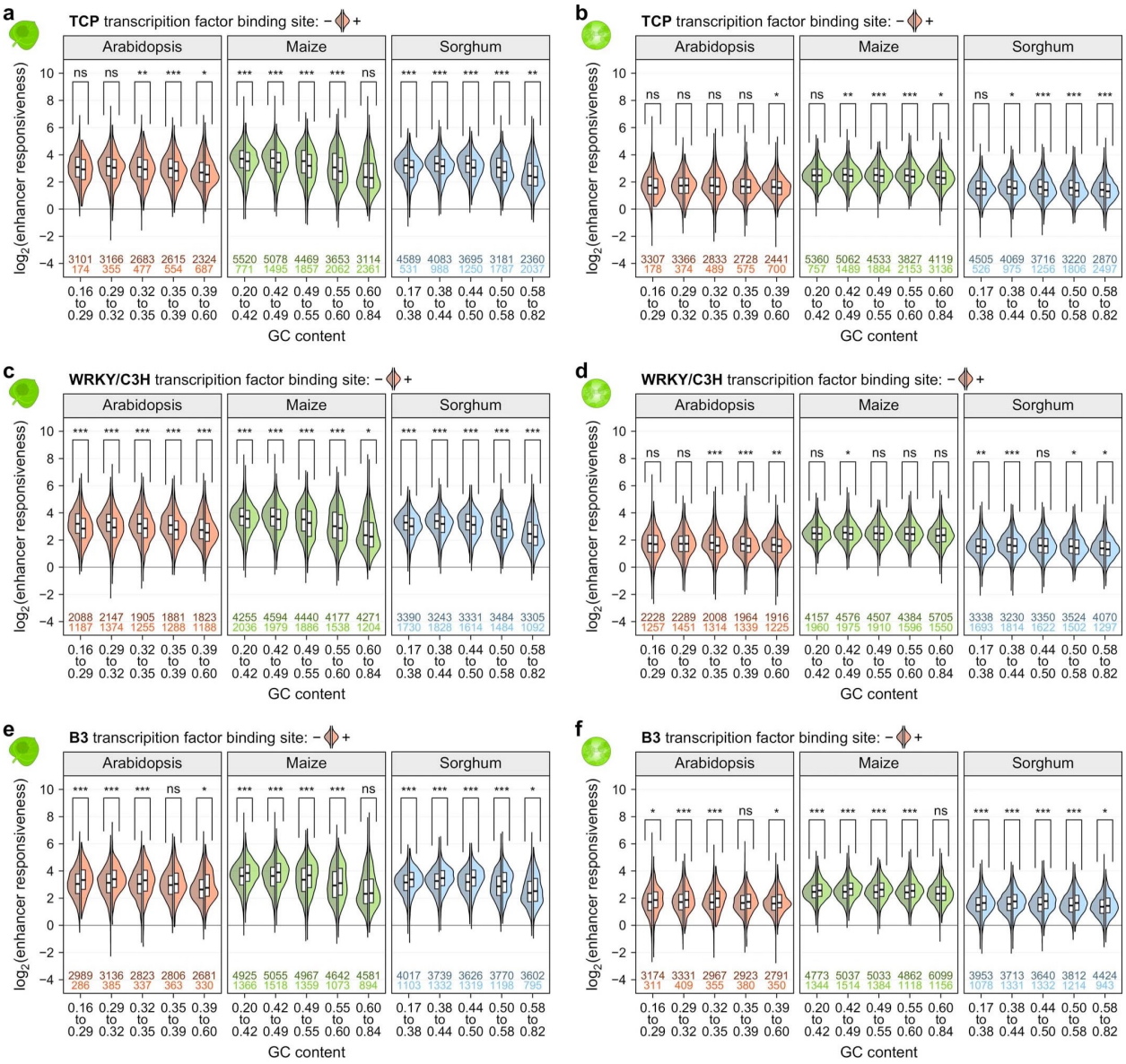


Supplemental Figure 4.15.

Transcription factor binding sites are more active upstream of the TATA box.

a-c, Histograms showing the number of promoters with a TCP (**a**), HSF (**b**), or NAC (**c**) transcription factor binding site at the indicated position. **d-i**, Violin plots, boxplots and significance levels (as defined in Fig. 2) of promoter strength for libraries without enhancer in tobacco leaves (**d-f**) or maize protoplasts (**g-i**). Promoters were grouped by the position of their

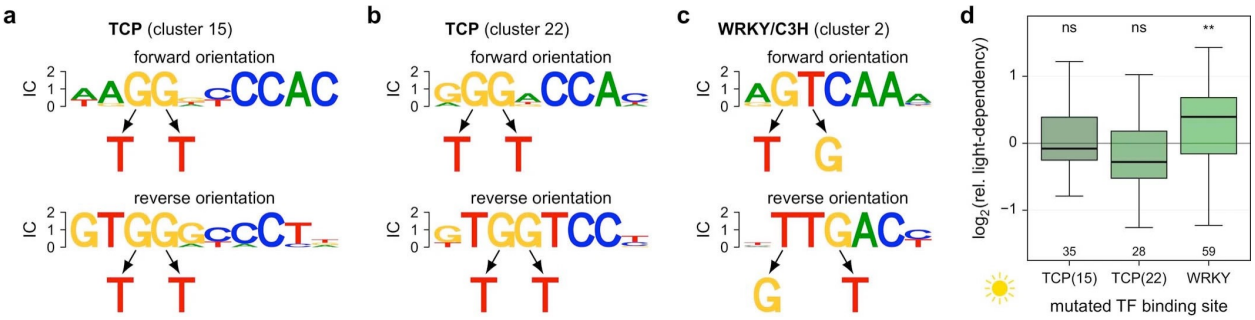
TCP (**d**,**g**), HSF (**e**,**h**), or NAC (**f**,**i**) transcription factor binding site relative to the TATA box: either upstream (up) or downstream (down).



Supplemental Figure 4.16.

Promoter-proximal transcription factor binding sites influence enhancer responsiveness.

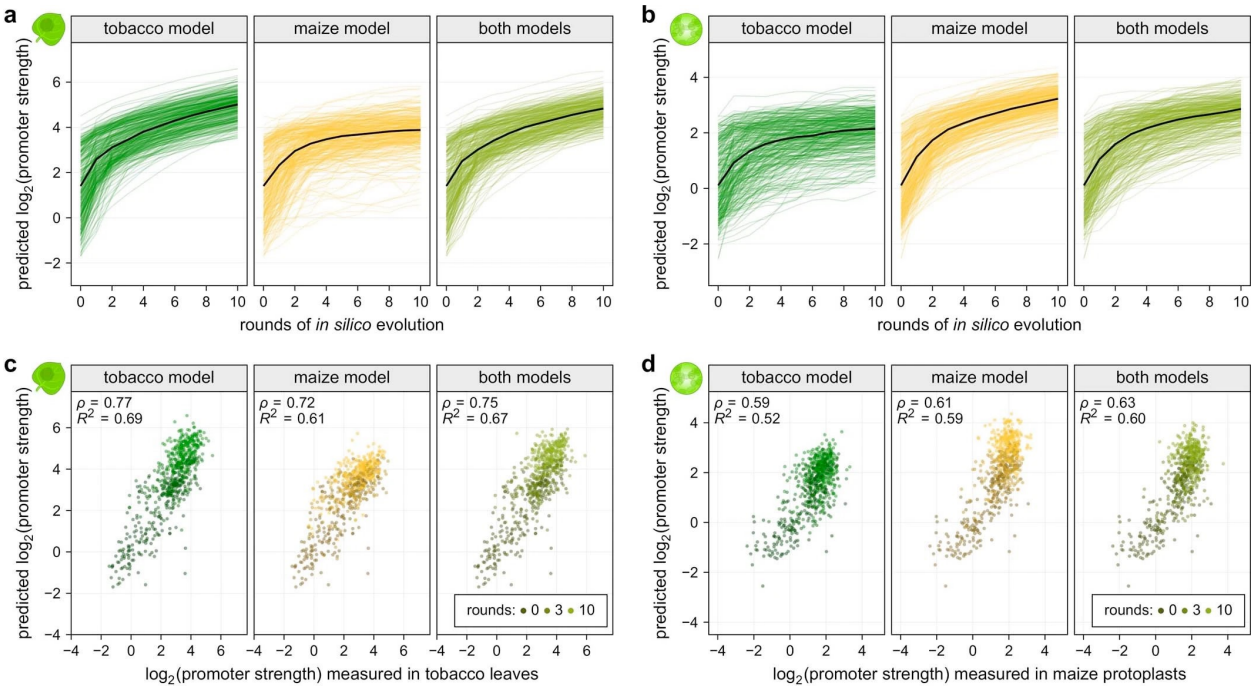
a-**f**, Violin plots of enhancer responsiveness in tobacco leaves (**a**,**c**,**e**) or maize protoplasts (**b**,**d**,**f**). Promoters were grouped by GC content and split into promoters without (left half, darker colour) or with (right half, lighter colour) a TCP (**a**,**b**), WRKY (**c**,**d**), or B3 (**e**,**f**) transcription factor binding site. Violin plots, boxplots and significance levels are as defined in Fig. 2. Only one half is shown for violin plots.



Supplemental Figure 4.17.

Mutations in transcription factor binding sites alter light-dependency.

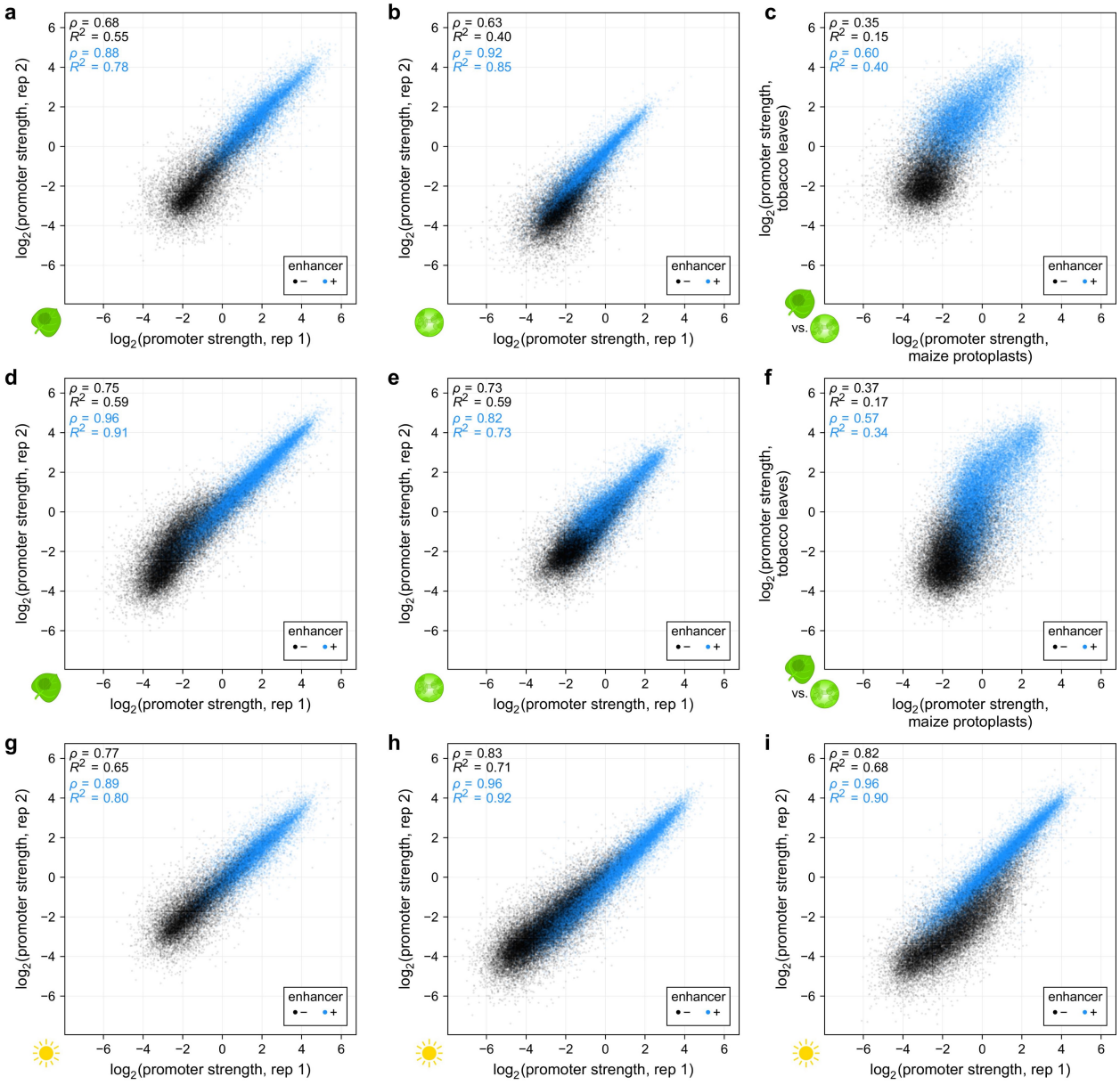
a-c, One or two T > G mutations were introduced in binding sites for TCP (**a,b**) or WRKY (**c**) transcription factors. The orientation of a binding site in the wild type promoter determined the bases that were mutated. **d**, Boxplots and significance levels (as defined in Fig. 4) for the relative light-dependency of promoters harbouring mutations in the indicated transcription factor binding site as shown in (**a-c**). The corresponding wild type promoter was set to 0 (horizontal black line).



Supplemental Figure 4.18.

The in silico evolution of promoters is most effective in early rounds.

a,**b**, 150 native and 160 synthetic promoters were subjected to 10 rounds of *in silico* evolution and the strength of the evolved promoters was predicted with the tobacco model (**a**) or the maize model (**b**). The black line represents the median promoter strength after each round. **c**,**d**, Correlation (Pearson's *R*2 and Spearman's ρ) between the predicted and experimentally determined strength of promoters after 0, 3, or 10 rounds of *in silico* evolution. Promoter strengths measured in tobacco leaves were compared to predictions from the tobacco model (**c**) and the data from maize protoplasts was compared to the predictions from the maize model (**d**). The models used for the *in silico* evolution are indicated on each plot.

