Transcript cleavage and polyadenylation in plants

Sayeh Gorjifard

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

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Christine Queitsch, Chair

Stanley Fields

Philip Green

Program Authorized to Offer Degree:
Genome Sciences

University of Washington

Abstract

Transcript cleavage and polyadenylation in plants

Sayeh Gorjifard

Chair of the Supervisory Committee:
Christine Queitsch
Genome Sciences

Eukaryotic gene expression is finely regulated at the post-transcriptional level by the untranslated regions of mRNA. The coding sequence (CDS) of mRNA is flanked by 5'- and 3'-untranslated regions (UTRs). The end boundary of the 3' UTR is defined by transcript cleavage and polyadenylation. The genic region that determine where the cleavage and polyadenylation complex (CPMC) binds and cleaves is called the terminator. Terminators overlap significantly with 3'UTRs but also include the sequences after the 3' UTR boundary. Elements in the resulting 3' UTR modulate stability, nuclear export, localization, and translation.

In this body of work, I will provide an overview of the historical exploration of terminator cleavage and polyadenylation, emphasizing the biotechnology that aided these discoveries. I will focus on how advances in DNA sequencing technologies expanded our understanding of terminator genetics and functionality across eukaryotes, with a particular emphasis on plants. Apart from transcriptome wide maps of cleavage and polyadenylation signaling, sequencing empowered functional genomics by enabling massively parallel reporter assays (MPRAs). These tools, in conjunction with computational machine learning, will allow the engineering of specific terminators for diverse applications in plant synthetic biology. Due to the limitations of plant systems, however, little work has been done to characterize plant terminator sequences on a genome wide basis for their strength in directing cleavage and fine tuning expression.

Following upon recent developments optimizing massively parallel reporter assays in transient tobacco leaves and maize protoplasts, I characterized nearly all *Arabidopsis thaliana* and maize terminator sequences for their strength in conferring expression and cleavage. The resulting data helped train a deep learning model to predict terminator strength, aiding in the *in silico* evolution of synthetic and species-specific terminators. In the final chapter, I will address existing limitations in the field and propose new experiments to fill in the gaps. Finally, I will turn to the elephant in the room. Do all these high throughput sequencing technologies and protocols help us get any closer to accurately predicting gene expression? Are we even capturing the data in a meaningful way if we lose higher order information among all the layers of gene regulation?

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

3' CLEAVAGE SITE: Site of cleavage of the pre-mRNA by the cleavage and polyadenylation machinery.

3' UNTRANSLATED REGION: Genic region immediately after the translation terminator codon up to the cleavage and polyadenylation site.

AAUAAA: Poly(A) signal hexamer AAUAAA is found in the Near Upstream Element of eukaryotic mRNAs and is involved in the accurate and efficient cleavage and polyadenylation of pre-mRNAs.

ALTERNATIVE POLYADENYLATION: A phenomenon that produces RNA molecules with different 3' ends due to distinct polyadenylation sites of a single gene.

CLEAVAGE AND POLYADENYLATION MACHINERY: A multi-protein core complex and dozens of associated factors that bind and recognize terminator elements to direct cleavage and polyadenylation.

DEEP LEARNING: A subset of machine learning methods based on artificial neural networks with representation learning. They are "deep" because they use of multiple layers in the network.

DOWNSTREAM SEQUENCE ELEMENT: Region downstream of the cleavage site of 3'end was shown to enhance cleavage and polyadenylation.

EXPRESSED SEQUENCE TAG: Short sub-sequence of a cDNA sequence often used to identify gene transcripts.

FUE-NUE-CS TRIPARTATE SIGNAL: The far upstream element (FUE), near upstream element (NUE), and the cleavage element (CE) are the polyadenylation signals in plants that recognize the cleavage and polyadenylation machinery.

GENE SILENCING: Gene silencing is a molecular defense mechanism that knocks down the gene expression in plants both in nature and in response to external stimuli.

MRNA: Messenger RNA is a single-stranded molecule of RNA that corresponds to the genetic sequence of a gene, and is read by a ribosome in the process of synthesizing a protein.

MASSIVELY PARALLEL REPORTER ASSAY: Assay that can functionally validate thousands of regulatory elements simultaneously using high-throughput sequencing and barcode technology.

MICRORNA: Small, single-stranded, non-coding RNA molecules that are typically 21 to 23 nucleotides. MiRNAs are involved in RNA silencing and post-transcriptional regulation of gene expression.

NEXT GENERATION SEQUENCING: A massively parallel DNA and RNA sequencing technology. NGS enables the interrogation of hundreds to thousands of genes at one time in multiple samples, as well as discovery and analysis of different types of genomic features in a single sequencing run.

OPEN READING FRAME: A span of DNA sequence between a start and stop codon.

POLY(A)-TAIL: The poly-A tail is a long chain of adenine nucleotides that is added to a messenger RNA (mRNA) molecule during RNA processing to increase the stability of the molecule.

POLYADENYLATION: The addition of the multiple adenosine monophosphates to an RNA transcript after termination of transcription.

POLYADENYLATION SIGNALS: Specific sequences in the pre-mRNA that bind the cleavage and polyadenylation complex.

PLANT STARR-SEQ: A massively parallel reporter assay in plants used to interrogate libraries of regulatory elements in a barcoded multiplexed fashion.

POST-TRANSCRIPTIONAL REGULATION: Control of gene expression at the RNA level, often modulated by RNA binding proteins that control splicing, capping, cleavage and alternative polyadenylation, nuclear degradation, RNA editing, stability, nuclear export, localization, and translation.

PRE-MRNA: pre-messenger RNA is the first (primary) transcript from a protein coding gene.

RNA BINDING PROTEINS: Cytoplasmic and nuclear proteins that bind to the double or single stranded RNA in cells and participate in forming ribonucleoprotein complexes. They modulate post-transcriptional control of RNAs, such as: splicing, polyadenylation, mRNA stabilization, mRNA localization and translation.

RNA SEQUENCING: an NGS technology specific to the transcriptome by the sequencing complementary DNAs (cDNA).

TERMINATOR: A genic element located downstream of the coding sequence that, once transcribed, is recognized by different protein complexes responsible for cleavage and polyadenylation in mRNA biogenesis.

UGUA: Poly(A) signal UGUA is found near in the Far Upstream Element involved in cleavage efficiency.

UPSTREAM SEQUENCE ELEMENT: Region upstream of the cleavage site of 3'end was shown to enhance cleavage and polyadenylation efficiency.

ACKNOWLEDGMENTS

I can't say the road to completing this dissertation was easy. What is not discussed in this work is all the failed project ideas that did not amount to anything, particularly a failed venture in trying to find dominant negative peptides of Cas9 and IL-6. Technology development is not for the faint of heart. Nevertheless, what I learned in the first projects I carried into the work of this dissertation. The bulk of this dissertation and all the data herein came from the final and a half year of my PhD.

All of this is to say that I owe a lot of my mental health and accomplishments to the wonderful mentors and friends I've made along the way.

None of this would have been possible without Dr Christine Queitsch. In the face of failure, you need someone to remind you that science is fun. She really is, as Kerry would say, a hurricane. I love her strength of will and the energy she brings into science, her unbridled enthusiasm about the work we do and about plants. I don't know if she remembers, but the first time I spoke to Christine in her office, I told her something along the lines of "If you helicopter mentor me, I will quit. I'm not here to learn how to pipette." She promised to make sure I have the freedom to think about science the way I want to, and she delivered! Because of her, I never had to worry about funding or lack of resources or opportunities. I always felt safe to question and follow the leads that interested me, even if at times, they were a distraction (they were).

Josh Cuperus is the pepper to Christine's salt. Just as much as you need someone that is excited about science, you need someone who thinks about the logistics and technicality of it. He has an intuitive sense for technology development, and in another life would have been a great engineer or cowboy.[1]

How lucky was I to work with Stan Fields? Stan is a fountain of science wisdom and

[1]I'm certain he's also the only one that read this thesis that was not on my reading committee.

best practices. He was always willing to lend an ear if I stopped by his office or caught him microwaving his lunch[2]. The best part about Stan, apart from his great sense of humor and evergreen curiosity about science, is that he will give you his best ideas, and then immediately forget that it was his idea. He is one of the greatest scientific minds of this department. I'm so thankful that we had Stan lead Genome Sciences during the pandemic.

Christine and Stan also have an incredible talent in hiring the nicest and most prodigious scientists. Kerry Bubb single-handedly gave me the best education in coding and statistics you can have without ever taking a course. Kerry is a pillar of the lab, a cornerstone of scientific integrity and bioinformatics, and an all around sweetheart. She isn't fooled easily and forces you to think deeply about your data, and when you're stuck, she generously helps untangle you. As an artist, I honestly appreciated how little pop and pizzazz swayed Kerry. No fancy diagram or color palette would distract her from bad data. The proof of the pudding is in the eating; one should be able to download the raw data of any high throughput experiment and reach the same conclusions the authors did. Outside of science, I'll miss sitting by Kerry and chatting about the latest seminar or what books she's been reading. I know how much she cares about all the students and we love her just as much too.

The rest of the lab members are all jewels. Ken was my first lab friend. Brilliant and kind, he always took time away from all his work to help me code when I was stuck, even past dinner time! He would never leave a soldier behind. I don't know what I would have done in the last two years without Jackson and Morgan. I'll always cherish going to Cornell and getting lost in the botanical gardens following Morgan and Jackson around, eating a massive block of cheese and thinking about deciduous trees. I loved how Jackson was always down to get lost in nature with me and feed my insatiable need for McDonalds ice cream cones. GS Hackathon would have never happened without Morgan or Joe either. Morgan could easily make cool evil robots, but instead, he wants to save the world with plants. Thank you Joe for helping to lead the Hackathon when I had a thesis to write. Your energy

[2]Even if it was fish!

and brilliance is lifesaving and you bring sunshine into my life.

Dearest sweetest Cris Alexandre. I always imagined a PhD to be like the old stories of Cold Spring Harbor Labs in the golden age of molecular biology. Walking around a collegiate campus, coffee in hand, and chatting with your colleagues about some grandiose theory that will shake the field. Each discussion would ricochet into a new idea, a new experiment to try. Working with Cris was even better. My fondest memories of grad school will be walking with her to get coffee and chat about everything–science, philosophy, art, literature–and connecting all the dots.[3] She has a mind like no other, and a heart that is equally as beautiful. I can't wait to see how she'll shake up the worlds of design and science.

The same goes for my dear friend Ruth Groza. I remember the first day she joined the Field's lab and I asked her how she was doing? Matter of fact, she responded "Are you just being polite, or do you *really* want to know?" And that's Ruth for you. She will never tell you the sky is blue when it's not, but she will stand by your side with an umbrella until the rain clears. With Ruth, I've been on top of mountains, glaciers, volcanoes, and crouched by every tide pool in between. I'll miss our "fireside" (bunsen burner) chats about science philosophy and COVID health policy, especially in those wee late hours of lab when it was only us in the building. Foege was our palace, and we were its queens.