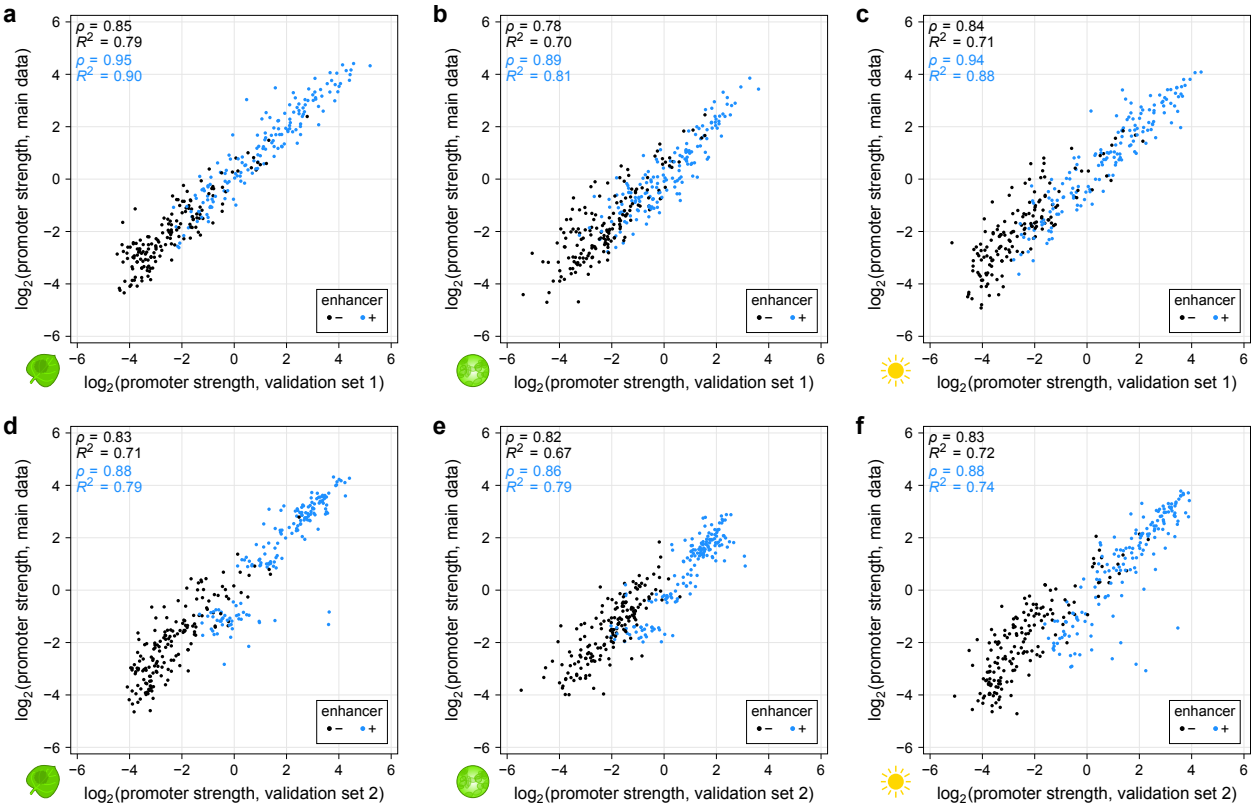


Supplemental Figure 4.19.

The promoter STARR-seq assay is highly reproducible but promoter strength depends on the assay system.

a,b, Correlation of two biological replicates of STARR-seq using the Arabidopsis promoter libraries in tobacco leaves (a) or in maize protoplasts (b). c, Comparison of the strength of Arabidopsis promoters in tobacco leaves and maize protoplasts. d,e, Correlation of two biological replicates of STARR-seq using the sorghum promoter libraries in tobacco leaves (d) or in maize

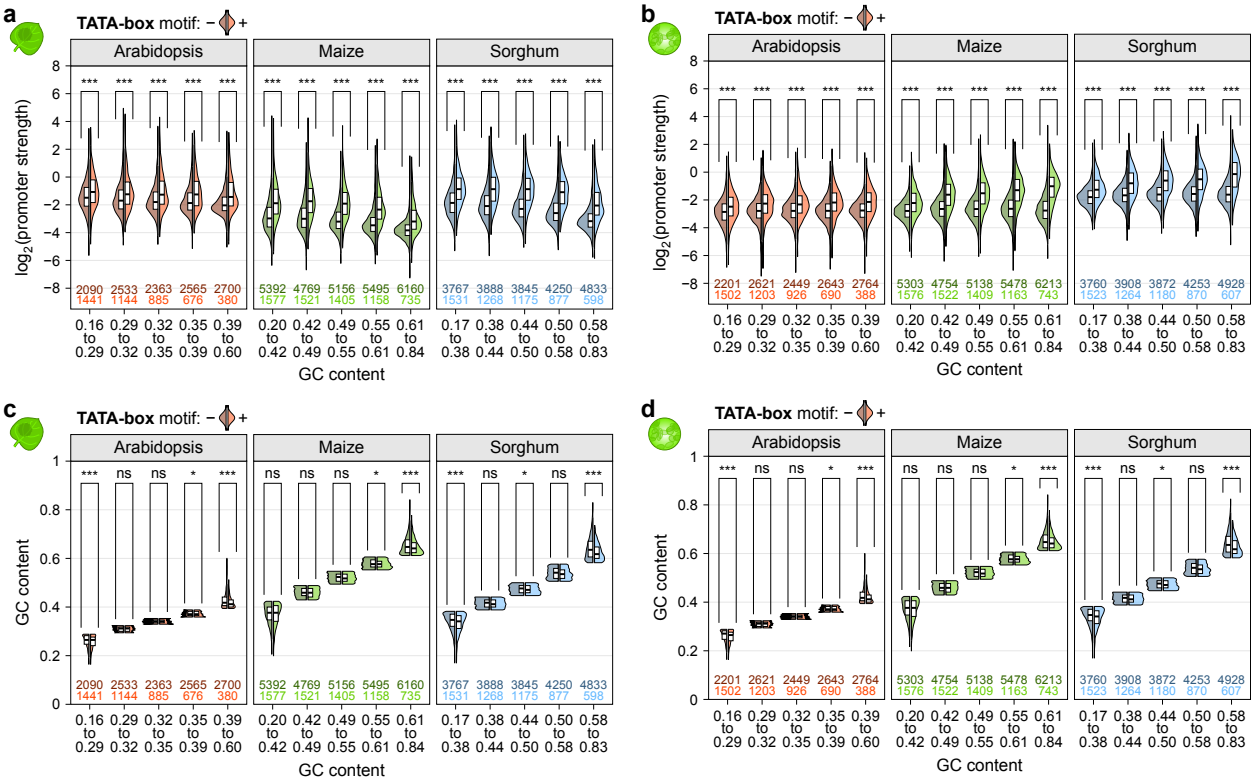
protoplasts (e). f, Comparison of the strength of sorghum promoters in tobacco leaves and maize protoplasts. g-i, Correlation of two biological replicates of STARR-seq using the Arabidopsis (g), maize (h), or sorghum (i) promoter libraries in tobacco leaves that were kept for two days in 16h light/8h dark cycles prior to mRNA extraction. Pearson's R_2 and Spearman's ρ are indicated in all plots.



Supplemental Figure 4.20.

Promoter strength in small validation libraries correlates highly with comprehensive data.

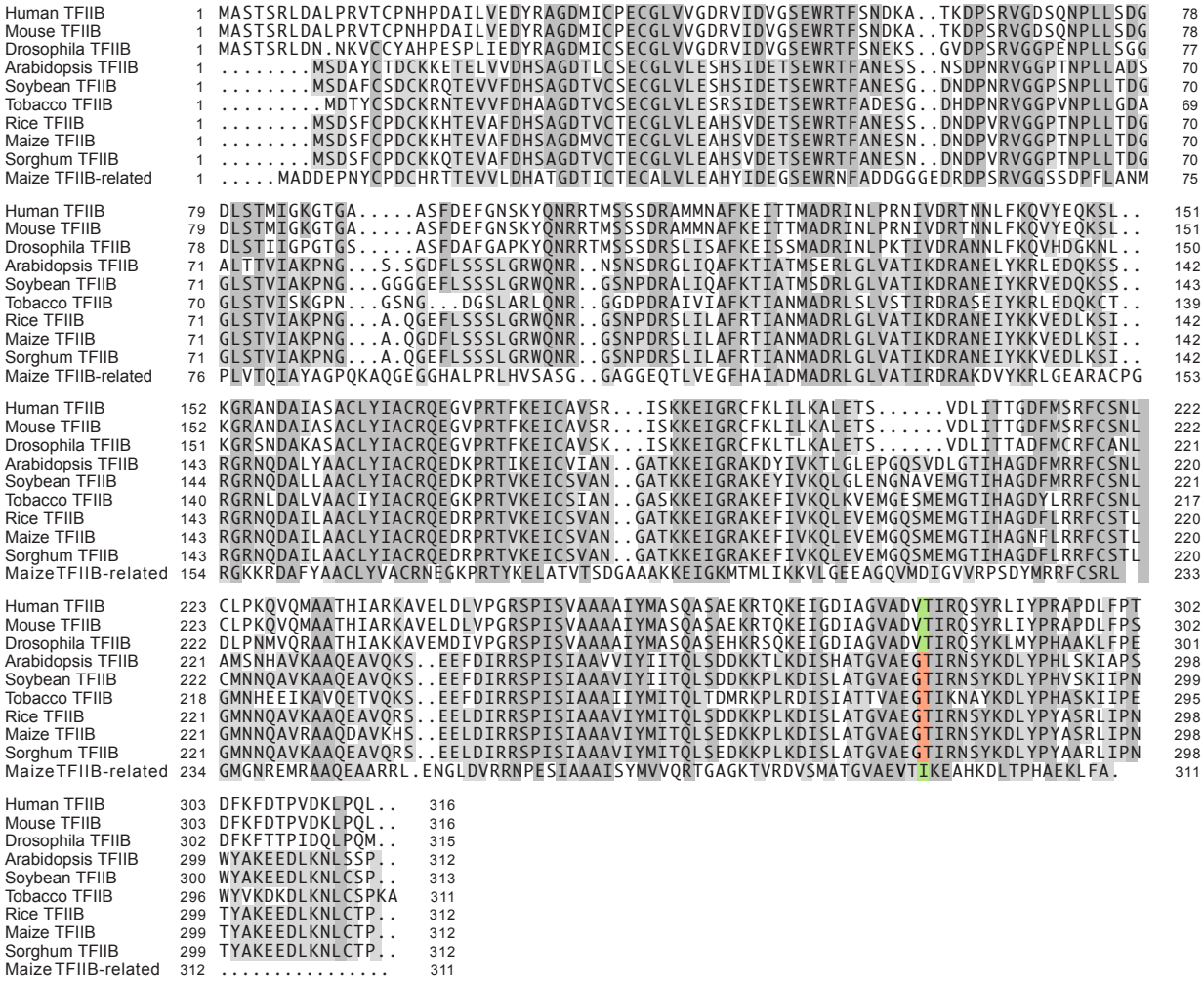
a-c, Correlation between the strength of promoters present in the comprehensive promoter libraries (main data) and in a separate, smaller validation library. The promoter strength was determined in tobacco leaves (a) and maize protoplasts (b) that were kept in the dark prior to mRNA extraction. Additionally, promoter strength was measured in tobacco leaves that were kept for two days in 16h light/8h dark cycles prior to mRNA extraction (c). d-f, As in (a-c) but for a second validation library. Pearson's R_2 and Spearman's ρ are indicated in all plots.



Supplemental Figure 4.21.

The effect of the TATA-box on promoter strength is not a result of decreased GC content.

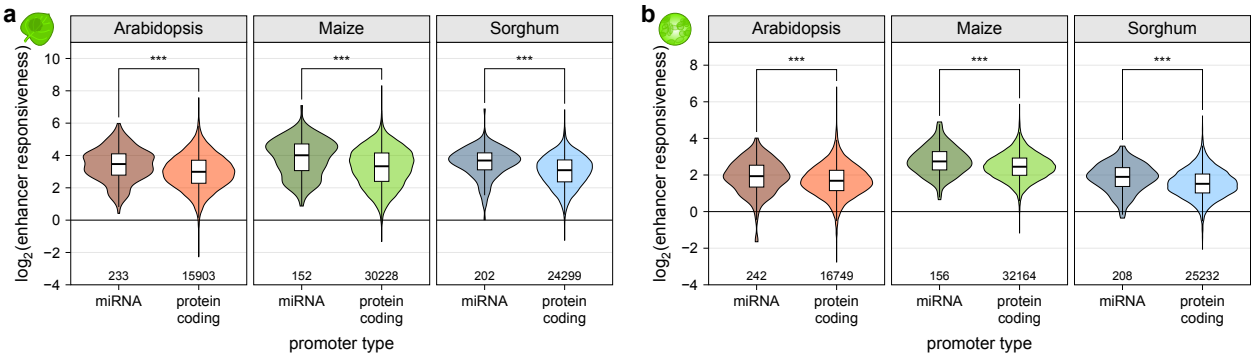
a-d, Violin plots of promoter strength (a,b) or GC content (c,d) in tobacco leaves (a,c) or maize protoplasts (b,d). Promoters were grouped by GC content and split into promoters without (left half, darker color) or with (right half, lighter color) a TATA-box. Violin plots, boxplots and significance levels are as defined in Figure 4.2. Only one half is shown for violin plots.



Supplemental Figure 4.22.

The maize genome encodes a TFIIB-related protein with a conserved valine residue required for BRE$_u$ recognition.

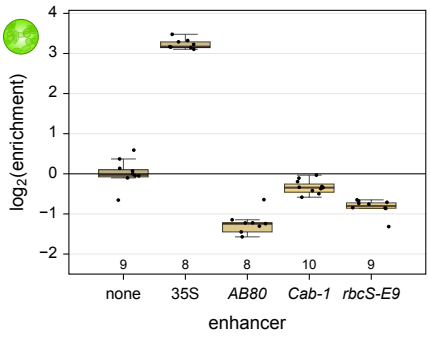
Alignment of TFIIB and TFIIB-like protein sequences from indicated species. Residues conserved in 80 or 50% of the sequences are highlighted in dark or light gray, respectively. The valine residue required for recognition of BRE$_u$ is highlighted in green.



Supplemental Figure 4.23.

Promoters of miRNA genes are more responsive to the 35S enhancer than those of protein-coding genes.

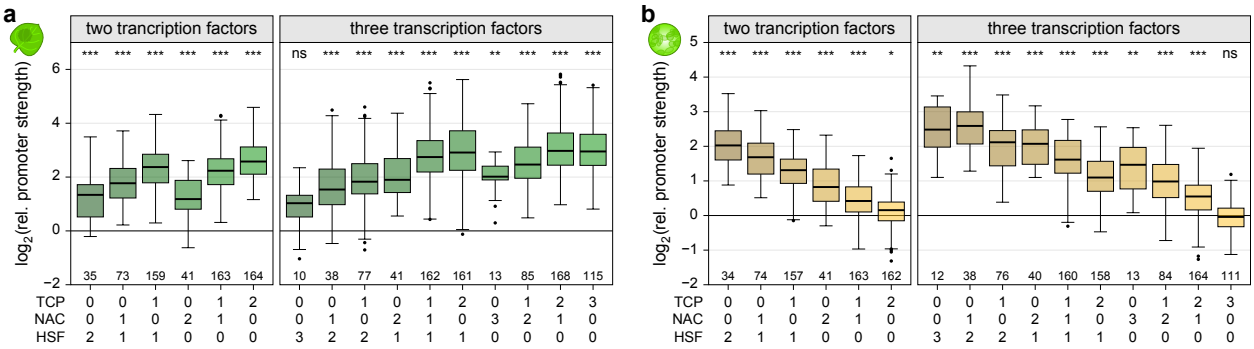
a,b, Violin plots, boxplots and significance levels (as defined in Figure 4.2) of enhancer responsiveness in tobacco leaves (a) or maize protoplasts (b). Promoters associated with miRNA or protein-coding genes are compared.



Supplemental Figure 4.24.

Light-responsive plant enhancers are not active in maize protoplasts.

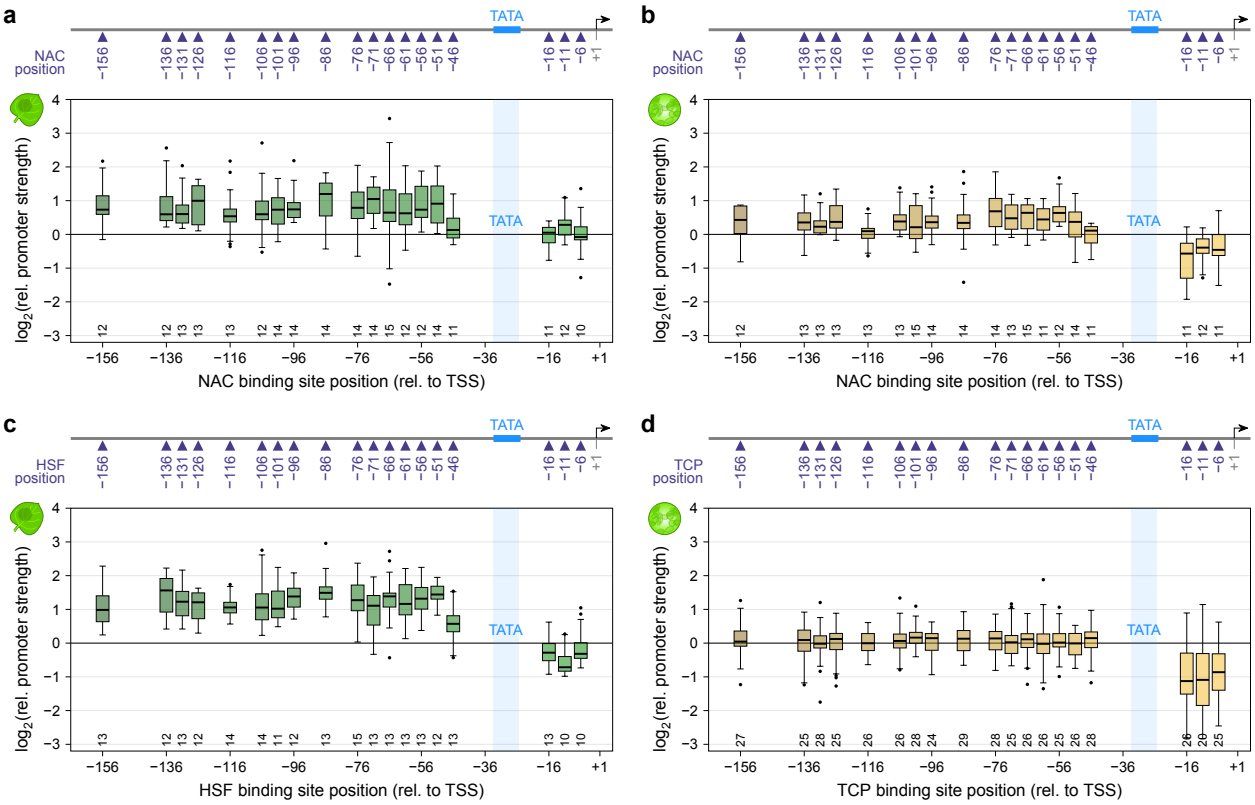
Constructs harboring no enhancer (none), a 35S enhancer, or one of three light-responsive plant enhancers (AB80, Cab-1, or rbcS-E9) upstream of the 35S minimal promoter were subjected to STARR-seq in maize protoplasts generated from dark-grown plants. Each boxplot (center line, median; box limits, upper and lower quartiles; whiskers, 1.5 × interquartile range; points, outliers) denotes the enrichment of all recovered mRNA barcodes over the DNA input. Individual data points are shown. The number of barcodes is indicated at the bottom of the plot. Only one experiment was performed.



Supplemental Figure 4.25.

Transcription factor binding sites affect promoter strength additively.

a,b, Boxplots and significance levels (as defined in Figure 4.4) of promoter strength for libraries without enhancer in tobacco leaves (a) or maize protoplasts (b) for synthetic promoters with the indicated numbers of binding sites for TCP, NAC, and HSF transcription factors. The corresponding promoter without any transcription factor binding site was set to 0 (horizontal black line).



Supplemental Figure 4.26.

Transcription factor binding sites increase promoter strength only when present upstream of the TATA-box.

a-d, A single NAC (a,b), HSF (c) or TCP (d) transcription factor binding site was inserted at the indicated position in the synthetic promoters containing a TATA-box. The strength of these promoters was measured in tobacco leaves (a,c) or maize protoplasts (b,d). Boxplots are as defined in Figure 4.4. The corresponding promoter without any transcription factor binding was set to 0 (horizontal black line).

Chapter 5. CHARACTERIZING AND CREATING LIGHT-RESPONSIVE PLANT ENHANCERS USING MASSIVELY PARALLEL REPORTER ASSAYS

5.1 INTRODUCTION

To offset crop losses from climate change[5,6] and meet the food demand of an increasing population[4], crop researchers need to improve agriculturally relevant traits. One way to do this is through engineering of plant gene expression. However, to make predictable changes in crops we need an understanding of how plants regulate themselves.

Key to controlling plant regulation of gene expression is cis-regulation, especially the interaction of enhancers with core promoters. Core promoters define the transcription start site (TSS) of a gene and allow for the initiation of transcription[42], but are insufficient to drive anything but low levels of transcription. Meanwhile, enhancers are DNA elements that interact with transcription factors (TFs), recruiting transcriptional machinery to interacting core promoters, increasing expression of a corresponding gene[42–44,156]. Enhancers confer tissue- and condition-specificity to the expression of genes they interact with independent of their position or orientation[43,45,48,49]. The evolution of enhancers has played a critical role in modulating gene expression and the development of phenotypic diversity[157,158].

The effect of an enhancer sequence on transcription can be assayed using massively parallel reporter assays (MPRAs). Our lab recently developed an MPRA to test the transcriptional effect of putative enhancers in plants by transiently transforming tobacco[33]. We utilized this technique to study light-responsive enhancers in plants. We pursued three light-responsive enhancers which

drive the expression of the genes *Cab-1*, *AB80*, and *rbcS-E9*. Each of these enhancers had been characterized to only a few hundred nucleotides required for light-responsive transcriptional activity[61,67,159]. Our goal is to identify which regions in these enhancers are necessary for light responsive transcription and use this information to better understand light-responsive regulation.

To accomplish this, we performed a deep mutational scan of each enhancer to identify the regions needed for transcriptional function. We found a set of mutationally sensitive regions necessary for the enhancer activity of the AB80, Cab-1, and rbcS-E9 enhancers. We took these mutationally sensitive regions and combined them to obtain a set of synthetic enhancers with a wide range of light-response. Then we confirmed the activity of each synthetic enhancer with our MPRA.

We modeled transcriptional activity in these synthetic enhancers using linear regression. Our model predicted well measurements in our assay and supported the billboard model of enhancers[160], in which activity associated with transcription factor binding sites combine additively. However, some combinations were not well explained by our linear model, especially those that had high expression in light conditions. This work provides a quick and comprehensive way to identify the important regions of plant enhancers. and to create synthetic enhancers of a wide range of expression or condition-specificity by combining regions from one or more known enhancers.

5.2 RESULTS

Three previously identified light-responsive plant enhancers were tested to measure their effect on transcription in our MPRA system using transient transformation in tobacco leaves that have either seen light (long day, 16H light: 8H dark) or dark (48 H dark) conditions (see methods

for more details). In our MPRA assay, we use next-generation sequencing to determine the frequency of each barcode both in the mRNA output and in the DNA input, resulting in a log2 enrichment score for each library component. This enrichment score in the tobacco leaves is a measure of the effect of tested regulatory DNA sequences on transcription (Figure 5.1 A).

Each of the three enhancers AB80, Cab-1, and rbcS-E9 had been previously characterized to a small region between 200 to 300 nucleotides in length. As well as the full length enhancers, we tested two 169 bp overlapping segments of each enhancer. Testing smaller parts of the enhancers allowed us to array-synthesize variants of each enhancer part. Synthesizing variants in the full length enhancer context was untenable due to oligonucleotide synthesis limitations.

We found each of the previously characterized light-responsive enhancers showed a light-responsive difference in enrichment in our MPRA (Figure 5.1 B-D). AB80 and Cab-1 both had an increase in enrichment in the light condition compared to their transcription in the dark condition. On the other hand, rbcS-E9 showed greater transcription in the dark condition. As a control, we tested the viral Cauliflower mosaic virus 35S enhancer. This viral enhancer is known to produce high levels of expression and is not expected to be light-responsive. In our assay we saw that this 35S enhancer showed extremely high enrichment in both the light and dark conditions. The 35S enhancer showed a slightly greater log2(enrichment) in the light condition but this preference was a smaller effect than the differences seen in our known light-responsive enhancers.

To synthesize enhancer variants for mapping variant effects, we split each enhancer into two parts each of 169 nucleotides allowing for some overlap. Part A of each enhancer represents the 169 nucleotides farthest from the gene for which the enhancer is named. Part B consists of 169 nucleotides closest to the enhancer's associated gene. Testing each shorter part of the enhancers showed that light-responsive effects were present without using the full natural enhancer (Figure

5.1 B-D). Specifically, part B of both AB80 and Cab-1 showed strong light-responsiveness. Part B of AB80 shows 16 times greater enrichment in the light condition than in the dark condition and part B of Cab-1 shows 4 times greater enrichment in light condition than dark condition. Contrarily, both parts of rbcS-E9 and the full enhancer showed greater transcription in the dark condition than in the light condition. Part A of rbcS-E9 showed more than 16 times enrichment in dark conditions than in the light. and part B showed more than 8 times enrichment than the wildtype and part B of rbcS-E9 showed high enrichment in both light and dark conditions.

After ensuring our MPRA's ability to measure light-responsive changes in enrichment, we identified which regions are necessary for transcription. To identify these sequences necessary for the enhancer function, we performed a deep mutational scan of each enhancer. Having confirmed the light-responsive effects could be seen in parts of the natural enhancer, we synthesized all single base pair insertions, deletions, and substitutions in each part of our natural enhancers. These synthesized enhancer variants were cloned directly upstream of the 35S minimal promoter in front of our barcoded reporter gene and tested in dark and light conditions as seen in Figure 5.2 A.

Three biological replicates were performed testing the library of enhancer variants (Figure 5.2 B). The replicates correlated well between replicates with Pearson's r squared between replicates between .77 and .92 within each set of enhancer fragments (Supplemental Figure 5.6). The one group that had lower Pearson's r squared between replicates is the variants of Cab-1 in the dark condition. The lower replicability is a result of Cab-1's low enhancer activity in the dark and so the measurements of enrichment are noisier in this condition.

We tested all insertions, substitutions, and deletions in part A and part B (Supplemental Figure 5.7) of each natural enhancer. We focused here on part B, which has higher enrichment in light conditions. Higher enrichment in light conditions provides a greater ability to see light-

specific deleterious effects of variants in our mutational scan since there is a larger range between the effect of the wild type enhancer and the background of the assay. We checked to see if any regions were specific to a single type of mutation (Supplemental Figure 5.8), but saw that insertions, substitutions and deletions all showed similar signatures of mutational sensitivity. We averaged over all the mutations at a nucleotide to get an average effect of mutations at that nucleotide. We smoothed this data using a rolling mean along the enhancer to identify regions of high mutational sensitivity (Figure 5.2 C).

Within each enhancer sequence there are regions that reduce enrichment when mutated. In AB80 and Cab-1 each region shows mutational sensitivity exclusively in the light condition. Comparatively, rbcS-E9 contains some regions that are mutationally sensitive in the light, some regions sensitive in both light and dark, and one region that is slightly more sensitive in the dark than in the light (R_B). Each of these light-responsive enhancers contains between 3-5 regions of mutational sensitivity. Many of the regions of high mutational sensitivity overlapped with predicted transcription factor binding sites (TFBS) in Cab-1 and rbcS-E9 (Supplemental Figure 5.9). However, AB80's mutationally sensitive regions had no overlap with our set of predicted TFBSs, as well as mutationally sensitive regions C_D and R_D which both showed strong mutational sensitivity (see methods).

Using our data from our mutational scan we identified regions of mutational sensitivity that we wanted to further explore (Figure 5.2 C). These mutationally sensitive regions ranged in size from 17 base pairs regions covering a single peak of mutational sensitivity to 47 base pairs covering three peaks of mutational sensitivity. We synthesized these mutational regions with a small 6 base pair overlapping region and randomly ligated these fragments together to create many combinations of mutationally sensitive regions. These combinations consist of 1 to 3 mutationally

sensitive regions and each combination can consist of regions from more than one enhancer. As controls, we randomized the nucleotides in the A_D region (A_D_rand) chose a region of Cab-1 that had low mutational sensitivity (C_Con).

After testing these combinations of fragments, we found the correlation between biological replicates to be between 0.83 and 0.95 (Pearson's r squared). We did see some skew between replicates with certain replicates showing higher expression overall. We attribute the skew either to differences in overall health of some of the plants between two biological replicates or to technical differences in infiltrating the tobacco leaves with agrobacterium.

Our experiment contained combinations covering a large fraction of possible combinations of three or fewer regions (Supplemental Figure 5.10). Within these synthetic combinations we saw a wide range of light-responses (Figure 5.3 B). This includes combinations that are active in both light and dark, show no activity, and that are active in a light- or dark-responsive manner. However, most the combinations showed low log2(enrichment) even though they consist mostly of pieces that are mutationally sensitive in their native context.

We used DESeq2[161] to identify fragments that changed significantly between light and dark sequences or didn't show significant change (Supplemental Figure 5.11). We used this data to identify four groups: light-responsive, dark responsive, light-insensitive and highly expressed, and light-insensitive and lowly expressed. For each of these groups we picked a hundred sequences and validated them in a smaller library to ensure that the effects we saw were independent of the library content. Additionally, we included part B of each natural enhancer as a comparison (Figure 5.3 C).

The smaller validation library replicated the log2(enrichment) seen in the original library context as shown in Figure 5.3 D. Many of the sequences that we tested had enrichments even

greater than our natural enhancer. We also saw a group of enhancers that had the same level of transcription as Cab-1 and AB80 but with even greater transcription in the light. Surprisingly, we saw one combination that had greater transcription than the 35S enhancer, one of the strongest known plant enhancers.

We wanted to identify which mutational regions were important in affecting the enrichment of a combination of mutationally sensitive regions. To do this we used linear-regression to model our log2(enrichment) data. For each light, dark, and for light response (log2(light enrichment/dark enrichment) we trained each model on 80% of our measured combinations. To evaluate the model, we used the trained model to predict the held out 20% of our data and computed an R^2 value between the predictions and the true measurements of the held-out set. The linear regression model we created takes into account both which mutational regions are present in a combination and how many of each region appear in the combination.

The comparison of the predictions of the linear regression model on the held-out test set with the measurements of those combinations are graphed in Figure 5.4 A, B, and C. We can see that our model predicts the dark condition extremely well with an R^2 between predictions and measurements of 0.89. We perform less well when predicting the light condition. Particularly, we often underestimate the highest log2(enrichment) combinations. This error is compounded when predicting light-response. Not only do we underestimate the most light-responsive combinations but we also have compounded the noise in our measurements by combining the light and dark enrichments.

We looked at the coefficients associated with each mutationally sensitive region in our models to identify which mutationally sensitive regions drove high log2(enrichment) scores in the light and dark (Figure 5.4 B). We saw that the included controls did not drive expression in our

model. The control region from Cab-1 (C_Con) was not associated with an increase in transcription while the randomized region from AB80 (A_D_rand) had a slight negative effect in both light and dark conditions. We did see a set of mutationally sensitive regions that drove transcription independent of light. These regions mostly came from rbcS-E9 with R_A+B+C and all sub pieces of it driving log2(enrichment) values both in the light and in the dark. Activity in these regions corroborates what is seen in the mutational scan, that these regions drive transcription both in the light and in the dark. Unsurprisingly, the regions that overlap more than one region of mutational sensitivity drive greater log2(enrichment) values. This is apparent with the combinations of all three sections from the rbcs9-E9 enhancer (R_A+B+C) being the greatest driver of log2(enrichment) followed by the combination of the A and B regions from rbcs9-E9 (R_A+B) and finally the set of individual mutational sensitive regions (R_A, R_B, and R_C).