I'd also want to thank the wonderful postdocs Tobias, Yash, Beth, Cole, and Bryan. Postdocs are the ground troops of science, helping us poor grad students left behind. Tobias is three scientists in one trench coat. His methodical approach to building STARR-seq off the ground showed me how tech-dev should be done and how to be as productive as possible. Yash is a protein wizard and academic. He brings back the fundamental biochemistry that we often ignore when studying genetics. I couldn't have purified proteins without him! Beth taught me how to run experiments with fastidious attention to detail and organization. Her lab notebooks are work of art and should be framed for their perfection. And finally, this last year would have been impossible without the support of Cole and Bryan. While I've only known them a short time, they've quickly become my support system in lab and beyond.

[3]My only complaint is her lack of appreciation for The Great Gatsby.

No, I will never read the manual or turn on the lights, but you guys make sure I get by anyway and that the lab doesn't catch on fire.

To the wonderful staff members Liz Kwan, Maureen Larson, and Brian Giebel. Liz is the most thoughtful, creative, and passionate scientists I've ever met. She has more talent in her pinky finger than I have in my whole body. Maureen is the real queen of Genome Sciences. All praise the Queen! Without her, Hackathon would have never happened. I loved our little chats and the energy she brings to GS. Brian Giebel is the foundation of everything. Without him, I would have never been able to graduate or keep up to date with all the paperwork and deadlines. He should be given a 200% raise every year. We cannot lose him.

I have the best friends outside of science and I love them all. The luckiest girl in the world, I've grown up with the kindest group of friends and family. My childhood friends Sabina, Alex, and Helena have seen me through thick and thin. They are no longer my friends, they are my sisters, and I can't wait to grow old with them and see their babies grow up too.

Meeting Anna Noreuil was the best part of Dartmouth. I didn't think I would meet another gremlin out there like me, but the lord graced me with her, another trash goblin from the abyss. I don't think I would have survived both my masters and grad school if I didn't have her to call when times were hard, or just to rant about anything. She always know how to make me laugh and what weird bug facts would cheer me up, not to mention fulfilling my girlish fantasy of going to Paris for our birthdays or seeing Forks, Washington. Even when we were miserably trapped in a tree on top of Aster Butte, I knew I was going to be okay because you were right there with me. I love her so much and also can't wait to grow old with her.

To my friend Marine, my sister from another mister. Meeting her in New York was the silver lining of New York. I love her openness to the world and others. She made my move to Seattle so much easier, opening her whole family to me. Never have you met a better clan, or more genuine and kind. I can't wait for us to have art shows and sell photographs

together, travel every corner of the globe, and read each other's books.

How lucky was I in finding the best house to live in Seattle? A lot. For 4 years, I lived by Ravenna park with the most compassionate and kind housemates (now friends too). Jim, Christy, Max, and Faye made COVID all the more bearable, especially since they had to eat all the sourdough I baked. Thank you to Luke, Bill, Melissa, Kaya, and Courtney for filling our house with love and joy. Our parties and dinners were the talk of the town and will probably be some of my most cherished memories living out here. I'm so glad you all put up with my silly goose self. And thanks to our landlord Wallace for always taking care of us like his own children.

I owe everything I have ever accomplished to my dear mom and dad. With two small kids, they left everything behind to immigrate to the United States, which wasn't without its sacrifice. Dad, I miss you a lot and wish you could have seen me get a PhD. You were gone too soon and there were so many times I wish I could have called you in the last 12 years. To my brother, Amir: You're so smart and so special, and I can't wait for you to show the world how brilliant you are. Even though we don't always see eye to eye, know that I'm proud of you for taking your own path.

And finally, to my dear Maman. I love you so much it doesn't make sense to put it in words. I wish I just had even 10 percent of the resilience and strength that you have. I watched you build our lives from nothing, overcoming setback after setback, sacrificing everything in your power for your kids. I love your strength of character, your work ethic, your high moral code, and your desire to serve the community before serving yourself. I don't know how I got so lucky to have a powerhouse mother like you. It's silly, but every time I eat cheetos puffs, I thank God that she made me your daughter. This PhD is as much mine as it is yours. You are and forever will be the best thing to have ever happened to me.

DEDICATION

To my dear Mom and Dad, who gave me everything.

Chapter 1

INTRODUCTION

1.1 mRNA: the medium is the message

Imagine you read the sentence:

Cole kicked Jimmy.

The message is clear. The action is obvious. For some reason, Cole kicked poor Jimmy! But what if we insert a few phrases or clauses?

Cole kicked Jimmy playfully.

Cole kicked Jimmy the soccer ball.

Seeing the incoming truck, **Cole kicked Jimmy** out of the way.

In each example, the action stays the same. Cole does indeed kick Jimmy. But how we interpret the message has now completely changed. By adding words at the end of the sentence, we can modify the context and consequently the understanding or meaning of the message. In a way, how the messages of genes are interpreted or translated are not that different. The additional words modifying the action in these sentences function like the noncoding bases in messenger RNA (mRNA), which modulate how the protein or "message" the mRNA encodes gets translated. They add context to the message, but they themselves are not the action. "Messenger" RNA acts as a form of communication between the nucleus and the ribosomes. How it communicates that message is just as important as the message itself. Even more critically, how and where the cell ends the message plays a critical role in regulation of gene expression. Indeed, the medium is the message.

Genetic regulation can happen at any stage as DNA is transcribed into mRNA and mRNA is translated into protein. Regulation on transcription initiation is the predominant form of regulation for most genes, and often the most studied. Yet transcription is surprisingly stochastic whereas protein production is precise [148]. Post-transcriptional con-

trols **(Figure 1.1)** modulate the amount of protein produced from an individual mRNA by altering rates of mRNA pre-processing, decay, and translation. Since transcription and translation are not directly coupled in eukaryotes, post-transcriptional regulation evolved to transform the nuclear primary RNA transcript (pre-mRNA) into a mature mRNA competent for translation in the cytoplasm. The processes that control the life and and death of mRNA account for nearly 60 percent of the variation in protein abundances [271, 37].

In eukaryotes, there are three major processing reactions responsible for the maturation of the newly transcribed pre-mRNA molecule [248]. First, the 5' ends of the pre-mRNA are enzymatically modified to generate a 7-methylguanosine cap, a process that protects the transcript from being broken down and helps the ribosome attach to the mRNA. Second, the splicing machinery (a large multi-protein and RNA complex known as the spliceosome) removes introns and stitches back the coding exonic regions of the mRNA to form a continuous **open reading frame (ORF)** ready for translation. Third, in order to be exported to the cytoplasm, the 3' end of the mRNA is then cleaved and polyadenylated. The final **poly(A) tail** is synthesized by a multiprotein complex that assembles on specific sequences of the pre-mRNA, called the **cleavage and polyadenylation signals (pA signals)**. These pA signals usually reside in the terminator region of the pre-mRNA and flank both sides of the cleavage and polyadenylation site. 3' end processing, particularly cleavage and polyadenylation, is perhaps the most critical aspect of nuclear post-transcriptional regulation. Where the transcript is cleaved and polyadenylated determines what other *cis*-regulatory elements remain in the 3' end. For many genes, there are **alternative cleavage and polyadenylation sites (PAS)** that can generate multiple mRNA isoforms with different 3' UTRs. This phenomenon, known as **alternative polyadenylation (APA)**, allows for inclusion and expulsion of *cis*-regulatory and *cis*-acting elements that control mRNA decay, subcellular localization, stability, and translation efficiency. APA is critical since it affords a direct linkage between mRNA processing and control of mRNA function. Shorter transcripts devoid of RNA regulatory elements (binding sites) and freed from that form of regulation, while longer transcripts are more likely to harbor sites of control. **RNA-binding proteins (RPBs)** and **microRNAs (miRNAs)** bind these elements in the 3' UTR and 5' UTR to perturb RNA stability and function. Depending on the RBP, the stability can be increased

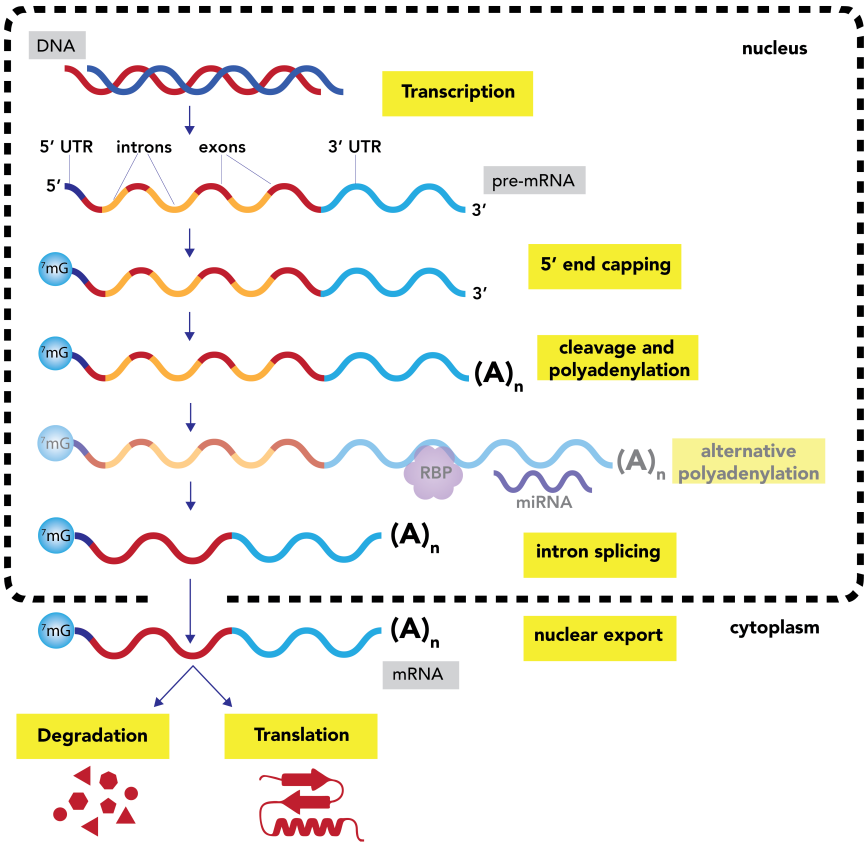


Figure 1.1: Post-transcriptional control can occur at any stage after transcription and before translation. Once RNA is transcribed, it must be processed to create a mature RNA that is ready to be translated. Nuclear processing involves 5' capping, intron splicing, and addition of a poly(A) tail to the 3' end. The **3'-UTR (untranslated region)** harbors cis-elements like **microRNA** and **RNA-binding protein (RBP)** binding sites. **Alternative polyadenylation (APA)** determines what regulatory elements are included, thereby having a significant effect on the stability and translational rate of mRNAs. mRNA is then transported to the cytoplasm for translation. The length of time mRNA resides in the cytoplasm before being degraded is called RNA stability. Higher RNA stability leads to longer residency time in the cytoplasm and more protein synthesis.

or decreased significantly; however, miRNAs always decrease stability and promote decay, a phenomenon known as **gene silencing**. MicroRNAs are small single-stranded non-coding RNA, typically 20-24 bases long, that bind to mRNA and silence them either via cleavage of mRNA, destabilization by shortening the poly(A) tail, or repression of translation initiation. RBP can also bind elements in the 5' UTR and 3' UTR to guide transport and cytoplasmic localization. The poly(A) tail and poly(A) binding protein interact with the methyl cap at the 5'-end to promote translation at the ribosome [336, 89].