We did see regions that drive transcription in our data in the light but not in the dark. Most notably we see that the highest regions that had the highest light-dependent mutationally sensitivity in our mutational scan (A_B+C, C_B+C, C_D+E, and R_D+E) drive predictions of log2(enrichment) more in the light than in the dark. Of the four regions, only C_B+C is associated with a positive effect on enrichment in the dark while the other three are associated with a decrease in log2(enrichment) compared to the no-enhancer control in the dark.

To better identify what drives the expression in our MPRA we also tested two other linear regression models. One is a presence absence model that is provided only whether a mutational region is present in a combination. The second is provided both which mutational region is present and the order of mutational regions in comparison to the 35S minimal promoter. To allow a comparison of the R^2 values between models we bootstrapped all the models. For each bootstrap we took different random samples of 80% of our data to use as the training and the remaining 20%

to use as the test set. We then similarly trained our model, predicted on the test set and calculated the R^2. Differences in the predictive power between the different linear regression models were significant however small. Specifically, having information of the number of mutationally sensitive regions present is more predictive than a presence absence model. Additionally, knowing the position of each region relative to the 35S minimal promoter improves the predictive power of the model. These effects seem to hold true regardless of whether you are predicting the light log2(enrichment), dark log2(enrichment), or the light-response variable. Regardless of the model we see that we best predict the dark enrichment followed by the light enrichment and finally light-response.

While much of the activity of combinations of mutationally sensitive regions is captured in our linear model we know there are further complexities. One complexity is that the order of regions can matter to the enrichment of a combination. When combining a region with little mutational sensitivity in the light (C_Con) with a region that is mutationally sensitive (C_B+C) we see that one combination of regions (C_B+C + C_Con) has higher activity than the reciprocal combination (C_Con + C_B+C) (Figure 5.5 B). This is not explained by proximity to the 35S minimal promoter since the last region of the combination is inserted next to the minimal promoter. Indeed, the expected active region (C_B+C) is farther from the promoter in the more active combination (C_B+C + C_Con).

Additionally, we know that the spacing between regions of mutational sensitivity matters to the overall activity of a synthetic combination. When looking at two regions that are mutationally sensitive R_A and R_B from rbcS-E9, we see that the combination of the two regions with a 6-nucleotide linker (R_A + R_B) increases the enrichment compared to either region alone (Figure 5.5 D). However, the combination (R_A + R_B) does not lead to levels of enrichment seen

when testing the overlap of the two regions with spacing (R_A+B). The effect of spacing implies that the distance between two mutationally sensitive regions can influence expression.

5.3 DISCUSSION

Using a plant based MPRA coupled with a deep mutational scan of the nucleotides in each of three known light-sensitive enhancers we identified regions that are necessary for transcription. We found that each of our natural enhancers is composed of multiple regions of mutational sensitivity, each of which contributed to the full function of the enhancer (Figure 5.2 B).

We combined regions that were mutationally sensitive and saw a wide range of transcriptional effects (Figure 5.3 B). The combinations we created showed light-responsive transcription along a gradient. Some combinations showed higher transcription in the light and similar transcription in the dark than the natural enhancers (Figure 5.3 C). The method shown here of quickly creating a large range of synthetic enhancers in plants from a small set of known condition specific enhancers is especially useful to plant regulatory engineering. As more condition-specific enhancers are characterized, our method will allow the quick development of enhancers to drive transcription at relevant levels in response to a wide range of stimuli.

We modeled the combinations of enhancers that we created using linear regression. Our linear regression models predicted well the transcriptional effect of combinations of mutationally sensitive regions on a held out set of enhancers (Figure 5.4 A). The linear regression models performed particularly well when predicting transcriptional effects in the dark condition. Better predictions for dark enrichments are probably due to our system having only a few regions associated with high transcription in dark conditions. Our linear regression models perform less well when predicting light expression. Specifically, at the highest levels of transcription in the light

condition they predict less transcription than is measured (Figure 5.4 A). This may be due to synergistic effects between enhancer regions that drive enrichment in the light.

When comparing types of linear regression models, we saw that adding information about the combinations increased the predictive power of the model (Figure 5.4 C). A model based on the presence or absence of mutationally sensitive regions predicted expression less well than models that took into account the number of mutational regions. Additionally, giving a model the position of order of each region from the 35S minimal promoter further improved predictions of transcription. However, the benefits in predictive power of increasing the complexity of a linear regression model were small which may be due to only a subset of combinations benefiting from the more complex models. We did not use more data intensive models because we had only on the order of 2000 measured synthetic combinations to train our model (Supplemental Figure 5.10 A).

The transcriptional effect of natural and synthetic enhancers replicate well between biological replicates in the transient tobacco system. However, before using these enhancers for real world applications, it is important to know how well the outputs of our MPRA translate to plants. We know that the agrobacterium-mediated tobacco model is under pathogen stress and so may be particularly sensitive to enhancers upregulated in pathogen response. To test if our results replicate well in plants we will create and assay stable transformants carrying reporter genes driven by enhancers sequences of interest.

5.4 METHODS

Library creation

We created libraries for performing STARR-seq in tobacco similarly to those in Jores et al., 2020. We started with the plasmid pPSup (Addgene 149416). This plasmid contains resistance genes, phosphinothricin resistance gene (BlpR) and a resistance gene conveying spectinomycin

resistance (SmR). We used GFP as a reporter gene that was terminated by the polyA site of the *A. thaliana* ribulose bisphosphate carboxylase small-chain 1A gene in the transfer DNA region. We used golden gate[95] to clone the 35S core promoter, a synthetic 5′ UTR synJ[94] (ACACGCTGG AATTCTAGTATACTAAACC), an ATG start codon and a 15-bp random barcode (VNNVNNVNNVNNVNN) in front of the second codon of our reporter GFP. We array-synthesized sequences of interest and cloned these into our backbone directly upstream of the 35S core promoter using golden gate cloning sites. We bottlenecked each library to contain on average 10-20 barcodes per tested sequence. The plasmid libraries were introduced into *Agrobacterium tumefaciens* GV3101 strain harboring the helper plasmid pSoup[93] by electroporation.

We grew Tobacco (*Nicotiana benthamiana*) in soil (Sunshine Mix no. 4) at 25°C in a long-day photoperiod (16 h light and 8 h dark; cool-white fluorescent lights [Philips TL-D 58W/840]; intensity 300 μmol/m^2/s). We transformed plants 3 to 4 weeks after germination. For transformation, we diluted an overnight culture of *A. tumefaciens* 1:10 into 100 mL YEP medium (1% [w/v] yeast extract, 2% [w/v] peptone) and grew it at 28°C to an OD of ~1. We took a 5 mL input sample of the agrobacterium cells transformed with our library and from it isolated plasmid which was sequenced to identify the beginning frequency of barcodes in our library. The cells were spun down and resuspended in 100 mL virus induction medium (M9 medium supplemented with 1% [w/v] Glucose, 10 mM MES, pH 5.2, 100 μM $CaCl_2$, 2 mM $MgSO_4$, and 100 μM acetosyringone). After overnight growth, we then resuspended in infiltration solution (10 mM MES, pH 5.2, 10 mM $MgCl_2$, 150 μM acetosyringone, and 5 μM lipoic acid) to an OD of 1 and infiltrated into the first two mature leaves of six tobacco plants. I held back 5 mL of the resuspended agrobacterium and from it isolated plasmid, which I sequenced to identify the frequency of barcodes in our library at the beginning of the assay. We further grew the plants for

48 hours under normal conditions (16 hours light 8 hours dark) or in the dark prior to mRNA extraction.

STARR-seq Assay

We performed three biological replicates for both the mutated variants (Figure 5.2) and the combinations of mutationally sensitive fragments (Figure 5.3 B), for each large library of ~500,000 sequences. For smaller libraries including the natural enhancers (Figure 5.1), validation of mutated variants (Supplemental Figure 5.6), and validation of combinations of mutationally sensitive fragments (Figure 5.3 C) we performed two biological replicates. Each biological replicate used tobacco plants that were germinated in different weeks. To perform our assay, we harvested tobacco leaves two days after agrobacterium infiltration which had either continued to see 16 H light, 8 hours dark or had been transferred to a dark incubator for 48 hours. After harvesting, we immediately froze them in liquid nitrogen. We then ground using a mortar and pestle and resuspended ground material in 24 mL of Trizol (company here) per 12 leaves. We pelleted debris by centrifugation (4000 x g, 5 min, 4°C) and transferred the supernatant to a new falcon tube. We added 5 mL of chloroform to the supernatant, vortexed the mixture for 30 seconds and centrifuged (4000 x g, 15 min, 4°C). We recovered the aqueous layer and again mixed with 5 mL of chloroform, vortexed (30 sec), and centrifuged (4000 x g, 15 min, 4°C). We then transferred 10 mL of the aqueous layer to a new falcon tube and mixed with 10 mL of isopropanol by inversion. We then added 10 mL of high salt buffer (0.8 M sodium citrate, 1.2 M NaCl) mixed by inversion, incubated at 25°C for 15 minutes, and pelleted nucleic acids by centrifugation (4000 x g, 30 min, 25°C). We washed the pellet with 25 mL of cold 70% (v/v) EtOH and centrifuged (4000 x g, 5 min, 4°C). Then we dried the pellet and resuspended RNA in 2400 μL of DEPC treated water. To isolate mRNA, we used two aliquots of 150 μL magnetic Oligo$(dT)_{25}$ beads (Thermo Fisher

Scientific) per 12 leaves, following the manufacturer's protocol. We resuspended each of the two aliquots of mRNA into 40 µL of 10mM Tris which we pooled for a DNAse I treatment (80 µL mRNA solution, 10 µL DNAse I buffer without MnCl2, 10 µL 100 mM MnCl2, 1 µL RNase OUT, 2 µL DNAse I) and incubated at 37°C for 1 hour. To precipitate RNA, we added 10 µL 8 M LiCl, 1 µL glycogen and 250 µL 95% (v/v) EtOH and incubated at -80°C for 15 minutes. We pelleted the solution (20,000 x g, 30 min, 4°C) and washed with 200 µL cold 70% (v/v) EtOH before being centrifuged again (20000 x g, 5 min, 4°C). We inverted the pellets, dried, and resuspended them in 100 µL DEPC treated water. We made cDNA from the mRNA by using SuperScript IV reverse transcriptase (Thermo Fisher Scientific) and a GFP specific primer (GAACTTGTGGCCGTTTAC G) according to the manufacturer's protocol. For each sample, we performed 8 reactions using 5 uL of mRNA per reaction. We added reverse transcriptase to half of the reactions while leaving out the reverse transcriptase enzyme from the other reactions as a control to check for DNA contamination in our extraction. After confirming low levels of DNA contamination, we combined cDNA reactions and purified them using DNA Clean and Concentrator-5 columns (Zymo Research). We then amplified the barcodes associated with our tested sequences with between 10-20 cycles of qPCR and sequenced the barcodes using next generation sequencing.

Computational analysis

We analyzed the MPRA experiments by counting the reads for each barcode in both the experimental cDNA samples and the input DNA samples. Barcode counts below five were discarded. For Figure 5.1 B, C, and D barcode enrichments were calculated by dividing the barcode frequency (barcode counts divided by all counts) in the cDNA sample by that in the input sample. For Figure 5.2 B, C and Figure 5.3 B, C, and D information from all barcodes associated with a tested element was aggregated. To determine enrichment of an element, the frequency of the sum

of all barcodes in the DNA input was divided by the frequency of the sum of all barcodes in the RNA output. Welch's two-sided t-tests were performed for Figures 5.1 B, C, and D as well as Figure 5.4 D and were multiple test corrected using a bonferroni correction. We calculated the positional mean at each position in Figure 5.2 B by taking the rolling mean of each variant with a 12 bp window around each position along in the tested fragment.

We identified synthetic combinations that changed expression in light or dark conditions using DESeq2[161]. Expression changes by light condition informed the subset of synthetic combinations that we validated in a smaller library (Figure 5.3 D).

Linear regression models in Figure 5.4 were fit on 80% of the measured combinations of light enhancer fragments using the lm() function in R v.4.0.2. We developed three linear regression models developed in our analysis. One was a presence absence model which only considered if a mutational region was present or not. The second was a model considered the mutational region and the number of regions but not the position of the region. Finally, we had a positional model which estimated the effect of each mutational region at either position 1, 2, or 3 in our combination. In the positional linear regression model, position 1 refers to the position closest to the 35S minimal promoter. To determine how well the model predicts log2(enrichment) of our synthetic combinations we used the linear regression models to predict the log2(enrichment) scores for the held out 20% of the dataset and calculated the R^2 between the measured and predicted values. In Figure 5.4 D, E, and F we resampled 80% of our data 100 times to make sure that differences between the model types were not due to the data used for training. The computational analysis as well as all data needed to replicate the analysis are on bitbucket (https://bitbucket.org/jackrtonnies/finalized_paper/).

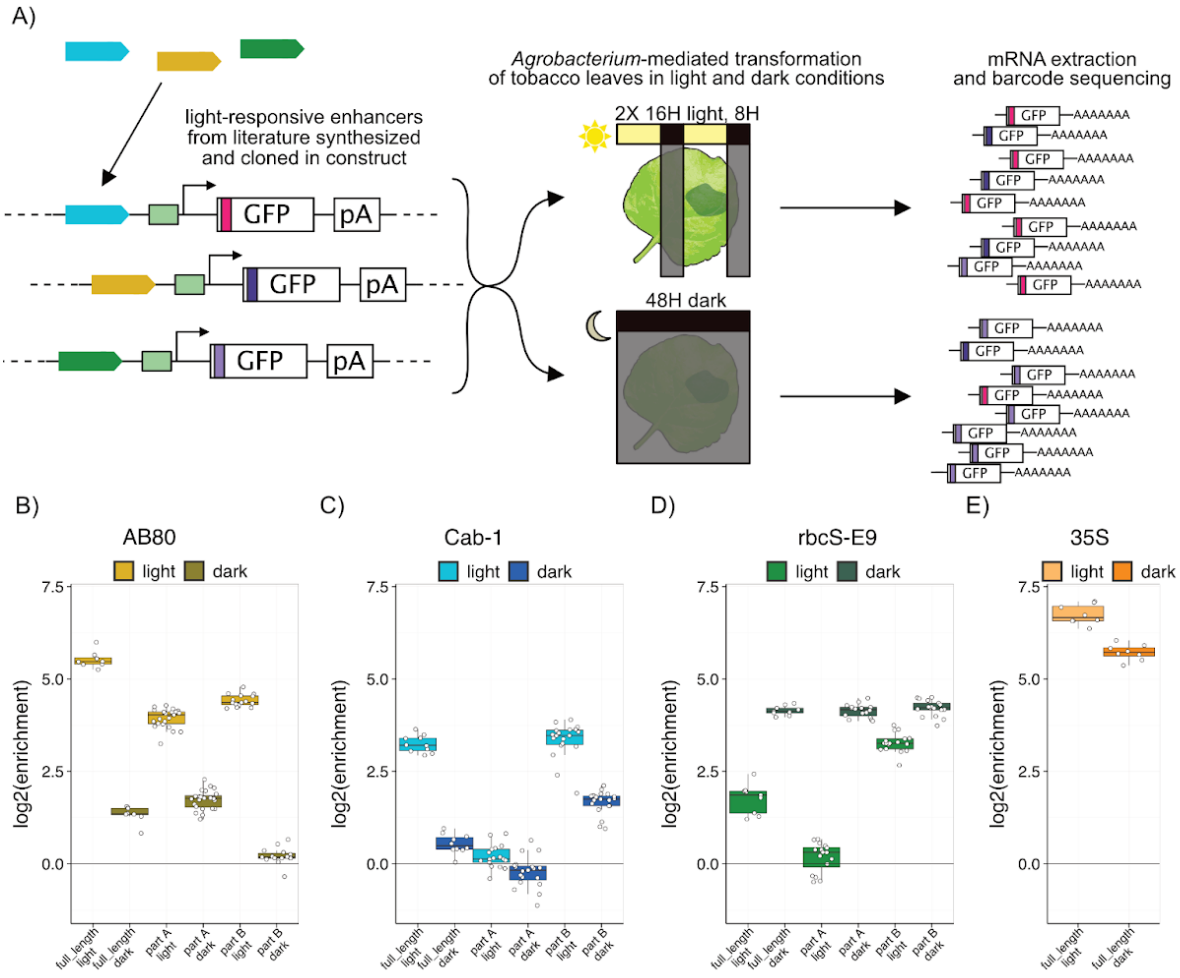


Figure 5.1.