1.2 Importance of untranslated regions of mRNA in post-transcriptional regulation

In 1953, Francis Crick and James Watson proposed the structure for deoxyribonucleic acid, or DNA. In their seminal pape on the structure, they suggested a means by which cells make an identical copy of DNA with each cell division; later, they suggested a code that determines the structure of proteins [330, 331]. How DNA encodes the structure of proteins was still unclear. In 1958, Francis Crick wrote in a letter, "Watson said to me, a few years ago, 'The most significant thing about the nucleic acids is that we don't know what they do.' By contrast, the most significant thing about proteins is that they can do almost anything." This viewpoint arguably influenced the protein-centric view of cell biology for the next half century. Crick might have been surprised to know that the noncoding regions of mRNA contain important regulatory and processing elements that can modulate protein function without altering amino acid sequence. Nucleic acids in the 3' UTR control critical post-transcriptional gene processes. From biochemistry, to recombinant technology, to deep sequencing, our understanding of how the noncoding bases in 3' UTRs impact gene regulation has expanded exponentially in the past few decades.

In this introduction, I will briefly overview how 3' UTR biology was historically investigated, highlighting technology that assisted in the discovery. I will focus on how advances in DNA sequencing technologies have expanded the discovery and characterization of 3' UTRs across eukaryotes, specifically in plants. I will end on how current biotechnological tools (from genome sequencing to a plethora of massively parallel reporter assays that interrogate function) coupled with computational machine learning will facilitate engineering specific

terminators for plant synthetic biology applications. Finally, I will end with current limitations in the field and propose solutions that can predict which combination of all genetic elements to use for specific, temporal, and tunable expression of genes.

1.2.1 Laying the groundwork of 3' untranslated region in cleavage and polyadenylation

By the summer of 1961, the concept of **messenger RNA (mRNA)** was born into the world. Through a series of experiments, Sydnew Brenner, Francis Crick, Francois Jacob, and Matthew Meselson uncovered mRNA's role in gene regulation and successfully isolated it [33, 132]. They proposed that messenger RNA was like a tape that copied information from DNA and then carried that information to the ribosome, the site of protein synthesis, which read and followed the instructions to make the encoding protein. With the discovery of mRNA, Francis Crick articulated a set of principles known as the Central Dogma of Molecular Biology to articulate the flow of information from DNA to proteins. DNA codes for RNA, and RNA codes for proteins. In other words, genes ultimately code for proteins.

Until the 1970s, the prevailing theory was that information transfer from DNA to proteins happens exclusively through translation of the coding region of mRNAs into amino acids of proteins. Yet early nucleotide sequencing revealed that mRNA contains additional nucleotides in the 5'- and 3'-end that are not translated [3, 198, 211] . Unlike bacterial genes that transcribe into mature RNA, eukayotic genes transcribe pre-mRNAs that are nonfunctional and much larger than the mature form. In 1971, a 3' poly(A) segment, later coined as the poly(A) tail, was discovered in polyribosomal mRNA through targeted enzymatic digestion. Translationally active ribosome-associated mRNA were isolated and then digested with RNAse that cuts only at C and U residents or G residues to reveal a resistant fraction of adenine residues [4, 75, 180]. Soon it was shown that the adenine residues were added to nuclear RNA by poly(A) polymerase (PAP), providing the first example that nuclear pre-mRNA is processed after transcription [60, 135, 337]. The discovery of poly(A) addition was followed by the discovery of the addition of the 5'-methylated cap at the 5' end mRNA [85, 86, 241, 276, 333]. In 1977, Richard Roberts and Phillip Sharp discovered that some protein-coding genes have intervening sequences of noncoding DNA. Micrographs of

DNA-RNA hybrids revealed spaces in the DNA that bound to no RNA, forming DNA loops that protrude away from the hybrid. These skipped regions came to be called **introns**. Collectively, these experiments showed that mRNA must be processed in the nucleus for proper translation, and that the non-coding segments in the 5', 3' and intron regions were critical for proper processing.

Understanding the noncoding bases in the 3' end was fundamental for mRNA technology development. Most notably, the discovery of the poly(A) tail was pivotal for isolating and sequencing mRNA. Ribosomal RNA makes up to 80%-90% of total RNA content in a cell, and mRNA is only 3-7%[66], making it challenging to study mRNA biology and gene regulation. Even though the function of the poly(A) tail was still unknown at the time, the poly(A) tail served as a natural tag to isolate mRNA by oligo(dT) affinity chromatography away from the bulk ribosomal RNA that lacks the poly(A) tail [16]. Before recombinant DNA technology [53], the only way to isolate individual mRNAs was to select a tissue that had pronounced and selective gene expression so that a particular mRNA is unusually abundant [247], significantly limiting the number of genes one could study.

Oligo(dT) capture of polyadenylated mRNAs facilitated the first example of mRNA sequencing. Nick Proudfoot and colleagues were also the first to use oligo(dTs) for *in vitro* complementary DNA (cDNA) synthesis, turning the mRNA into a complementary single-stranded DNA [243]. They extracted six globin mRNAs to sequence the 3' end of the mRNA for the first time. Using the 2D chromatographic "fingerprinting" technique of DNA sequencing techniques from Fred Sanger's laboratory [35, 266], Proudfoot discovered the presence of the **poly(A) signaling hexamer (AAUAAA)** close to the 3' end of the six globin mRNAs, later proving that it signals for polyadenylation and transcription termination [246, 248, 244]. These early sequencing results also showed that the stop codon does not define the mRNA 3' end, but rather it is the 3' UTR that possessed the gene's polyadenylation signal [245].

The discovery of polyadenylation was quickly followed by the discovery of alternative polyadenylation. Scientists discovered alternative polyadenylation in the 3' UTR of eukaryotic mRNA through northern blotting and nuclease protection assays. Northern blotting provides a gel fractionation image of the specific gene's mRNA output that gives information

on the size and quantity of each mRNA. Briefly, mRNA isoforms from genes were recovered and hybridized to the cloned DNA of the gene and then digested with S1 nuclease to map sites that are protected from cleavage, allowing visualizations of the complete set of mRNA isoforms generated from a gene. With this technology, scientists could then map introns, and the 5' and 3' UTR ends of transcribed genes. Soon after, scientist found that multiple polyadenylation sites exist in 3' untranslated end of a transcript, discovering alternative polyadenylation [72, 311].

1.2.2 Dissecting poly(A) signals in mammals

Advances in recombinant DNA technology also led to deletion constructs and site-directed mutagenesis approaches to probe how much sequences inside 3' UTR influence polyadenylation [80]. Soon after the discovery of AAUAAA, more cleavage and polyadenylation elements were discovered: the GU-rich **downstream sequence element (DSE)** region downstream of the cleavage site of 3'end was shown to enhance cleavage and polyadenylation [94, 95, 205], and the **upstream sequence element (USE)** was shown to enhance 3'-end processing efficiency [40, 216, 316, 317]. Finally, the nucleotides of the **3' cleavage site (CS)** can also influence the efficiency of cleavage [48]. Together, the positioning and patterning of these signals together can drastically affect the location of cleavage and polyadenylation by thousands of nucleotides [278, 306].

These early experiments in investigating mRNA polyadenylation signals were all in mammals. Extensive biochemical work in other eukaryotes revealed that the mammalian pattern of **USE-AAUAAA-DSE** is generally conserved across eukaryotes, with a slight exception for plants.

1.3 Dissecting poly(A) signals in plants

Plant cleavage and polyadenylation patterns do not share the exact structure as other eukaryotes **(Figure 1.2)**. Plants do not cleave and polyadenylate transgene products containing animal-derived polyadenylation signals or 3' UTRs [126]. Plant gene transcripts, on average, also have a higher instance of multiple poly(A) signals than mammalian genes

[109, 272]. Nevertheless, early work in plants also found that AAUAAA is a polyadenylation signal [64, 69], albeit not as conserved as expected. Through *Agrobacterium*-mediated transfection into tobacco leaves, scientists studied how viral genes were transcribed and processed in plants. S1 nuclease mapping from the T-DNA derived isolates revealed that the poly(A) signal AAUAAA acts as a polyadenylation signal. However, the sequence is not a universal feature near many polyadenylation sites in plants. With 46 plant gene partial 3' UTR sequences available at that time, scientists scanned for putative poly(A) signals and found that only a third of these genes had the AAUAAA near the cleavage site and a significant portion had no AAUAAA-life motif at all [143].

Despite the growing number of sequenced plant genes, early functional studies on cleavage and polyadenylation in plants were done on only four genes, of which only two were plant genes: cauliflower mosaic virus (CaMV) 35S, the pea rubisco (ribulose bisphosphate carboxylase and oxygenase) small subunit gene (rbcS) E9, the *Agrobacterium tumefaciens* octopine synthase (ocs) gene, and maize storage protein coding gene (zein 27-kDa). Nuclease protection assays and reverse transcription/polymerase chain reactions (RT-PCR) on these genes allowed the effects of deletions and mutations on the function of particular poly(A) sites to be evaluated. While only two of the genes are eukaryotic genes (not viral or bacterial), functional assays on all genes led to similar models, suggesting that the results can be generalized to all plant poly(A) signals. It is important to note that RNA analysis conducted using northern blotting fails to provide sufficient information on the effect of mutations on poly(A) site function [124]. Nevertheless, the polyadenylation *cis*-regulatory elements in plants were divided into three distinct elements: The **far upstream elements (FUEs)**, the **near upstream elements (NUEs)**, and the **cleavage site element (CS)**.

1.3.1 FUE-NUE-CS tripartite polyadenylation signal in plants

The NUEs are A-rich sequences 6-10 nucleotides that are situated 10-40 nucleotides from their corresponding poly(A) site. NUEs are functionally analogous to the mammalian polyadenylation signal AAUAAA [175, 341]. However, systematic point mutation analysis indicated that there is much variation in the AAUAAA motif in the CaMV poly(A)

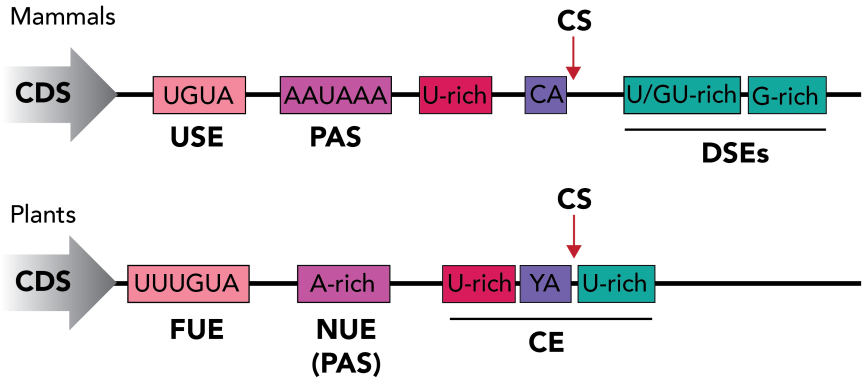


Figure 1.2: Schematic representation of the 3' UTR regions of mammals and plants. The cis-acting elements recognized by the 3' end processing machinery are displayed and consensus sequences are given. The following elements are indicated: CDS, coding sequence; CE, cleavage element; CS, cleavage site; DSE, downstream element; FUE, far upstream element; NUE; near upstream element; PAS, polyadenylation signal; USE, upstream element.

signal, and that these mutations are tolerated more in plants than in mammalian systems [125, 262]. The FUEs are required for efficient usage of poly(A) sites, as deletion of an FUE can decrease poly(A) site efficiency by an order of magnitude. Large deletions in the FUE have dramatic effects on polyadenylation, but smaller mutations (point, deletion, or linker scanning) have very subtle effects. Unlike the USE of mammals, the relative location of the FUE to the NUE can be variable (from 13 to 100 nts upstream of the NUE), and FUEs lack highly conserved consensus sequences FUEs. Plant CSs are usually situated in U-rich regions of the 3' UTR and contain a Y/(C,A) dinucleotide and behave very similarly to those in animals and yeast [321]. Unlike the mammalian DSE, sequences after the polyadenylation site in plants do not appear to play a role in mRNA polyadenylation [124, 262]. However, plant FUEs seem to share sequence properties with DSE required for polyadenylation in mammals, a decided UG/U-richness, suggesting that there might be an evolutionary conserved *trans*-acting factors that can recognize DSEs in animals and FUEs in plants.