A). Experimental design using our MPRA to test the transcriptional effect of natural light-responsive sequences (blue, green, and yellow). The enhancer sequences are inserted in front of the 35S minimal promoter (light green) and drive transcription of GFP containing barcodes (pink and purples). pA, poly-adenylation site. Figure panels 1 B, C, and D show the log2(enrichment) of each tested light-responsive enhancer from the literature and two overlapping synthesized parts of each enhancer of length 169 nucleotides. All measurements in these graphs are normalized to a no-enhancer control set to 0.

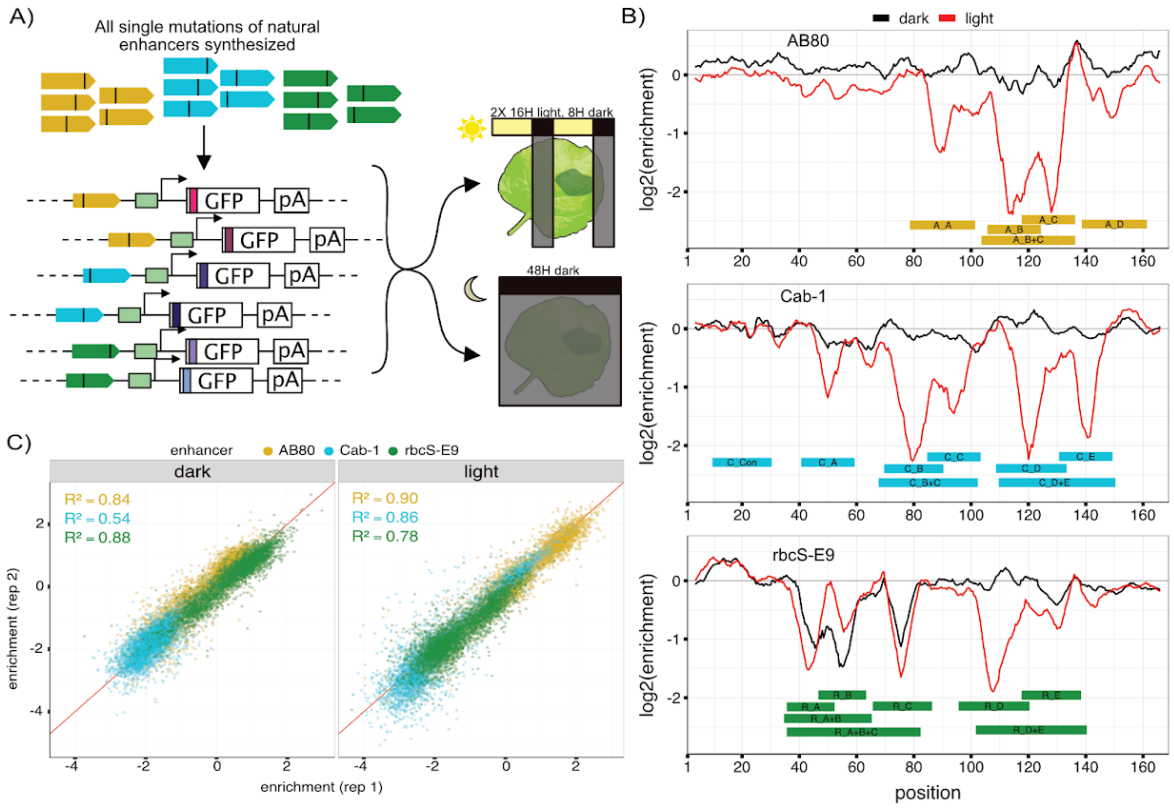


Figure 5.2.

A). Experimental design of the mutational scanning experiment to test the light-responsive expression of single base pair mutations by measuring resulting enrichment in light and dark conditions. B). The average positional effect of a mutation at each position along part B of wild type enhancers. Each line represents the effect of mutations compared to the wild type enhancer in light (red) or dark (black) conditions. In each graph, 0 is the log2(enrichment) of the wildtype enhancer part B. We identified and named these regions with the first letter of the enhancer they originate from and an identifying second letter. C). The comparison between two biological replicates of the enhancer variants library. Pearson's r squared (R^2) was calculated for each group of sequences associated with an enhancer.

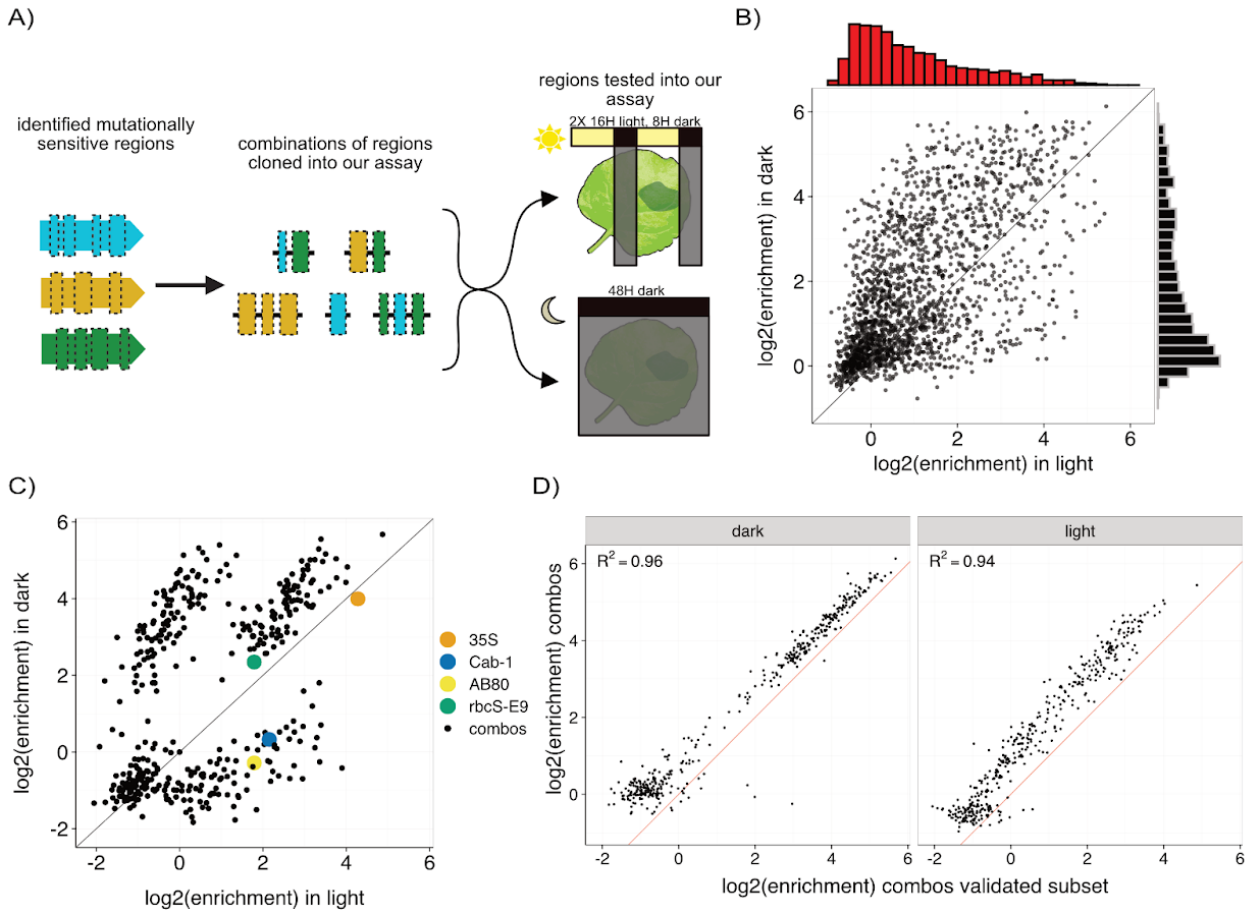


Figure 5.3.

A). Experimental design of combining the identified mutationally sensitive regions and testing them in light and dark conditions using our MPRA. B). The log2(enrichment) of each combination of mutationally sensitive regions in both light and dark conditions normalized to a no-enhancer control. Along each axis is a histogram showing the fraction of combinations at each enrichment range. C). Validation sequences showing log2(enrichment) of combinations compared to part B of natural enhancer sequences normalized to a no-enhancer control. Controls spiked in include part B of Cab-1, AB80, and rbcS-E9 as well as the 35S enhancer. D). Comparison of the log2(enrichment) of validation sequences in the original experiment and validation experiment.

Pearson's R^2 between log2(enrichment) comparing sequences in the original library and the validation set.

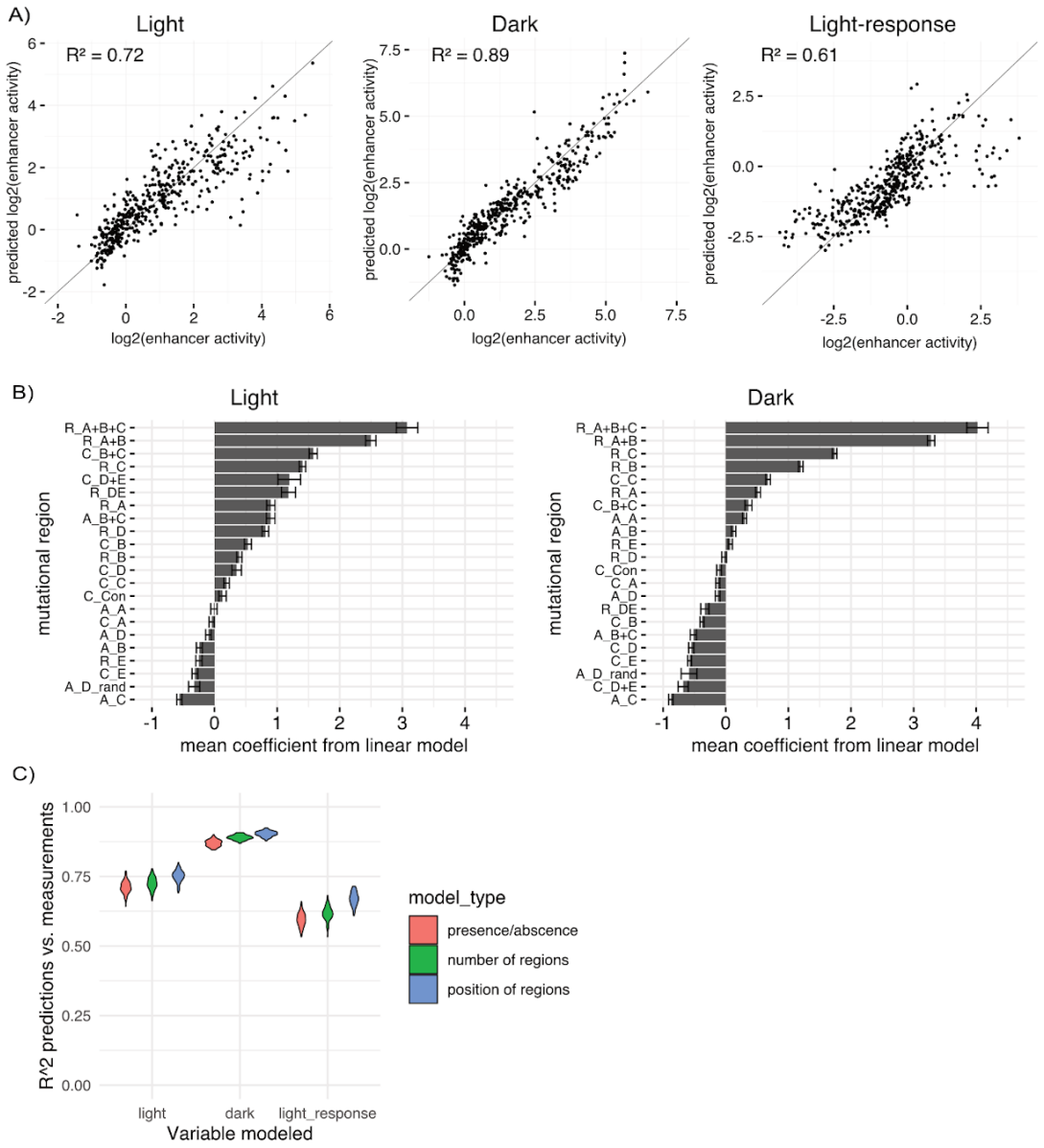


Figure 5.4.

A). Predictions of a linear regression model compared to measured values of a held out test set of 20% of our measured synthetic combinations. A linear-regression model given the number of each mutationally sensitive region in a combination of mutationally sensitive regions was trained on data from our MPRA. Separate linear regression models were trained to predict the log2(enrichment) of combinations in the light condition, dark condition, and light-response

(log2(light enrichment/dark enrichment)). B). The coefficients extracted from the linear regression models shown from Figure 5.4 A. 100 linear-regression models were trained with different random samples of the training and test data. The error bars represent two standard deviations away from the mean of all the coefficients extracted from 100 models. C). The coefficient of determination (R^2) of three different linear regression models on a held out test set. Each violin plot consists of 100 bootstrapped calculations of the R^2 between predicted log2(enrichment) of held out data and the true measurement.