1.3.2 Cleavage and polyadenylation machinery

The 3' regulatory regions described above guide the cleavage and polyadenylation protein machinery. In eukaryotes, a multiprotein complex of more than 20 proteins, named the Cleavage and Polyadenylation Molecular Complex (CPMC) or Polyadenylation Complex (PAC), recognizes and interacts with *cis*-elements to cleave and polyadenylate pre-mRNA. These proteins have been deeply characterized in yeast and mammals [196]. The CPMC in mammals is composed by four subcomplexes: Cleavage and Polyadenylation Specificity Factor (CPSF), Cleavage Stimulatory factor (CSTF), mammalian Cleavage factor I (CFIm), and Cleavage factor II (CFIIm). Additional core factors include scaffolding and enzymatic subunits such poly(A) polymerase, FIP1, symplekin, and nuclear poly(A) binding protein.

CPSF and CSTF work cooperatively to promote cleavage; CPSF recognizes the poly(A) signal (AAUAAA-like variants) and CSTF recognizes the U/GU rich regions in the DSEs to guide cleavage between respective binding sites [369]. CPSF recruitment is central, as it constitutes the core processing complex required for both the cleavage and subsequent polyadenylation reactions. Unlike CPSF, CSTF is essential only for cleavage reactions [297]. CFIm is a tetramer of two subunits that binds the UGUA motifs upstream of the cleavage site and is indispensable for proper cleavage, where binding to the UGUA motif promotes recruitment of CPSF and CSTF. CFIm also helps direct the choice of PAS [152, 163]. CFIIm plays a role in alternative polyadenylation, with both of its subunits (Clp1 and Pcf11) promoting the use of proximal poly(A) sites (PAS). Depletion of both CFIIm subunits lead to increased use of distal PAS across transcripts [41, 176].

The polyadenylation reaction requires the cleaved pre-mRNA template, CPSF, poly(A) polymerase, and poly(A) binding protein. Poly(A) polymerase (PAP) adds adenosines to the 3' hydroxyl group of RNA in a template-independent manner [164, 320], with the length of adenosines added varying among species (200-250 in humans, 70-80 in yeast). The speed of assembly of the functional cleavage and polyadenylation complex is dependent on the strength of the pA site and the sequence architecture around the pA site [45, 227].

Biochemical characterization of plant cleavage and polyadenylation complex is challenging due to the low expression levels in plant tissue and the lack of efficient purification

approaches able to isolate protein factors while maintaining their natural interacting partners. As a result, the gears and mechanisms of the plant cleavage and polyadenylation machinery are still not fully elucidated. Despite the limitation, protein interaction assays (including yeast two-hybrid, in vitro pull-down, immunoprecipitation, and affinity purification assays) provided evidence that the interaction topology of polyadenylation factors in *Arabidopsis* is similar to that of yeast and humans [125], although some unique features have been noted in higher plants.

Plants possess most, but not all, of the orthologs encoding the CPMC [125, 130, 368]. Unlike the CPMC in mammals, in which each of the subunits are encoded by a single gene, the subunits in plants are encoded by gene families. Some plant genes encode at least two isoforms of the same subunit [28, 130]. Unlike the mammalian homologs, some plant CPMC subunits are dispensable for growth and development (such as CSFP30, CstF77, CstF64, and several poly(A) polymerase isoforms) [128]. This unique distinction between plants and mammals allowed the study of plant mutants that lack CPMC proteins. The impact of mutations (including full knockouts) of different CPMC subunits demonstrated a wide array of phenotypes that are associated with global changes in gene regulation, such as altered root and flower development, altered hormone response, and different responses to stress (Hunt, 2020). While these phenotypes of CPMC mutants could be the result of alternative poly(A) site usage that in turn directly affects expression of genes responsible for phenotype, the large scope of changes in poly(A) site usage and overall gene expression makes it difficulty to tease out connections and order of operations. In other words, is alternative polyadenylation responsible for phenotypes or are changes in gene expression responsible, or a complicated feedback of both? It is hard to assess the cause-and-effect in most cases since CPMC associated phenotypes are controlled by a large network of genes [129].

The *Arabidopsis* genome encodes all the protein factors analogous to the mammalian CPSF, but studies aimed at studying the *in vivo* composition and interaction patterns among the plant CPSF components are limited to low expression level of CPSF and the lack of efficient purification approaches able to isolate the protein factors involved and maintain their *in vivo* interactions. *Arabidopsis* cell cultures were used instead. The *Arabidopsis*

homolog of CPSF30 (AtCPSF30) has been confirmed as an RNA-binding protein with an affinity for U-rich sequences in the FUE. AtCPSF30 is implicated in binding NUE-residing motifs, and mutations in the subunit resulted in the choice of unusual poly(A) signals [304]. AtFY, another ortholog of the CPSF subunit protein WDR33, is involved in the recognition of the canonical NUEs in *Arabidopsis* [362]. The choice of poly(A) signal relies on AtCPSF30 and AtFY interactions [362], with double mutants in plants generating up to 50% more alternative polyadenylation events. Unlike their mammalian counterparts, AtCPSF73 and AtFY are also involved in plant-specific processes like flowering and plant female development, suggesting that plants evolved variants of CPSF to deal with specific environmental cues. *Arabidopsis* orthologs of CFI, AtCFI, recruit the *Arabidopsis* ortholog of poly(A) binding protein as well [81]. The C' terminus of AtCSTF77 interacts with AtCPSF30 *in vivo*. The *Arabidopsis* orthologs of CSTF subunits interact *in vitro* and have the ability to bind RNA [26]. This function might be plant specific since RNA binding has not been reported for either mammalian or yeast CSTF77 subunit. Genome-wide poly(A) site usage of CSTF77-2 mutants in *Arabidopsis* revealed extensive usage of distal poly(A) sites and diminished transcription termination efficiencies [363].

Investigations of the evolutionary conservation of the plant CPMC across 10 plant species found that the number of genes encoding plant polyadenylation factors increased from "lower" to "higher" plants, with gene expansion in higher order plants biased to some polyadenylation factors [130]. Plant polyadenylation complex consists of a relatively constant core (shared with mammals) and a panoply of peripheral subunits that are somewhat distinct for each species. The evolutionary expansion of the complex might explain why plants are more forgiving in recognizing variations of the poly(A) signal. An additional layer of complexity of the plant CPMC is the existence of partial protein isoforms, due to either through alternative RNA processing or coding by separate genes [130]. These partial proteins possess some but not all of the functionality of their respective "complete" versions, potentially affecting the functioning of other subunits and thus perturbing the complex architecture.

1.3.3 Sequencing accelerates understanding of 3' end processing and polyadenylation

Traditional genetics and biochemical approaches helped elucidate some of mechanisms behind poly(A) signal selection and cleavage. The surge of improved sequencing technologies at the beginning of the second millennium catapulted the endeavour. It is abundantly clear that sequencing expanded the horizons of genetic exploration, along with the development of bioinformatic tools and algorithms. However, genome-wide analyses often only provide correlation and lack direct functional experimentation to prove causation. Sequencing is useless unless we can interpret it properly.

Once Sanger sequencing was invented [267], more mammalian and plant mRNAs were sequenced to determine conservation of more *cis*-regulatory elements in untranslated regions. In 1980, the first full 3' UTR sequence of mRNA was published [209]. As sequencing became less expensive, the number of genes sequenced grew exponentially. Having the DNA sequences from genes of different organisms allowed for the first comparative genomic analysis [226], in which the rate of substitutions reflects the degree of functional constraint. Early comparative genomic studies found a higher degree of substitutions in untranslated regions than in coding regions, with the exception of short sequence elements involved in polyadenylation and splicing. Despite the high substitution rate, they also found a high degree of sequence homology in the 3' UTRs [212]. Later studies also found that the 3' UTR sequences of homologous genes coding for actin proteins are highly conserved across organisms, but the 3' UTR sequences are highly divergent across actin isoforms expressed across different tissue and function [352]. Conservation in the 3' UTR region highlights their regulatory roles, but divergence in the 3' UTR sequence of isoforms suggests additional genetic information to distinguish highly similar proteins [203].

Advances in sequencing improved genome-wide understanding of polyadenylation and the curious case of conservation. By the late 1990s, the availability of genomic, full-length cDNA and **expressed sequence tag (EST) sequences** from large scale genome sequencing projects made it possible to search for poly(A) signals using bioinformatics tools. ESTs are short segments of DNA that represent coding regions. In 1998, Gautheret et al. demonstrated that one can find the distribution of alternative mRNA forms through EST sampling

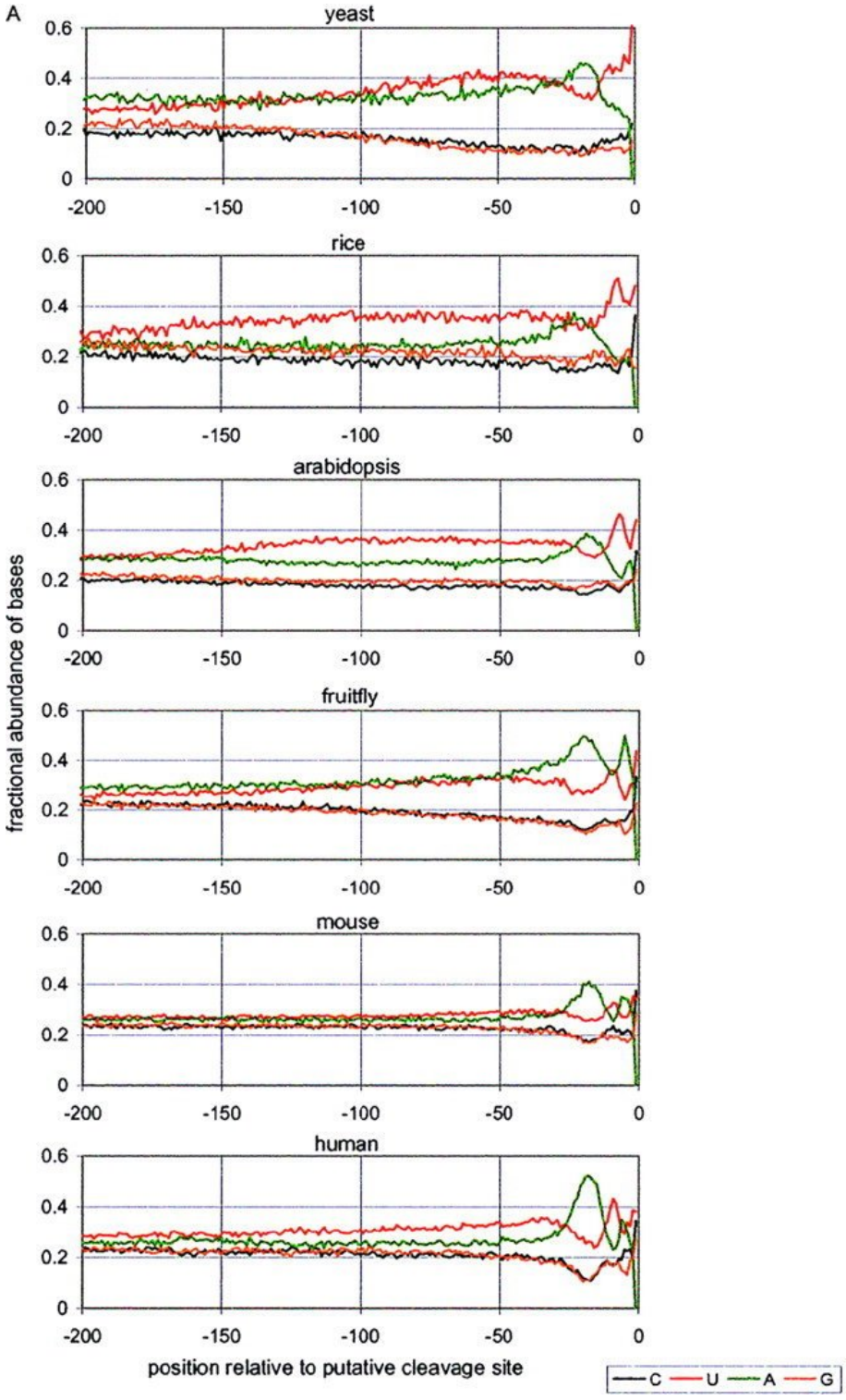


Figure 1.3: From [99], this figures shows the single-nucleotide frequencies preceding the 3'-end-processing sites as determined by alignment of 3'-ESTs generated from yeast (1,352), rice (1,246), *Arabidopsis* (4,069), fruitfly (3,236), mouse (6,029), and human (4,427) cDNAs. Positions are given relative to the putative 3'-end-processing site. (A) Sequences aligned on the 3'-most end of the ESTs.

(Figure 1.3). They compared 164,000 human 3' ESTs and clustered them into homogenous groups, where clusters of overlapping ESTs were analyzed for distinct poly(A) sites in humans [92]. Outside of human genetics, EST databases also allowed comparative genomic studies on poly(A) signals. Mining EST databases across 6 eukaryotic species (yeast, rice, *Arabidopsis*, fruit fly, mouse, and human) revealed that the use and conservation of the canonical AAUAAA element varied widely and was especially weak in plants and yeast. The same study also found that plant polyadenylation signals are more similar to those in yeast than in animals, with both content and arrangement of the elements. Across species, they found that polyadenylation signals appear to consist of aggregates of multiple elements. They suggest that no single exact sequence is universally required for 3' end processing. Rather, the total efficiency is a function of all elements, where an inefficient part of one element can be compensated by strong contribution of another [99]. Together, these findings suggest that conservation style approaches to finding *cis*-regulatory 3' UTR elements is not that informative for function.

1.3.4 Whole genome sequencing and bioinformatic modeling increases understanding of polyadenylation signals in plants

By 2000, the completion of the *Arabidopsis thaliana* genome assembly necessitated genome annotation, which is generally based on the identification of functional RNA and coding sequences. The *Arabidopsis* genome was annotated using *in silico* gene-finding methods, comparison to EST and protein databases, and manual reconciliation of the data [122]. The availability of genomic, full-length cDNA and EST data also made it possible to search for poly(A) signals using bioinformatic tools [25, 92, 99, 106]. By 2005, using two large datasets of 3' UTR sequences in *Arabidopsis* (one with 8,160 ESTs with authenticated poly(A) sites, the other with 16,211 full-length cDNA downloaded from the *Arabidopsis* Information Resource TAIR), scientists computationally scanned for poly(A) signals derived from conventional genetics [21, 186, 263]. This genome-wide scanning approach confirmed the **FUE-NUE-CS tripartite signal** in plants, while further dissecting the nucleotide distribution patterns around the CS and poly(A) sites **(Figure 1.4)**. Similar analysis

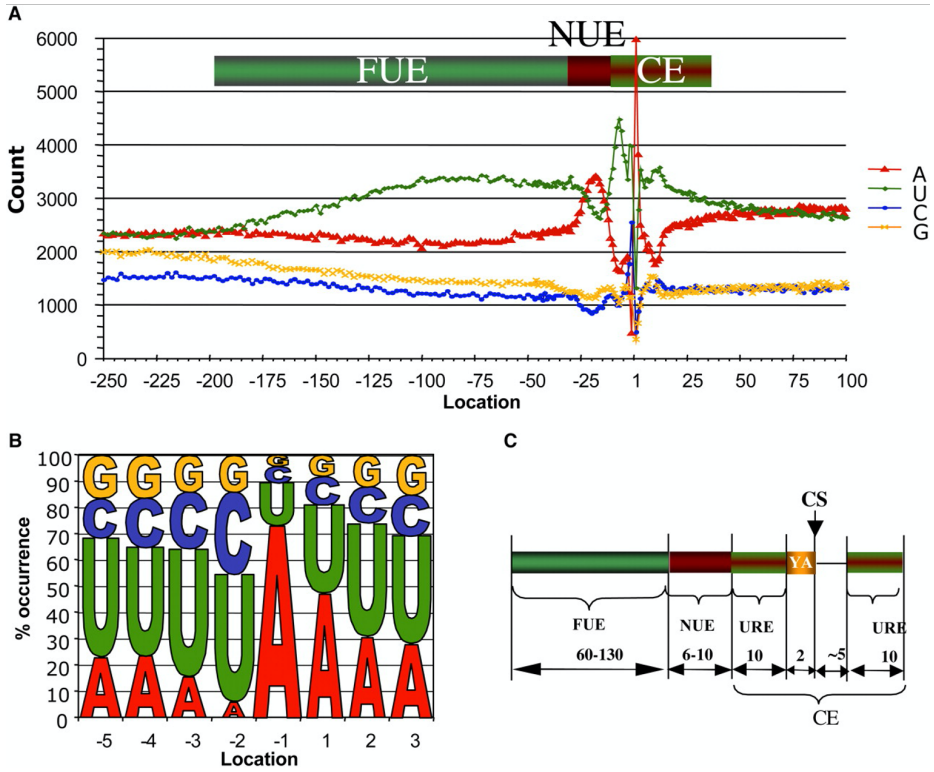


Figure 1.4: From [186], this figures shows the single-nucleotide profile of 3'-UTR and a (then) current model of plant poly(A) signals. A) Single-nucleotide scan from positions -250 to +100 in the whole UTR + downstream region. Distinct profiles flanking the CS are now named CEs. B) Sequence logo generated from the actual percentage of each of the four nucleotide's occurrence in the 8-K dataset, indicating preferred nucleotides flanking the CS (-5 to +3 nt). C) A current model for *Arabidopsis* mRNA poly(A) signals. URE, U-rich regions, which are found flanking both upstream and downstream of the CS.

followed for polyadenylation in rice to show the tripartite signal is robust regardless of plant model [280]. Once again, these studies in *Arabidopsis* and rice confirmed how the poly(A) signal AAUAAA is not strictly conserved in 3' UTRs of plants. Compared to 50% of mammalian genes that have the AAUAAA, only 10-15% of genes in *Arabidopsis* and rice have AAUAAA as their poly(A) signal.

1.3.5 Next generation sequencing improves 3' UTR analysis

The development of next generation sequencing (NGS) of DNA and RNA in the early 2000s revolutionized our ability to study and analyze genomes. Because they are massively parallel, NGS technologies provide highly efficient, rapid, and low cost sequencing beyond the reach of the traditional sequencing developed in the late 1970s. Next generation sequencing technology, like the Illumina platform, coupled with bioinformatics algorithms in downstream genome assembly, has also significantly increased the number of sequenced plant genomes [298]. First, NGS technologies helped establish genome-wide analysis related to gene expression and transcript profiles, which are collectively known as RNA-seq [22]. Sequences derived from RNA samples are mapped to a reference genome, where the number of reads that map to each gene correspond to its expression level[165]. In practice, a population of RNA is converted to a library of cDNA fragments with adapters attached to one or both ends for sequencing. With or without amplification, the molecules are sequenced from one end (single end sequencing) or both ends (paired-end sequencing). Following sequencing on an NGS platform, the resulting reads are either aligned to a reference genome or a reference transcript. RNA-seq is the first sequencing based method that allows the transcriptome to be surveyed in a high throughput and quantitative manner. RNA-sequencing not only offered both single-base resolution for annotation and 'digital' gene expression levels, but it was also cheaper and more scalable than large-scale Sanger EST sequencing [328]. Besides gene expression, RNA-seq can be adapted to analyze transcript boundaries, intron/exon junctions, profiling of noncoding RNA, nascent transcripts, ribosome associated mRNA, and polyadenylation [273]. For our purposes, the 3' boundaries of many transcripts were mapped by searching for poly(A) tags in RNA-seq datasets [326]. However, in its original

Method	Fragmentation	Adapting	Internal priming	Ease	Sequencing platform
3SEQ	Heat shearing	DNA ligation	Yes	Medium	Illumina
Mangone et al.	DpnII	DNA ligation	Yes	Medium	454
3P-seq	–	RNA ligation	No	Low	Illumina
PAS-seq	Heat shearing	RT + template switching	Yes	High	Illumina
SAPAS	Heat shearing	RT + template switching	Yes	High	454/Illumina
Wu et al.	NlaIII or TaiI	DNA ligation	Yes	Medium	Illumina
MAPS	–	RT + second strand synthesis	Yes	Medium	Illumina
Poly(A)-seq	–	RT + second strand synthesis	Yes	High	Illumina
3' seq	Heat shearing	DNA ligation	Yes	Medium	Illumina
A seq	RNase I	RNA ligation	Yes	Medium	Illumina
3' T fill	Heat shearing	DNA ligation	Yes	Medium	Illumina
3' READS	Heat shearing	RNA ligation	No	Medium	Illumina
3PC	Heat shearing	Circularization	Yes	Medium	Illumina
PA-seq	Heat shearing	DNA ligation	Yes	Medium	Illumina
EXPRSS	Covaris shearing	DNA ligation	Yes	Medium	Illumina
PAT-seq	RNase T1	RNA ligation	No	Medium	Illumina
WTTS-seq	Heat shearing	RT + second strand synthesis	Rare	Medium	Ion Torrent

Table 1.1: Details of each 3'-enriched RNA-seq method for global pA site profiling

form, RNA-seq is not as suitable for identifying poly(A) sites precisely and extensively, a result of relatively low overall read coverage of the 5' and 3' end of genes. Thus, better library construction protocols were designed to capture 3' ends of mRNA for direct profiling of genome-wide poly(A) sites, including Direct RNA sequencing (DRS) [231, 281], 3P-seq [134, 313], 3'READS [118], PAT-seq [107], and TAIL-seq [42]. Collectively, these technologies are known as 3' enriched RNA-seq, which can be classified into two categories based on the strategy used to enrich the 3' end: oligo(dT) capture methods or RNA manipulation methods [118].