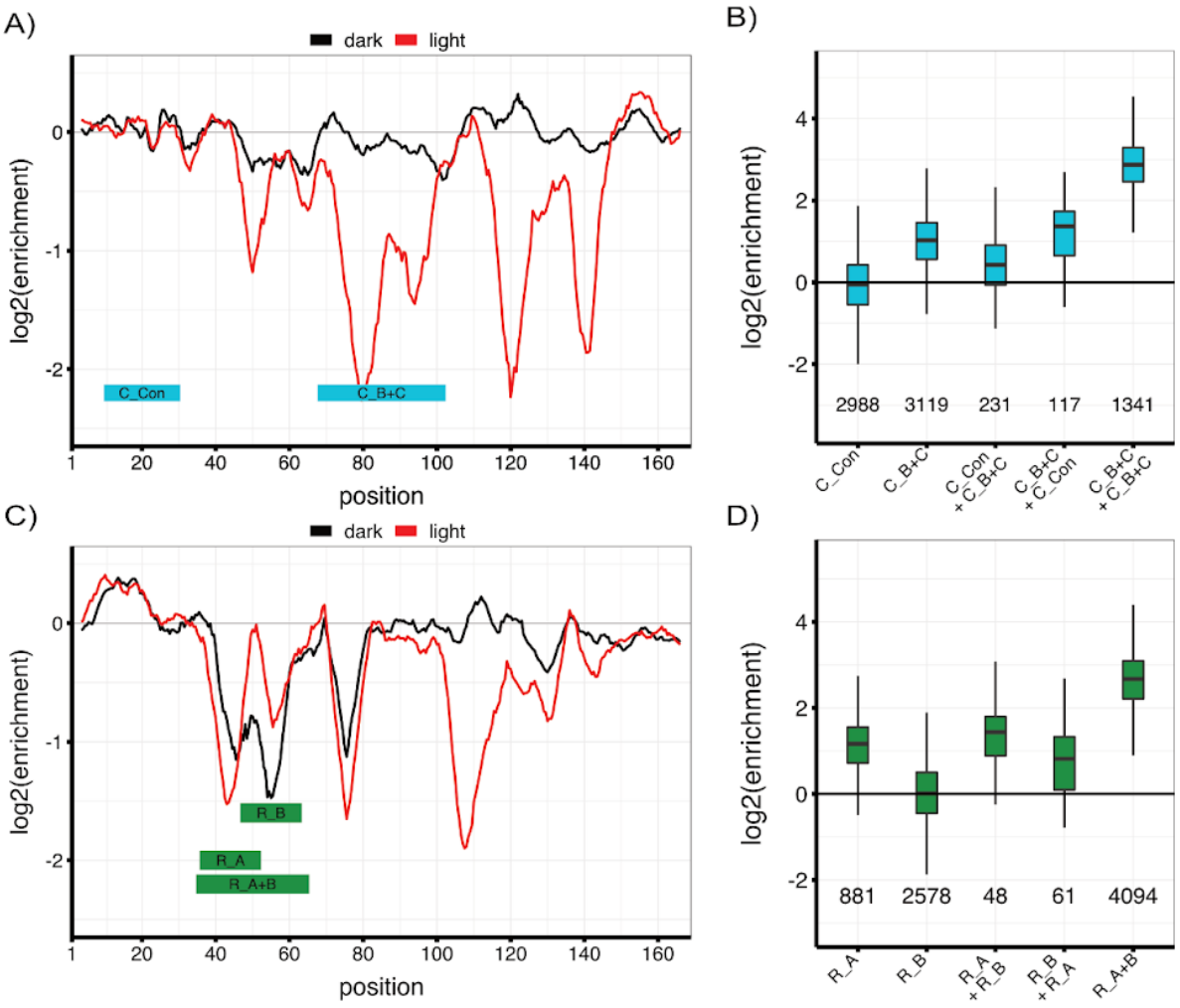
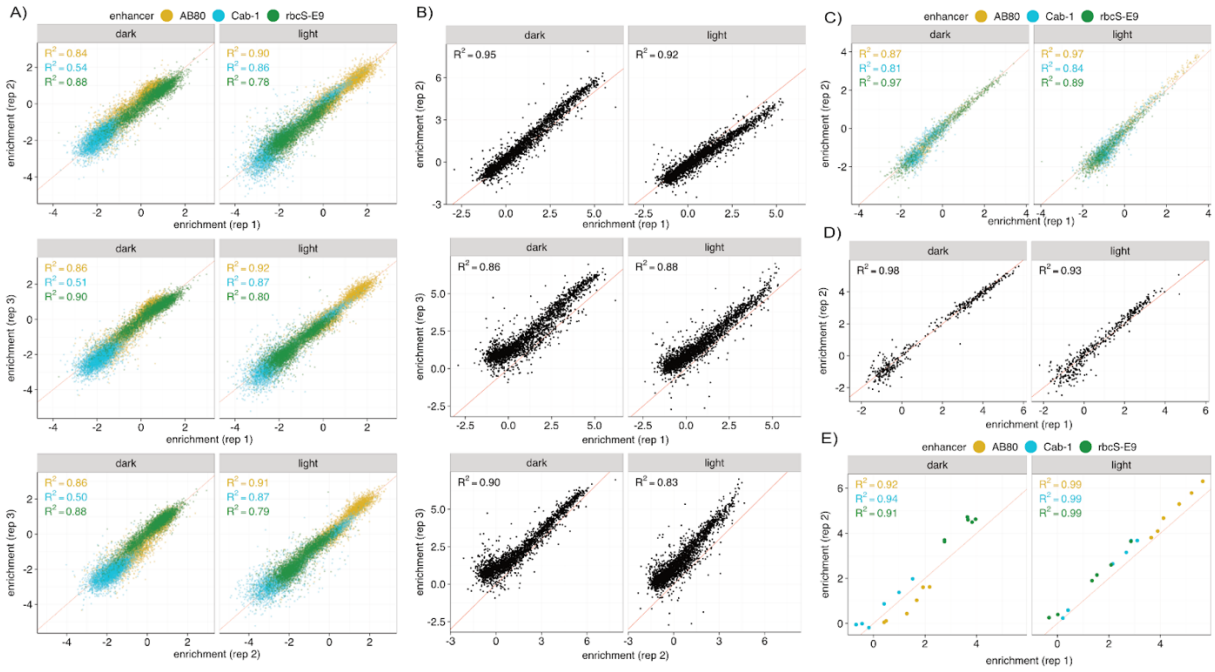


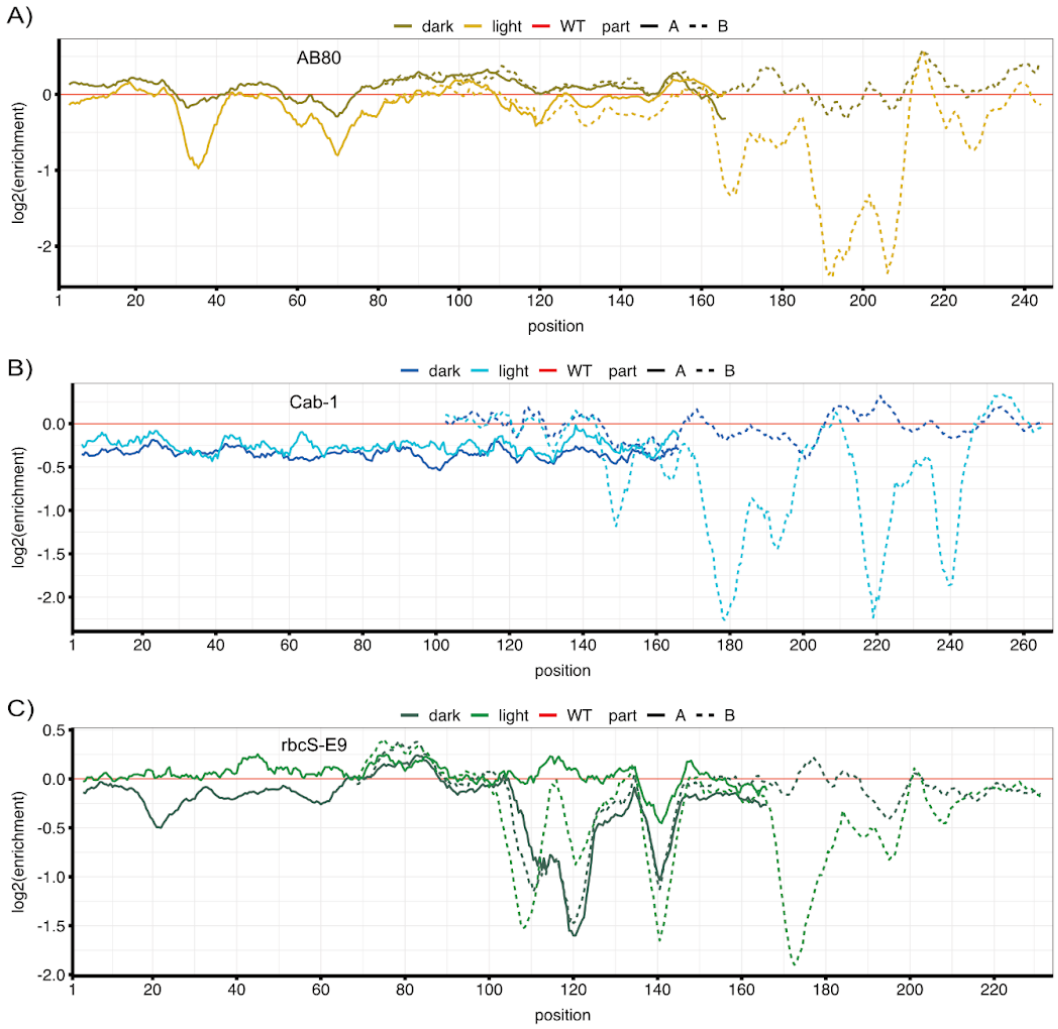
Figure 5.5.

A). Mutational scan as seen in Figure 5.2 B highlighting two regions specific regions of mutational sensitivity in the enhancer Cab-1. B). Boxplot representing the mean enrichment in the light between replicates for each barcode linked to synthetic combinations of the regions seen in Figure 5.5 A. C). Mutational scan as seen in Figure 5.2 B highlighting two regions specific regions of mutational sensitivity in the enhancer rbcS-E9. D). Boxplot representing the mean enrichment in the light between replicates for each barcode linked to combinations of the regions seen in Figure 5.5 C.



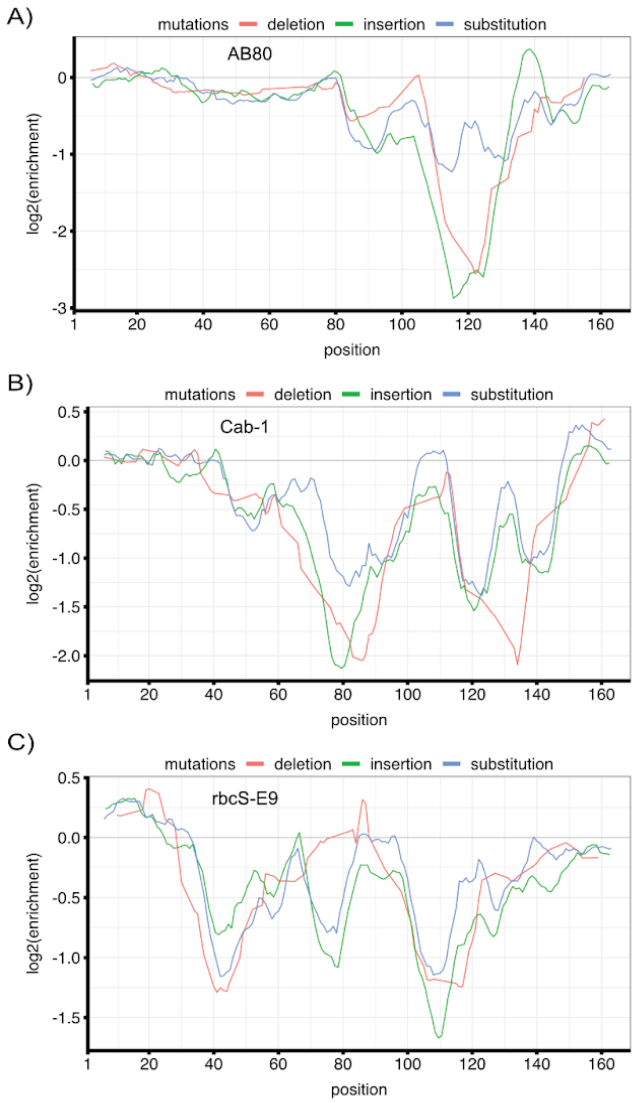
Supplemental Figure 5.6.

A). Correlations between biological replicates for each the mutational library of enhancers AB80, Cab-1, and rbcS-E9. Each dot represents the mean of all barcodes encoding for a single enhancer variant. Pearson's r-squared (R^2) is reported for each group of variants. B). Correlations between biological replicates for the combinations of mutationally sensitive regions. C). Correlations between biological replicates for the validation of select mutations from the mutational library. D). Correlations between biological replicates for the validation of select combinations of synthetic combinations. E) Correlations between replicates of the full length and each part of the wild type enhancers.



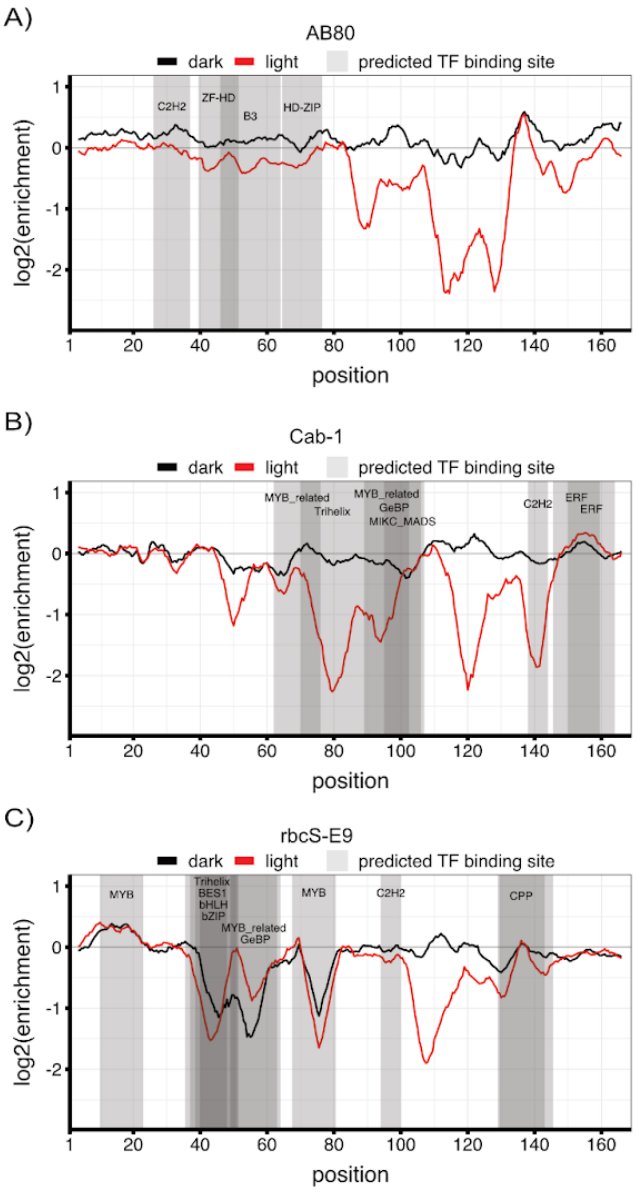
Supplemental Figure 5.7.

A). Mutational sensitivity showing the overlap of parts A and B for the enhancer AB80 B). Cab-1 C). rbcS-E9.