NGS sequencing established methods and genomic assays that were able to map the 3' end of mRNA transcriptome-wide, revealing that more than half of the human and mouse genes generate alternative mRNA isoforms but encode proteins with identical amino acids [68, 103, 118, 178] **(Table 1.1)**. Similar work sequencing plant mRNA isoforms also es-

tablished the expansive degree of alternative polyadenylation in plant genomes. By 2011, scientists conducted the 3' end deep sequencing protocol to query the junctions of 3' UTR and poly(A) tails and confidently maps the poly(A) tags to the annotated genome of *Arabidopsis* [304, 343]. To study *Arabidopsis* poly(A) sites on a genome-wide basis, short DNA tags that include the mRNA-poly(A) site junction (called poly(A) tags, or PATs) were prepared and sequenced. However, with 3' end sequencing, the intrinsic template switching and DNA-dependent DNA-polymerase activities of reverse transcriptases and the oligo(dT)-dependent internal priming can cause well-established artifacts when sequencing antisense RNAs, splicing events, and RNA 3' ends [223]. To address this and get an accurate genome wide approach, another group conducted direct RNA sequencing (DRS), which uses native RNA as the sequencing template where the sequence is read by imaging complementary fluorescent nucleotides. Each technique discovered that greater than 70% of the genes in *Arabidopsis thaliana* have two or more poly(A) sites, in which roughly 17% are located within 5' UTRs, coding sequences (CDS), and introns [281, 343]. Other crop genomes, like rice and maize, were analyzed either through PAT-seq, RNA-seq, or DRS for poly(A) site mapping [82, 133, 322], demonstrating the extensive degree of alternative polyadenylation across plant genomes. The impact of alternative polyadenylation on gene expression in plants is profound, affecting developmental control of flowering time, seed dormancy, root/leaf development, and stress response [36, 57, 181, 348, 361, 366, 367]. Interestingly, there were also subtle differences between plant species.

1.3.6 Use of long read sequencing to resolve complete plant 3' UTRs

A major limitation of NGS sequencing is that it is short read based, with reads spanning 100 to 250 base pairs, making it harder to infer structures of full-length transcripts. The average 3' UTR length in maize is 350 bases alone and almost 470 in rice [293]. The PacBio isoform sequencing (Iso-Seq) method provides long read sequencing reads with uniform coverage, able to define full-length transcripts with no assembly required. Historically, the cost of Iso-seq has been prohibitive. Despite the higher cost, Iso-seq has enhanced understanding of the transcriptome of several plant species, including sugarcane [116], arabica coffee [51],

opium poppy [350], and jojoba [6]. Recently, several studies collected and reanalyzed long reads from Iso-seq into comprehensive databases such as Plant ISOform sequencing database (PISO) [79] and ISOdb [347]. Currently, there are 19 plant species listed in PISO: *Amborella trichopoda, Arabidopsis thaliana, Beta vulgaris subsp. Vulgaris, Chenopodium quinoa, Coffea arabica, Fragaria vesca, Gossypium barbadense, Hevea brasiliensis, Panax ginseng, Phyllostachys edulis, Sorghum bicolor, Triticum aestivum, Zea mays, Allium sativum, Astragalus membranaceus, Dipteryx oleifera, Nepenthes ampullaria, Nepenthes rafflesiana and Salvia miltiorrhiza.* Using Iso-seq, 7700 genes containing two or more polyadenylation sites have been detected in *Sorghum bicolor* (great millet) [2]. In allopolyploid cotton, 6935 genes have at least five poly(A) sites [323]. Despite the strength of Iso-seq to resolve all APA isoforms in longer plant genomes, the quantification of APA site preference still depends on NGS due to the low sequencing depth of Iso-seq [370].

Genome size, including the coding and noncoding parts of the genome, has dramatically increased during the evolution from worms to humans. The sequence space occupied by the 3' UTRs has also dramatically expanded during the evolution of higher organisms and correlated with cellular complexity of organisms [202]. For example, while humans and worms encode a similar number of genes, the average 3' UTR length in worms is 140 nucleotides and 1,200 in humans. Across plant genomes, average 3' UTRs length can vary from a few to thousands of nucleotides [242, 325], implicating the need for long-read sequencing to resolve the 3' UTR space.

The tallies of whole plant transcriptomes significantly improved our understanding of poly(A) signals. The profiles of these signals can be used to build computer models that can predict poly(A) sites in newly sequenced genomes, potential APA sites in genes of interest, and identify/mutate unwanted poly(A) sites in target transgenes to facilitate crop improvements.

1.3.7 Computational modeling and predicting poly(A) sites from genomes

Improper regulation of transgene sequences in target plants is a major issue in plant processing, especially if the transgene is derived from exogenous sources [348]. For example,

initial efforts to express *Bacillus thuringiensis* toxin in plants failed because of truncated mRNA generated by improper cleavage and polyadenylation of the bacterial gene transcript [71]. Thus, it is critical to be able to predict unwanted polyadenylation sites of a gene inserted into plants. To address the need to predict poly(A) sites, computer algorithms using conventional machine learning were designed based on the established plant poly(A) signals from EST databases [138, 159]. Such algorithms include Hidden Markov Models (HMM), Support vector machines (SVM), Bayesian networks, and random forests [342]. Poly(A) Site Sleuth (PASS), the first model for predicting poly(A) sites in plants, was a generalized HMM [138]. An updated version from the same group had a more versatile classifier-based model in which polyadenylation parameters from different species can be incorporated in the model [137, 136].

The deluge of 3' end sequencing data also lent itself to more advanced modeling frameworks like deep learning models to predict poly(A) sites. **Deep learning (DL)** is a class of ML algorithms that uses multiple layers to progressively extract higher-level features from raw input [170]. In recent years, DL models have been shown to outperform traditional machine learning methods, owing to their direct and automatic feature extraction and scalability with large amounts of genomic data. The input frameworks of early ML methods like SVM or random forest rely heavily on manually designed features, whereas DL-based methods learn hidden features without prior knowledge of sequence motifs. Most DL methods predicting poly(A) sites in genomes use convolutional neural networks (CNNs), such as DeepPolyA [90]), ConvNet [173], DeeReCT-PolyA [346], DeepPASTA [12], DeepGSR [145], and APARENT [32]. CNNs work well with inputs that have spatially invariant patterns, such as images or DNA sequences, and are effective at detecting spatial patterns in input data. The important features in regulatory sequences are thought to be specific combinations of consecutive base pairs (let's call them "motifs"), which makes CNNs well-suited to the task of identifying significant motifs, like cleavage and polyadenylation signals. Additionally, DL-based computational tools have been developed to identify and quantify PAS by leveraging RNA-seq datasets across experiments and tissues, helping to consolidate data collected worldwide [12, 90, 104, 345, 356]. The advances of single cell RNA sequencing (scRNA-seq) techniques, especially those with 3' tag-based protocols, have also spurred off

more diverse computational tools to profile APA in single cells, improving our understanding of APA across development [5, 91, 93, 344].

Despite the immense computational and sequencing progress, most of these DL studies and tools were conducted on mammalian genomes. Tools that are tailored for plant poly(A) site prediction are limited, especially since conservation of poly(A) signals in plants is very low [49]. For example, the most dominant AAUAAA appears in less than 10% of polyA sites [305]. To date, the most recent tool developed to profile genome-wide polyadenylation found that plant poly(A) sites have much higher micro-heterogeneity than animal ones and more species-specific patterning of poly(A) motifs. [357, 359]. Still, as sequencing becomes less expensive and better plant genome assemblies are published, the next 10 years will see a surge of plant-specific DL models to predict plant polyadenylation sites across the plant kingdom.

1.4 Massively parallel reporter assays empower functional genomics

Although we have extensive information about poly(A) site location and how cleavage and polyadenylation occurs, we still lack an interpretable and quantitative model that integrates *cis*-regulatory sequences and predicts cleavage positions and isoform abundances. In other words, we lack models that predict how specific sequences contribute to proper cleavage and polyadenylation. Such models are a critical need as alternative polyadenylation is highly prevalent in plant genomes and plays a huge role in development, environmental response, and growth [127]. Understanding and predicting why some sites are chosen over others will also benefit future plant genetic engineering projects [62].

Validation and functional analysis of every 3' UTRs from a plant for activity *in vivo* would be a major bottleneck if these elements had to be cloned and tested individually via traditional methods such as transient transgenic reporters, gene-targeted reporters, and knockouts. The low-throughput high-cost nature of these approaches diminishes their utility. Using NGS, **massively parallel reporter assays (MPRAs)** bypass these limitations and allow for high-throughput functional experimentation [207, 237, 275, 291, 334]. MPRAs are a functional genomics technique that uses a reporter assay with a sequencing-based readout to measure the activity of thousands of *cis*-regulatory elements in a single experiment. The

power of MPRAs is derived from their multiplexing approach. The general protocol involves first pairing your tested region to a unique barcode, with multiple barcodes per variant. The barcoded variants are cloned into a reporter plasmid to make the final pooled library that will be tested in a cell type or tissue of interest. Barcode abundances are subsequently quantified through RNA-seq (and normalized to the input DNA barcode abundances). Since the output sequencing only reads the short barcodes associated with each tested construct, this approach lets you assay up to hundreds of thousands of different sequences simultaneously. A variation of the MPRA, called **STARR-seq (self-transcribing active regulatory region sequencing)**, has sequences of interest cloned downstream the coding sequence and upstream of the poly(A) tail [15]. Originally designed to test enhancer activity, STARR-seq tests libraries of sequences in the 3' UTR, so measurements are likely confounded by RNA stability. However, in the plant version of STARR-seq, enhancer elements are best tested upstream of the minimal promoter and not self-transcribed (even though the name is STARR-seq, the assay is a barcoded *cis*-regulatory MPRA) [141].

While used extensively to catalog enhancer and promoter activity, MPRAs have been adapted to test mRNA functionality as well. Massively parallel reporter assays identified cis-regulatory elements for DNA transcription [15, 61, 101, 141, 142, 151], mRNA stability [182, 230, 252, 274, 355, 371], splicing [232, 259], and translation efficiency [105, 264]. MPRAs can also test random sequences to explore the effects of sequences outside of native sequence contexts [105, 238, 270]. An ultimate goal of decoding *cis*-regulatory elements in this high throughput scalable manner is to also develop computational models that explain and predict the phenotype tested [158]. With an increasing number of available MPRA datasets, one can develop data-driven models that, given a DNA sequence, can predict activity [218]. For example, models learned on MPRA data can predict alternative splicing frequencies [259], enhancer activity [101, 141], and promoter activity [83, 142, 155].

1.4.1 Massively parallel reporter assays conducted in plants

Massively parallel reporter assays in plants have strictly focused on regulators of transcription by testing either promoters or enhancers derived from accessibility assays [141, 142,

255, 296, 302, 308]. What has been fundamentally lacking in plant genomics are MPRAs that functionally assay sequences in the 3' UTR, 5' UTR, and intronic space for perturbing mRNA functionality. Despite the pivotal role the 3' UTR plays in mRNA stability and degradation, no one had yet conducted a massively parallel reporter assay of 3' UTR activity in plants.

A major limitation has been the low transfection rate of plant model systems that cannot scale for large libraries. However, recent development in optimizing a massively parallel reporter assay in tobacco leaves, called plant STARR-seq [141, 142], opened the door to testing thousands of 3' UTR sequences in parallel. By efficient *Agrobacterium*-mediated transformation, plant STARR-seq libraries are transiently expressed in tobacco leaves. The reporter consists of a barcoded Green Fluorescent Protein (GFP) reporter under the control of the Cauliflower mosaic virus 35S minimal promoter and a 35S core enhancer. Originally designed to test enhancer and promoter sequences, plant STARR-seq can be easily modified to test sequences in the 3' UTR.

Another major limitation for conducting MPRAs in plants has been the transfection dependency of protoplasts [255, 310]. Howeer, with a recent development in the scalable transfection of maize mesophyll protoplasts [310], one can now test thousands of sequences in maize with the same reporter construct used in tobacco STARR-seq [142]. Together, these technologies facilitated the largest plant promoter MPRA to date, revealing species-specific rules of transcription initiation regulation. The data were used to train a CNN model to predict species-specific promoter strength, aiding the design of strong synthetic plant promoters for crop engineering efforts [142].

The goal of MPRAs in plants is two-fold: first, to learn about the biology of the tested sequences and second, to use what is learned for engineering sequences with desired outcomes. Scientists are currently working to effectively generate complex traits in plants, including plant metabolic pathways, synthetic switches, and regulatory circuits. The design and characterization of the promoters or 3' UTRs are key stages in the design/build/test/learn cycle. However, the complex genetic architecture and long life cycle of plants necessitate the use of iterative rounds of testing and modification, which are laborious and inefficient. Thus, these highly parallel assays are particularly important for optimization at scale.

1.5 Hypothesis and scope of disseration

Inspired by the results of plant STARR-seq in tobacco leaves and maize protoplasts, I sought to test how the 3' end of each *Arabidopsis* and maize gene governs cleavage and polyadenylation position and efficiency. In the plant world, the region downstream of the coding sequence that contains the information necessary for mRNA maturation or cleavage and polyadenylation is called the terminator. The 3' UTR includes part of the terminator sequence, but the terminator can span past the annotated 3' UTR end to govern cleavage. The most commonly used terminators in plant transgenics are derived from viral or bacterial 3' UTRs. Yet when paired with strong constitutive or viral promoters, these non-plant terminators were shown to be improperly cleaved or not polyadenylated [70, 194]. In the same study, Diamos and Mason et al. [70] compared the expression level of a GFP reporter using native terminator sequences from various plants and found that in most cases, plant terminators led to higher GFP protein production than the 35S or NOS (*Agrobacterium tumefaciens* nopaline synthase) terminators. Still, no one had yet functionally characterized all terminators from *Arabidopsis* and maize, the most studied plant models.

Characterizing plant terminators for strength and efficiency will resolve the lack of available terminators for crop engineering. A major goal of plant transgenics is to highly express a gene or many genes of interest. Early plant transgenic constructs have already shown that pairing strong terminators with strong promoters helps deliver optimal transgene expression in tobacco [70, 131, 257], potato [9], tomato [115], and *Arabidopsis* and rice [220]. A fundamental goal of next generation crop engineering is to express multiple transgenes (for multiple desired traits) for crop improvement [324]. Since each transgene usually requires a unique promoter and terminator pair for expression to combat gene silencing and unwanted plasmid recombination, multiple unique elements of each are required to express different trangenes at varying levels within one molecular stack [251, 324]. The lack of diverse terminators tested in plants leads to repeated use of the same promoter and terminator combinations, a practice that often leads to unwanted gene silencing [239].

Increasing the repertoire of native plant terminators will deter transgene silencing, a phenomenon frequently observed in transgenic constructs driven by a strong promoter

[24, 194, 240]. Transcripts lacking terminators are major sources for small RNAs (sRNA) involved in gene silencing, while transcripts with double terminators led to significant decreases in sRNA production and improved expression [194, 229].

Plant terminators determine which poly(A) site the gene's mRNA uses. The precision at which transcripts are cleaved and polyadenylated is also a major determinant of terminator strength [62, 63]. Early mutagenesis screens designed to investigate transgene susceptibility to silencing identified genes that were directly involved with polyadenylation and transcription termination (The step at which RNA polymerase II releases the pre-mRNA) [110]. Transgenes with the NOS terminator showed diffuse patterns of polyadenylation whereas the same transgene with the heat shock protein HSP terminator showed polyadenylation only at one site [78]. As expected, the transgene with the HSP terminator produced greater gene expression and protein production than the transgene with NOS.

Finally, early work studying polyadenylation in plants found species-specific differences, mainly between the two clades of angiosperms known as monocotyledons (monocots) and dicotyledons (dicots). Monocots and dicots differ in many respects when it comes to *cis*-regulatory mechanisms of transcription. Early genetic studies found that the dicot pea rbcS mRNA was efficiently polyadenylated in transgenic dicot petunia [221], whereas the monocot wheat rbcS mRNA was improperly polyadenylated in transgenic dicot tobacco plants [149]. The expression of the same disease-resistance gene in three monocot species resulted in truncated messengers by differential polyadenylation, implying that there may be even more differences in poly(A) signal recognition within each clade [18]. Thus, it would be necessary for crop engineering to characterize terminators from different plant species representing dicots (e.g. *Arabidopsis*, tobacco) and monocots (e.g. maize, rice, wheat) for a complete picture of the plant polyadenylation.

In summary, I proposed a massively parallel reporter assay designed to evaluate the strength of native plant terminator sequences originating from two distinct plant species: dicot *Arabidopsis* and monocot maize. These sequences would be assessed in the context of two recipient plant species, dicot tobacco and monocot maize. Since strong terminators have been shown to increase gene expression, an RNA/DNA enrichment-based read-out would be the ideal measurement of terminator strength. I also sought to conduct 3' end

RNA-sequencing on the reporter constructs to assess how cleavage patterns correlate to terminator strength. Together, these data would be used to train a deep learning model that can accurately predict terminator strength given any DNA sequence, unlocking a new tool in the plant synthetic biology chest.

Chapter 2

FEATURES THAT GOVERN TERMINATOR STRENGTH IN PLANTS

Sayeh Gorjifard[1], Tobias Jores[1], Jackson Tonnies[1,2], Nicholas A Mueth[1], Kerry Bubb[1], Travis Wrightsman[3], Edward S Buckler[3,4], Stanley Fields[1,5], Josh T Cuperus[1,], Christine Queitsch[1,6]

Abstract

The 3' end of a gene, often called a terminator, modulates mRNA stability, localization, translation, and polyadenylation. Here, we adapted Plant STARR-seq, a massively parallel reporter assay, to measure the activity of over 50,000 terminators from the plants *Arabidopsis thaliana* and *Zea mays*. We characterize thousands of plant terminators, including many that outperform bacterial terminators commonly used in plants. Terminator activity is species-specific, differing in tobacco leaf and maize protoplast assays. While recapitulating known biology, our results reveal the relative contributions of polyadenylation motifs to terminator strength. We built a computational model to predict terminator strength and used it to conduct *in silico* evolution that generated optimized synthetic terminators. Additionally, we discover alternative polyadenylation sites across tens of thousands of terminators; however, the strongest terminators tend to have a dominant cleavage site. Our results establish features of plant terminator function and identify strong naturally occurring and synthetic terminators.

---

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195.

[2]Graduate Program in Biology, University of Washington, Seattle, WA 98195.

[3]Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY 14853.

[4]Institute for Genomic Diversity, Cornell University, Ithaca, NY 14853.

[5]Department of Medicine, University of Washington, Seattle, WA 98195.

[6]Corresponding Author: Christine Queitsch, *queitsch@uw.edu*

2.1 Introduction

A critical challenge is producing enough food for a growing world population that is likely to reach over 9 billion people by the year 2050 [67]. Food security can be bolstered through crop engineering, which can improve yields and increase tolerance to pathogens and environmental stresses [77, 111]. However, for crop engineering to be successful, precise control over transgene expression is required. In plants, past efforts to optimize critical cis-elements have mostly focused on the identification, characterization, and manipulation of upstream regulatory elements such as promoters [141, 258, 327, 373]. Although these upstream elements can produce large effects on gene expression, other sequence elements also contribute. For example, the 3' end of a gene contains sequence motifs necessary for mRNA 3' end maturation, cleavage, and polyadenylation [166, 124, 127, 62, 186, 305, 263, 262]. In keeping with established nomenclature [34, 62, 307], throughout this manuscript, we refer to sequences surrounding a cleavage and polyadenylation site as terminators. Terminators also affect mRNA stability, nuclear export, and translation [62]. Moreover, terminators play a central role in transgene silencing mediated by small RNAs [7, 19, 70, 78, 250].

Messenger RNA cleavage and polyadenylation are tightly linked to and initiate transcription termination [54, 30, 247, 335]; however, termination occurs up to 1 kb downstream of the mRNA cleavage site [213]. The current model of how mRNA 3' end processing leads to transcription termination combines features of two previously proposed models [261, 193, 73, 162]: After mRNA cleavage, conformational changes of RNA polymerase II allosterically reduce its elongation rate (allosteric model). At the same time, the cleavage site enables a 5' to 3' exonuclease (Rat1 in yeast, XRN2 in humans, and XRN3 in *Arabidopsis*) to degrade the downstream RNA and initiate transcription termination when it catches up with the RNA polymerase (torpedo model). The exact site of transcription termination can further be affected by RNA polymerase II pause sites, R loops formed between DNA and the nascent RNA, and DNA-binding proteins [73, 360].

Although the choice of terminator significantly impacts transgene expression, only a few terminators — often those derived from viruses or bacteria — are commonly used in crop engineering. These viral or bacterial sequences make transgene silencing more likely

[58, 70]. Compared to the commonly used *Agrobacterium tumefaciens* nopaline synthase terminator (tNOS), the terminator from the *Arabidopsis thaliana* heat shock protein 18.2 results in increased transgene expression and mRNA stability, and decreased silencing [63, 78, 114, 220, 240]. This finding highlights the need to characterize and optimize native plant terminators as building blocks for crop engineering applications [70, 131, 308].

Plant terminators have been characterized by analysis of wild-type and mutated versions of a handful of terminators, from both plant and non-plant sources, or by searching for motifs enriched in genomic sequences surrounding mRNA cleavage sites. These studies established three main cis-acting elements required for 3' end processing [28, 78, 62, 124, 186, 280]. The first element is the far upstream element (FUE), which resides in a U-rich region located 30 – 150 nucleotides upstream of the cleavage site and often contains one or more UGUA sequences. The second is the near upstream element (NUE), which resides in an A-rich region located 10 – 30 nucleotides upstream of the cleavage site and contains the polyadenylation signal AAUAAA. The third is the cleavage element (CE), which contains the cleavage site, formed by a UA or CA dinucleotide embedded in a U-rich region. Although the general pattern and nucleotide preferences of cis-acting elements in plant terminators are known, their relative contributions to terminator strength and species specificity have not been determined.