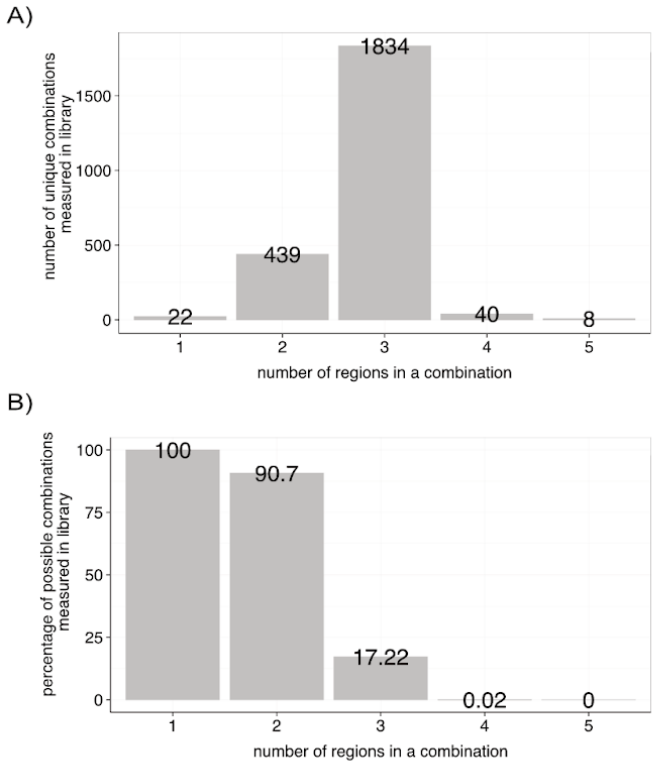
Supplemental Figure 5.8.

A). Mutational sensitivity when determined by a subset of mutations either deletions, insertions, or substitutions in AB80. B). Cab-1 C). rbcS-E9. The mutational sensitivity at each position is calculated using a rolling mean as in Figure 5.2.



Supplemental Figure 5.9.

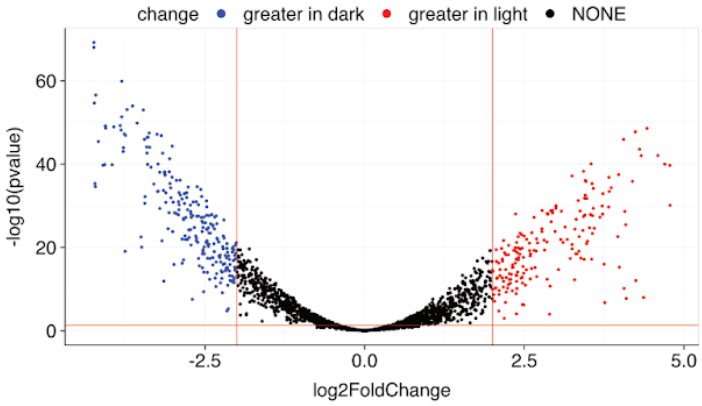
A) Overlap of called motifs with mutationally sensitivity seen in enhancer fragments part B of AB80. B). Cab-1 C). rbcS-E9.



Supplemental Figure 5.10.

A). The number of unique combinations measured in our library of synthetic combinations of mutationally sensitive regions separated by the number of regions combined. B). The percentage of possible combinations covered in our library separated by the number of regions combined.

A)



Supplemental Figure 5.11.

A). Determining whether synthetic combinations of enhancers change expression in response to light using DESeq2 determined p-values. P-value cutoff = 0.05, fold change cutoff = 4.

Chapter 6. CONCLUSIONS AND FUTURE DIRECTIONS

During my graduate school career, I have created tools to better understand plant regulation and worked on characterizing and creating plant regulatory elements. The maize mesophyll protoplast system that I optimized has allowed for library scale transformations in crop species. This tool is already in use in multiple other projects in our lab, including identifying plant insulator elements and characterizing terminator sequences in plants.

Work creating synthetic light-responsive enhancers will be continued in the lab. Future experiments are planned to test these synthetic enhancers *in planta.* My method to quickly create a wide variety of synthetic enhancers from a set of known enhancers will hopefully enable future scientists to create a variety of condition specific enhancers for bioengineering purposes.

While all my projects were interesting, I find highly translational science most compelling. I get excited when learning about projects that will directly affect plant phenotypes and will change plant cultivation systems within a matter of years. This insight has led me to transition to the agritech industry. To this end, I have accepted a job with Sound Agriculture to work on creating new phenotypes in plants. Specifically, we will methylate regions of the plant genome in seeds. This methylation will knockdown expression of genes likely to affect agriculturally relevant phenotypes. By testing the effect of gene knockdowns within a single generation we will quickly inform breeders of which loci are critical for specific phenotypes. At my job, I will continue to use the skills developed during my graduate studies and hopefully will have a positive impact on plant-based agriculture.

Bibliography

1. Publishing, B. E. *The History of Agriculture*. (Britannica Educational Publishing, 2012).
2. Kissinger, G., Herold, M. & Sy, V. de. *Drivers of deforestation and forest degradation: A synthesis report for REDD+ policymakers*. (2012).
3. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Rodés-Guirao, L. World Population Growth. *Our World Data* (2013).
4. United Nations, Department of Economic and Social Affairs, & Population Division. *World population prospects Highlights, 2019 revision Highlights, 2019 revision*. (2019).
5. Zhao, C. *et al.* Temperature increase reduces global yields of major crops in four independent estimates. *Proc. Natl. Acad. Sci.* **114**, 9326–9331 (2017).
6. Jägermeyr, J. *et al.* Climate impacts on global agriculture emerge earlier in new generation of climate and crop models. *Nat. Food* **2**, 873–885 (2021).
7. Robinson, T. THE EVOLUTION TOWARDS MORE COMPETITIVE APPLE ORCHARD SYSTEMS IN THE USA. *Acta Hortic.* 491–500 (2008) doi:10.17660/ActaHortic.2008.772.81.
8. Gallardo, R. K. & Brady, M. P. Adoption of labor-enhancing technologies by specialty crop producers: The case of the Washington apple industry. *Agric. Finance Rev.* **75**, 514–532 (2015).
9. Varshney, R. K. *et al.* Designing Future Crops: Genomics-Assisted Breeding Comes of Age. *Trends Plant Sci.* **26**, 631–649 (2021).
10. Langner, T., Kamoun, S. & Belhaj, K. CRISPR Crops: Plant Genome Editing Toward Disease Resistance. *Annu. Rev. Phytopathol.* **56**, 479–512 (2018).

11. Lemmon, Z. H. *et al.* Rapid improvement of domestication traits in an orphan crop by genome editing. *Nat. Plants* **4**, 766–770 (2018).

12. Kupferschmidt, K. EU verdict on CRISPR crops dismays scientists. *Science* **361**, 435–436 (2018).

13. Liu, W. & Stewart, C. N. Plant synthetic biology. *Trends Plant Sci.* **20**, 309–317 (2015).

14. Gottesfeld, J. M., Murphy, R. F. & Bonner, J. Structure of transcriptionally active chromatin. *Proc. Natl. Acad. Sci.* **72**, 4404–4408 (1975).

15. Weintraub, H. & Groudine, M. Chromosomal Subunits in Active Genes Have an Altered Conformation: Globin genes are digested by deoxyribonuclease I in red blood cell nuclei but not in fibroblast nuclei. *Science* **193**, 848–856 (1976).

16. Wu, C., Wong, Y.-C. & Elgin, S. C. The chromatin structure of specific genes: II. Disruption of chromatin structure during gene activity. *Cell* **16**, 807–814 (1979).

17. Keene, M. A., Corces, V., Lowenhaupt, K. & Elgin, S. DNase I hypersensitive sites in Drosophila chromatin occur at the 5'ends of regions of transcription. *Proc. Natl. Acad. Sci.* **78**, 143–146 (1981).

18. Feng, J. & Villeponteau, B. High-resolution analysis of c-fos chromatin accessibility using a novel DNase I-PCR assay. *Biochim. Biophys. Acta BBA-Gene Struct. Expr.* **1130**, 253–258 (1992).

19. Elgin, S. C. DNAase I-hypersensitive sites of chromatin. *Cell* **27**, 413–415 (1981).

20. Gross, D. S. & Garrard, W. T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).

21. Alexandre, C. M. *et al.* Complex Relationships between Chromatin Accessibility, Sequence Divergence, and Gene Expression in Arabidopsis thaliana. *Mol. Biol. Evol.* **35**, 837–854 (2018).

22. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* **339**, 1074–1077 (2013).

23. Sheen, J. Molecular mechanisms underlying the differential expression of maize pyruvate, orthophosphate dikinase genes. *Plant Cell* **3**, 225–245 (1991).

24. Jiang, F., Zhu, J. & Liu, H.-L. Protoplasts: a useful research system for plant cell biology, especially dedifferentiation. *Protoplasma* **250**, 1231–1238 (2013).

25. Wang, S., Tiwari, S. B., Hagen, G. & Guilfoyle, T. J. AUXIN RESPONSE FACTOR7 Restores the Expression of Auxin-Responsive Genes in Mutant Arabidopsis Leaf Mesophyll Protoplasts. *Plant Cell* **17**, 1979–1993 (2005).

26. Hirner, A. *et al.* Arabidopsis LHT1 Is a High-Affinity Transporter for Cellular Amino Acid Uptake in Both Root Epidermis and Leaf Mesophyll. *Plant Cell* **18**, 1931–1946 (2006).

27. Hwang, I. & Sheen, J. Two-component circuitry in Arabidopsis cytokinin signal transduction. *Nature* **413**, 383–389 (2001).

28. Tan, M.-L. M. C., Rietveld, E. M., van Marrewijk, G. A. M. & Kool, A. J. Regeneration of leaf mesophyll protoplasts of tomato cultivars (L. esculentum): factors important for efficient protoplast culture and plant regeneration. *Plant Cell Rep.* **6**, 172–175 (1987).

29. OECD & Food and Agriculture Organization of the United Nations. *OECD-FAO Agricultural Outlook 2021–2030*. (OECD, 2021).

30. Gallois, P. & Marinho, P. Leaf Disk Transformation Using Agrobacterium tumefaciens-Expression of Heterologous Genes in Tobacco. in *Plant Gene Transfer and Expression Protocols* (ed. Jones, H.) 39–48 (Springer New York, 1995). doi:10.1385/0-89603-321-X:39.

31. Zhang, X., Henriques, R., Lin, S.-S., Niu, Q.-W. & Chua, N.-H. Agrobacterium-mediated transformation of Arabidopsis thaliana using the floral dip method. *Nat. Protoc.* **1**, 641–646 (2006).

32. Schäffner, A. R. & Sheen, J. Maize rbcS promoter activity depends on sequence elements not found in dicot rbcS promoters. *Plant Cell* **3**, 997–1012 (1991).

33. Jores, T. Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell* **32**, 2120–2131 (2020).

34. Jores, T. *et al.* Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants* **7**, 842–855 (2021).

35. Ricci, W. A. Widespread long-range cis-regulatory elements in the maize genome. *Nat Plants* **5**, 1237–1249 (2019).

36. Jeon, J. M. *et al.* Efficient transient expression and transformation of PEG-mediated gene uptake into mesophyll protoplasts of pepper (Capsicum annuum L.). *Plant Cell Tissue Organ Cult.* **88**, 225–232 (2007).

37. Pindel, A. Optimization of isolation conditions of Cymbidium protoplasts. *Folia Hort* 10 (2007).

38. Ren, R. *et al.* Highly Efficient Leaf Base Protoplast Isolation and Transient Expression Systems for Orchids and Other Important Monocot Crops. *Front. Plant Sci.* **12**, (2021).

39. Yoo, S.-D., Cho, Y.-H. & Sheen, J. *Arabidopsis* mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nat. Protoc.* **2**, 1565–1572 (2007).

40. Scheben, A., Wolter, F., Batley, J., Puchta, H. & Edwards, D. Towards CRISPR/Cas crops—bringing together genomics and genome editing. *New Phytol* **216**, 682–698 (2017).

41. Swinnen, G., Goossens, A. & Pauwels, L. Lessons from domestication: Targeting cis-regulatory elements for crop improvement. *Trends Plant Sci* **21**, 506–515 (2016).

42. Andersson, R. & Sandelin, A. Determinants of enhancer and promoter activities of regulatory elements. *Nat. Rev. Genet.* **21**, 71–87 (2020).

43. Weber, B., Zicola, J., Oka, R. & Stam, M. Plant Enhancers: A Call for Discovery. *Trends Plant Sci.* **21**, 974–987 (2016).

44. Marand, A. P., Zhang, T., Zhu, B. & Jiang, J. Towards genome-wide prediction and characterization of enhancers in plants. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1860**, 131–139 (2017).

45. Banerji, J., Rusconi, S. & Schaffner, W. Expression of a β-globin gene is enhanced by remote SV40 DNA sequences. *Cell* **27**, 299–308 (1981).

46. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33**, 729–740 (1983).

47. Chandrasekharappa, S. C. & Subramanian, K. N. Effects of position and orientation of the 72-base-pair-repeat transcriptional enhancer on replication from the simian virus 40 core origin. *J Virol* **61**, 2973–2980 (1987).

48. Amano, T. *et al.* Chromosomal dynamics at the Shh locus: Limb bud-specific differential regulation of competence and active transcription. *Dev Cell* **16**, 47–57 (2009).

49. Studer, A., Zhao, Q., Ross-Ibarra, J. & Doebley, J. Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.* **43**, 1160–1163 (2011).

50. Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. *Nat Genet* **46**, 685–692 (2014).

51. Liu, Y. *et al.* Functional assessment of human enhancer activities using whole-genome STARR-sequencing. *Genome Biol* **18**, 219 (2017).

52. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer risk. *Genome Biol* **18**, 194 (2017).

53. Wang, X. *et al.* High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat Commun* **9**, 5380 (2018).

54. Sun, J. *et al.* Global quantitative mapping of enhancers in rice by STARR-seq. *Genomics Proteomics Bioinformatics* **17**, 140–153 (2019).

55. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).

56. Lehti-Shiu, M. D., Panchy, N., Wang, P., Uygun, S. & Shiu, S.-H. Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families. *Biochim Biophys Acta Gene Regul Mech* **1860**, 3–20 (2017).

57. Wilhelmsson, P. K. I., Mühlich, C., Ullrich, K. K. & Rensing, S. A. Comprehensive genome-wide classification reveals that many plant-specific transcription factors evolved in streptophyte algae. *Genome Biol Evol* **9**, 3384–3397 (2017).

58. Fang, R. X., Nagy, F., Sivasubramaniam, S. & Chua, N. H. Multiple cis regulatory elements for maximal expression of the cauliflower mosaic virus 35S promoter in transgenic plants. *Plant Cell* **1**, 141–150 (1989).

59. Benfey, P. N., Ren, L. & Chua, N. H. Tissue-specific expression from CaMV 35S enhancer subdomains in early stages of plant development. *EMBO J* **9**, 1677–1684 (1990).

60. Odell, J. T., Knowlton, S., Lin, W. & Mauvais, C. J. Properties of an isolated transcription stimulating sequence derived from the cauliflower mosaic virus 35S promoter. *Plant Mol Biol* **10**, 263–272 (1988).

61. Fluhr, R., Kuhlmeier, C., Nagy, F. & Chua, N.-H. Organ-specific and light-induced expression of plant genes. *Science* **232**, 1106–1113 (1986).

62. Simpson, J., Schell, J., Montagu, M. V. & Herrera-Estrella, L. Light-inducible and tissue-specific pea lhcp gene expression involves an upstream element combining enhancer- and silencer-like properties. *Nature* **323**, 551–554 (1986).

63. Nagy, F., Boutry, M., Hsu, M. Y., Wong, M. & Chua, N. H. The 5′-proximal region of the wheat Cab-1 gene contains a 268-bp enhancer-like sequence for phytochrome response. *EMBO J.* **6**, 2537–2542 (1987).

64. Giuliano, G. *et al.* An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. *Proc. Natl. Acad. Sci.* **85**, 7089–7093 (1988).

65. Fejes, E. *et al.* A 268 bp upstream sequence mediates the circadian clock-regulated transcription of the wheat Cab-1 gene in transgenic plants. *Plant Mol. Biol.* **15**, 921–932 (1990).

66. Argüello, G. *et al.* Characterization of DNA sequences that mediate nuclear protein binding to the regulatory region of the Pisum sativum (pea) chlorophyl a/b binding protein gene AB80: Identification of a repeated heptamer motif. *Plant J* **2**, 301–309 (1992).

67. Gotor, C., Romero, L. C., Inouye, K. & Lam, E. Analysis of three tissue-specific elements from the wheat Cab-1 enhancer. *Plant J.* **3**, 509–518 (1993).

68. Kertész, S. *et al.* Both introns and long 39-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic Acids Res* **34**, 6147–6157 (2006).

69. May, J. P., Yuan, X., Sawicki, E. & Simon, A. E. RNA virus evasion of nonsense-mediated decay. *PLoS Pathog* **14**, 1007459 (2018).

70. Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C. & Cohen, B. A. Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proc Natl Acad Sci USA* **109**, 19498–19503 (2012).

71. Inoue, F., Kreimer, A., Ashuach, T., Ahituv, N. & Yosef, N. Identification and massively parallel characterization of regulatory elements driving neural induction. *Cell Stem Cell* **25**, 713-727 10 (2019).

72. Ido, A. *et al.* Arabidopsis Pol II-dependent in vitro transcription system reveals role of chromatin for light-inducible rbcS gene transcription. *Plant Physiol* **170**, 642–652 (2016).

73. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: Scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

74. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. & Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res* **48**, 1104–1113 (2020).

75. Lam, E., Benfey, P. N., Gilmartin, P. M., Fang, R. X. & Chua, N. H. Site-specific mutations alter in vitro factor binding and change promoter expression pattern in transgenic plants. *Proc Natl Acad Sci USA* **86**, 7890–7894 (1989).

76. Niggeweg, R., Thurow, C., Kegler, C. & Gatz, C. Tobacco transcription factor TGA2.2 is the main component of as-1-binding factor ASF-1 and is involved in salicylic acid- and auxin-inducible expression of as-1-containing target promoters. *J Biol Chem* **275**, 19897–19905 (2000).

77. Lam, E. & Chua, N. H. ASF-2: A factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in Cab promoters. *Plant Cell* **1**, 1147–1156 (1989).

78. Callis, J., Fromm, M. & Walbot, V. Introns increase gene expression in cultured maize cells. *Genes Dev* **1**, 1183–1200 (1987).

79. Rose, A. B. & Last, R. L. Introns act post-transcriptionally to increase expression of the Arabidopsis thaliana tryptophan pathway gene PAT1. *Plant J* **11**, 455–464 (1997).

80. Rose, A. B. The effect of intron location on intron-mediated enhancement of gene expression in Arabidopsis. *Plant J* **40**, 744–751 (2004).

81. Samadder, P., Sivamani, E., Lu, J., Li, X. & Qu, R. Transcriptional and post-transcriptional enhancement of gene expression by the 59 UTR intron of rice rubi3 gene in transgenic rice cells. *Mol Genet Genomics* **279**, 429–439 (2008).

82. Laxa, M. *et al.* The 59UTR intron of Arabidopsis GGT1 aminotransferase enhances promoter activity by recruiting RNA polymerase II. *Plant Physiol* **172**, 313–327 (2016).

83. Laxa, M. Intron-mediated enhancement: A tool for heterologous gene expression in plants? *Front Plant Sci* **7**, (2017).

84. Bruce, W. B., Christensen, A. H., Klein, T., Fromm, M. & Quail, P. H. Photoregulation of a phytochrome gene promoter from oat transferred into rice by particle bombardment. *Proc Natl Acad Sci USA* **86**, 9692–9696 (1989).

85. Christensen, A. H., Sharrock, R. A. & Quail, P. H. Maize polyubiquitin genes: structure, thermal perturbation of expression and transcript splicing, and promoter activity following transfer to protoplasts by electroporation. *Plant Mol Biol* **18**, 675–689 (1992).

86. Wroblewski, T., Tomczak, A. & Michelmore, R. Optimization of Agrobacterium-mediated transient assays of gene expression in lettuce, tomato and Arabidopsis. *Plant Biotechnol J* **3**, 259–273 (2005).

87. Andrieu, A. *et al.* An in planta, Agrobacterium-mediated transient gene expression method for inducing gene silencing in rice (Oryza sativa L.) leaves. *Rice* **N. Y.) 5**, 23 (2012).

88. Zheng, L., Liu, G., Meng, X., Li, Y. & Wang, Y. A versatile Agrobacterium-mediated transient gene expression system for herbaceous plants and trees. *Biochem Genet* **50**, 761–769 (2012).

89. Bond, D. M. *et al.* Infiltration-RNAseq: Transcriptome profiling of Agrobacterium-mediated infiltration of transcription factors to discover gene function and expression networks in plants. *Plant Methods* **12**, 41 (2016).

90. Sheen, J. Metabolic repression of transcription in higher plants. *Plant Cell* **2**, 1027–1038 (1990).

91. Zhang, Y. *et al.* A highly efficient rice green tissue protoplast system for transient gene expression and studying light/chloroplast-related processes. *Plant Methods* **7**, 30 (2011).

92. Nanjareddy, K., Arthikala, M.-K., Blanco, L., Arellano, E. S. & Lara, M. Protoplast isolation, transient transformation of leaf mesophyll protoplasts and improved Agrobacterium-mediated leaf disc infiltration of Phaseolus vulgaris: Tools for rapid gene expression analysis. *BMC Biotechnol* **16**, 53 (2016).

93. Hellens, R. P., Edwards, E. A., Leyland, N. R., Bean, S. & Mullineaux, P. M. pGreen: a versatile and flexible binary Ti vector for Agrobacterium-mediated plant transformation. *Plant Mol Biol* **42**, 819–832 (2000).

94. Kanoria, S. & Burma, P. K. A 28 nt long synthetic 59UTR (synJ) as an enhancer of transgene expression in dicotyledonous plants. *BMC Biotechnol* **12**, 85 (2012).

95. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3, e3647**, (2008).

96. Halperin, S. O. *et al.* CRISPR-guided DNA polymerases enable diversification of all nucleotides in a tunable window. *Nature* **560**, 248–252 (2018).

97. Subramaniam, A. R., Pan, T. & Cluzel, P. Environmental perturbations lift the degeneracy of the genetic code to regulate protein levels in bacteria. *Proc Natl Acad Sci USA* **110**, 2419–2424 (2013).

98. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

99. Lomonossoff, G. P. & D'Aoust, M.-A. Plant-produced biopharmaceuticals: a case of technical developments driving clinical deployment. *Science* **353**, 1237–1240 (2016).

100. Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu Rev Biochem* **72**, 449–479 (2003).

101. Grosschedl, R. & Birnstiel, M. L. Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *Proc Natl Acad Sci USA* **77**, 1432–1436 (1980).

102. Wasylyk, B. Specific in vitro transcription of conalbumin gene is drastically decreased by single-point mutation in T-A-T-A box homology sequence. *Proc Natl Acad Sci USA* **77**, 7024–7028 (1980).

103. Smale, S. T. & Baltimore, D. The "initiator" as a transcription control element. *Cell* **57**, 103–113 (1989).

104. Ince, T. A. & Scotto, K. W. A conserved downstream element defines a new class of RNA polymerase II promoters. *J Biol Chem* **270**, 30249–30252 (1995).

105. Burke, T. W. & Kadonaga, J. T. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA box-deficient promoters. *Genes Dev* **10**, 711–724 (1996).

106. Lagrange, T., Kapanidis, A. N., Tang, H., Reinberg, D. & Ebright, R. H. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev* **12**, 34–44 (1998).

107. Lewis, B. A., Kim, T.-K. & Orkin, S. H. A downstream element in the human β-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. *Proc Natl Acad Sci USA* **97**, 7172–7177 (2000).

108. Lim, C. Y. The MTE, a new core promoter element for transcription by RNA polymerase II. *Genes Dev* **18**, 1606–1617 (2004).

109. Deng, W. & Roberts, S. G. E. A core promoter element downstream of the TATA box that is recognized by TFIIB. *Genes Dev* **19**, 2418–2423 (2005).

110. Parry, T. J. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev* **24**, 2013–2018 (2010).

111. Molina, C. & Grotewold, E. Genome wide analysis of Arabidopsis core promoters. *BMC Genom* **6**, 25 (2005).

112. Yamamoto, Y. Y. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res* **35**, 6219–6226 (2007).

113. Bernard, V., Brunaud, V. & Lecharny, A. TC-motifs at the TATA box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genom* **11**, 166 (2010).

114. Blake, M. C., Jambou, R. C., Swick, A. G., Kahn, J. W. & Azizkhan, J. C. Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter. *Mol Cell Biol* **10**, 6632–6641 (1990).

115. Patwardhan, R. P. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotechnol* **27**, 1173–1175 (2009).

116. Sharon, E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**, 521–530 (2012).

117. Lubliner, S. Core promoter sequence in yeast is a major determinant of expression level. *Genome Res* **25**, 1008–1017 (2015).

118. Arnold, C. D. Genome-wide assessment of sequence-intrinsic enhancer responsiveness at single-base-pair resolution. *Nat Biotechnol* **35**, 136–144 (2017).

119. Arensbergen, J. Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* **35**, 145–153 (2017).

120. Weingarten-Gabbay, S. Systematic interrogation of human promoters. *Genome Res* **29**, 171–183 (2019).

121. Boer, C. G. Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat Biotechnol* **38**, 56–65 (2020).

122. Kotopka, B. J. & Smolke, C. D. Model-driven generation of artificial yeast promoters. *Nat Commun* **11**, 2113 (2020).

123. Kumari, S. & Ware, D. Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots. *PLoS ONE* **8, e79011**, (2013).

124. Morton, T. Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell* **26**, 2746–2760 (2014).

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

126. Kiran, K. The TATA box sequence in the basal promoter contributes to determining light-dependent gene expression in plants. *Plant Physiol* **142**, 364–376 (2006).

127. Srivastava, R. Distinct role of core promoter architecture in regulation of light-mediated responses in plant genes. *Mol Plant* **7**, 626–641 (2014).

128. Cai, Y.-M. Rational design of minimal synthetic promoters for plants. *Nucleic Acids Res* **48**, 11845–11856 (2020).

129. Srivastava, A. K., Lu, Y., Zinta, G., Lang, Z. & Zhu, J.-K. UTR-dependent control of gene expression in plants. *Trends Plant Sci* **23**, 248–259 (2018).

130. Yahraus, T., Chandra, S., Legendre, L. & Low, P. S. Evidence for a mechanically induced oxidative burst. *Plant Physiol* **109**, 1259–1266 (1995).

131. Walley, J. W. Integration of omic networks in a developmental atlas of maize. *Science* **353**, 814–818 (2016).

132. Wang, B. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Res* **28**, 921–932 (2018).

133. Mergner, J. Mass-spectrometry-based draft of the Arabidopsis proteome. *Nature* **579**, 409–414 (2020).

134. Singh, R., Ming, R. & Yu, Q. Comparative analysis of GC content variations in plant genomes. *Trop Plant Biol* **9**, 136–149 (2016).

135. Rensink, W. A. Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts. *BMC Genom* **6**, 124 (2005).

136. Tsai, F. T. F. & Sigler, P. B. Structural basis of preinitiation complex assembly on human Pol II promoters. *EMBO J* **19**, 25–36 (2000).

137. Gehrig, J. Automated high-throughput mapping of promoter–enhancer interactions in zebrafish embryos. *Nat Methods* **6**, 911–916 (2009).

138. Yanai, I. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).

139. Heerah, S., Katari, M., Penjor, R., Coruzzi, G. & Marshall-Colon, A. WRKY1 mediates transcriptional regulation of light and nitrogen signaling pathways. *Plant Physiol* **181**, 1371–1388 (2019).

140. Cuperus, J. T. Deep learning of the regulatory grammar of yeast 5′ untranslated regions from 500,000 random sequences. *Genome Res* **27**, 2015–2024 (2017).

141. Klein, J. C. A systematic evaluation of the design and context dependencies of massively parallel reporter assays. *Nat Methods* **17**, 1083–1091 (2020).

142. Hong, C. K. & Cohen, B. A. Genomic environments scale the activities of diverse core promoters. (2021) doi:10.1101/2021.03.08.434469.

143. Dorrity, M. W. The regulatory landscape of Arabidopsis thaliana roots at single-cell resolution. *Nat Commun* 23675- (1038).

144. Marand, A. P., Chen, Z., Gallavotti, A. & Schmitz, R. J. A cis-regulatory atlas in maize at single-cell resolution. *Cell Httpsdoiorg101016j* (2021).

145. Zhang, T.-Q., Chen, Y., Liu, Y., Lin, W.-H. & Wang, J.-W. Single-cell transcriptome atlas and chromatin accessibility landscape reveal differentiation trajectories in the rice root. *Nat Commun* **12**, (2021).

146. Cheng, C.-Y. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. *Plant J* **89**, 789–804 (2017).

147. McCormick, R. F. The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *Plant J* **93**, 338–354 (2018).

148. Mejía-Guerra, M. K. Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. *Plant Cell* **27**, 3309–3320 (2015).

149. Jiao, Y. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).

150. Masella, A. P., Bartram, A. K., Truszkowski, J. M., Brown, D. G. & Neufeld, J. D. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinform* **13**, 31 (2012).

151. Raudvere, U. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update. *Nucleic Acids Res* **47**, 191–198 (2019).

152. Madeira, F. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res* **47**, 636–641 (2019).

153. Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M. & Solovyev, V. V. PlantProm: a database of plant promoter sequences Nucleic Acids Res. vol. 31 114–117 (2003).

154. Fornes, O. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **48**, 87–92 (2020).

155. Onimaru, K., Nishimura, O. & Kuraku, S. Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information. *PLoS ONE* **15, e0235748**, (2020).

156. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).

157. Wittkopp, P. J. & Kalay, G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2012).

158. Carroll, S. B. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* **134**, 25–36 (2008).

159. Garcia-Hernandez, E., Gariglio, P., Herrera-Estrella, L. & Simpson, J. Characterization of DNA sequences that mediate nuclear protein binding to the regulatory region of the Pisum safiwum (pea) chlorophyl a/b binding protein gene A680: identification of a repeated heptamer motif. *Plant J.* **2**, 301--309 (1992).

160. Arnosti, D. N. & Kulkarni, M. M. Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? *J. Cell. Biochem.* **94**, 890–898 (2005).

161. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

VITA

Jackson Tonnies was born in Johnson City, Tennessee and grew up in Kingsport nearby. He lived there throughout high school and then moved to Berkeley where he acquired a Bachelor of Sciences from University of California Berkeley in Genetics and Plant Biology. At Berkeley, he learned sterile technique and plant tissue culture while working in the lab of Peggy Lemaux. Jackson then came to the University of Washington, Seattle to pursue a PhD in plant biology.

He loves plants! He has worked as a park ranger in Tennessee, at the Berkeley Botanical Gardens, and volunteered at the University of Washington Native Plant Nursery. He is always looking for new hikes and to forage wild mushrooms. To date, he has found and eaten nine different types of mushrooms in the forests of Washington.