Here, we characterized over 50,000 terminators from the model plant *Arabidopsis thaliana* and the crop plant maize (*Zea mays*). We identified sequence features contributing to terminator strength and quantified their effect on transcript levels. By measuring terminator strength in two assay systems—tobacco leaves and maize protoplasts—we detected similarities and differences in the terminator grammar of monocotyledonous and dicotyledonous plants. Leveraging our large dataset, we trained computational models that accurately predict terminator strength. We used these models to design robust synthetic plant terminators for future crop engineering efforts.

2.2 Results

2.2.1 Measuring the strength of plant terminators with Plant STARR-seq

To assess the strength of plant terminators at high throughput, we used Plant STARR-seq, a massively parallel reporter assay that measures the activity of cis-regulatory elements [141, 142]. The average 3' untranslated region (UTR) length in Arabidopsis and maize is 242 bp and 310 bp, respectively [293, 133]. However, due to technical limitations in DNA array-synthesis, we were limited to sequences of 170 nucleotides. Previous studies on plant terminators revealed that most elements required for efficient pre-mRNA 3' end processing reside within approximately 150 bp upstream of the cleavage and polyadenylation site [186, 124, 280, 28, 62, 262]. Furthermore, studies in yeast and animals revealed that sequence elements downstream of the cleavage site can also affect polyadenylation [305, 28, 62]. Although such elements have not been reported in plants, we included the 20 nucleotides downstream of the cleavage site in our candidate sequences to test if they have an effect of terminator activity. For these reasons, we defined a terminator as the 170 nucleotide sequence from position −150 to +20 relative to a cleavage and polyadenylation site (position 0) in this study. We used experimentally derived primary cleavage sites to select terminator sequences from wild-type *Arabidopsis* and maize [133, 304, 343]. For the 3,754 *Arabidopsis* genes that were not represented in these datasets, we used the end of the 3' untranslated region (UTR) annotation in TAIR10 as the cleavage site. Since alternative polyadenylation plays an important role in gene regulation by creating different mRNA isoforms [127], we sought to investigate the strength of both primary and secondary polyadenylation sites. Therefore, we included experimentally-derived, secondary cleavage sites supported by at least 30% of the total reads of a gene [133, 304, 343]. As expected, the 24,529 *Arabidopsis* terminators and the 30,092 maize terminators displayed the distinctive nucleotide composition preferences of terminator sequences [28] and predominantly resided in the annotated 3' UTR region **(Supplementary Figure 2.1a-d)**.

The library of plant terminators was array-synthesized and cloned downstream of the coding sequence of a barcoded green fluorescent protein (GFP) reporter gene driven by the cauliflower mosaic virus 35S promoter **(Figure 2.1a)**. The plasmid library was transiently

expressed in tobacco leaves or maize protoplasts. After 1 – 2 days of incubation, the reporter mRNA was extracted, and next-generation sequencing was used to count the reporter barcodes in the input DNA and the extracted RNA. Strong terminators yield higher transcript levels due to improved 3' end processing or transcript stability as compared to weak terminators [62, 28]. The number of transcripts (i.e. RNA reads) per DNA template (i.e. DNA reads) is therefore a direct measure of the strength of a terminator. We define terminator strength as the enrichment of barcodes in RNA over DNA normalized to the enrichment of a control construct containing the 35S terminator. Thus, terminator strength in our assay reflects both transcriptional activity and RNA stability, and tends to have low values due to the normalization to the highly active 35S terminator.

We performed two biological replicates in tobacco leaves and maize protoplasts. The results were highly correlated in both assay systems **(Figure 2.1 b, c)**. Similarly, retesting of over 400 terminators in a second, independent library showed that the replicates in this validation experiment were highly correlated with each other and with the results of the first, large-scale experiment **(Supplemental Figure 2.2)**. Therefore, we used the average terminator strength from both replicates of the large-scale experiment for all further analyses. Terminator strength spanned a wide range of activity, allowing us to disentangle the signals that contribute to this strength. In tobacco leaves, we observed more than 64-fold difference between strong and weak terminators. Similar to previous studies [141, 142], the dynamic range was lower in maize protoplasts, in which we detected a 16-fold difference between strong and weak terminators.

To ensure that we measured terminator strength, we included negative controls derived from coding regions in *Arabidopsis* and maize and from randomized sequences. The random sequences were generated such that their overall (global random) or per-position (positional random) nucleotide frequencies resembled that of an average *Arabidopsis* or maize terminator **(Supplemental Figure 2.1e-g)**. In both tobacco leaves and maize protoplasts, sequences derived from plant terminators outperformed these coding sequences and random controls **(Figure 2.1 d, e)**. Of all control sequences, the positional random sequences showed the greatest strength, likely because they most closely resemble actual terminators. Together, these findings demonstrate that our assay captures *bona fide* terminator strength.

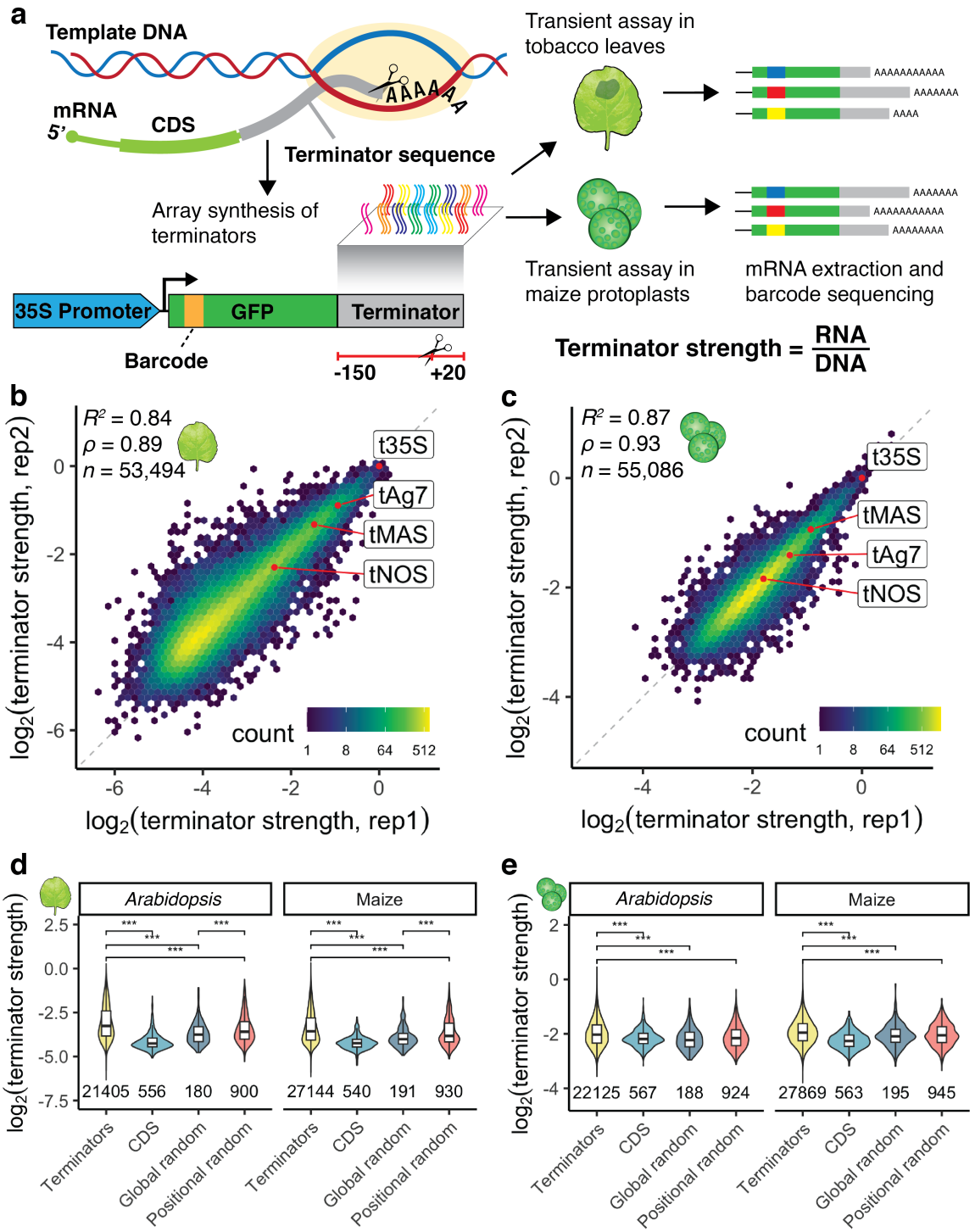

a
Template DNA
mRNA
5'
CDS
AAAAAA
Terminator sequence
Array synthesis of terminators
35S Promoter
GFP
Terminator
Barcode
-150
+20
Transient assay in tobacco leaves
Transient assay in maize protoplasts
mRNA extraction and barcode sequencing
Terminator strength = RNA / DNA
b
R² = 0.84
ρ = 0.89
n = 53,494
t35S
tAg7
tMAS
tNOS
count
1 8 64 512
log2(terminator strength, rep2)
log2(terminator strength, rep1)
c
R² = 0.87
ρ = 0.93
n = 55,086
t35S
tMAS
tAg7
tNOS
count
1 8 64 512
log2(terminator strength, rep2)
log2(terminator strength, rep1)
d
Arabidopsis
Maize
log2(terminator strength)
21405 556 180 900
27144 540 191 930
Terminators
CDS
Global random
Positional random
e
Arabidopsis
Maize
log2(terminator strength)
22125 567 188 924
27869 563 195 945
Terminators
CDS
Global random
Positional random

Figure 2.1: **Plant STARR-seq measures terminator strength in tobacco leaves and maize protoplasts.** a Terminator sequences (bases −150 to +20 relative to the cleavage and polyadenylation site) were array-synthesized and cloned downstream of a barcoded GFP reporter gene driven by the 35S promoter. After transient expression of the plasmid library in tobacco leaves or maize protoplasts, mRNA was extracted for barcode sequencing. We define terminator strength as the enrichment of barcodes in the extracted mRNA over the input DNA normalized to the strength of the 35S terminator. b, c Hexbin plots (color represents the count of points in each hexagon) of the correlation between two biological replicates of Plant STARR-seq in tobacco leaves (b) or maize protoplasts (c). Commonly used terminators are highlighted in red. Pearson's R^2, spearman's ρ, and number (n) of terminators are indicated. d, e Violin plots, box plots, and significance levels of terminator strength in tobacco leaves (d) or maize protoplasts (e) for plant terminators (Terminators) compared to sequences from coding regions (CDS) and randomized sequences with an overall (Global random) or per-position (Positional random) nucleotide frequency similar to an average Arabidopsis or maize terminator. Violin plots represent the kernel density distribution and the box plots inside represent the median (center line), upper and lower quartiles and 1.5x the interquartile range (whiskers) for all corresponding terminators. Numbers at the bottom of each violin indicate the number of terminators in each group. Significant differences between two samples were determined by the two-sided Wilcoxon rank-sum test and are indicated: *p $\leq$ 0.05, **p $\leq$ 0.01, ***p $\leq$ 0.001, NS, not significant.

We sought to confirm our measure of terminator strength by Plant STARR-seq with an orthogonal assay. To do so, we conducted a dual-luciferase assay in both tobacco leaves and maize protoplasts for several weak, intermediate, and strong terminators. We observed a strong correlation between terminator strength measured by the two assays **(Figure 2.3 b, d)**. Strong terminators yielded higher nanoluciferase activity than weak terminators **(Figure 2.3 a, c)**.

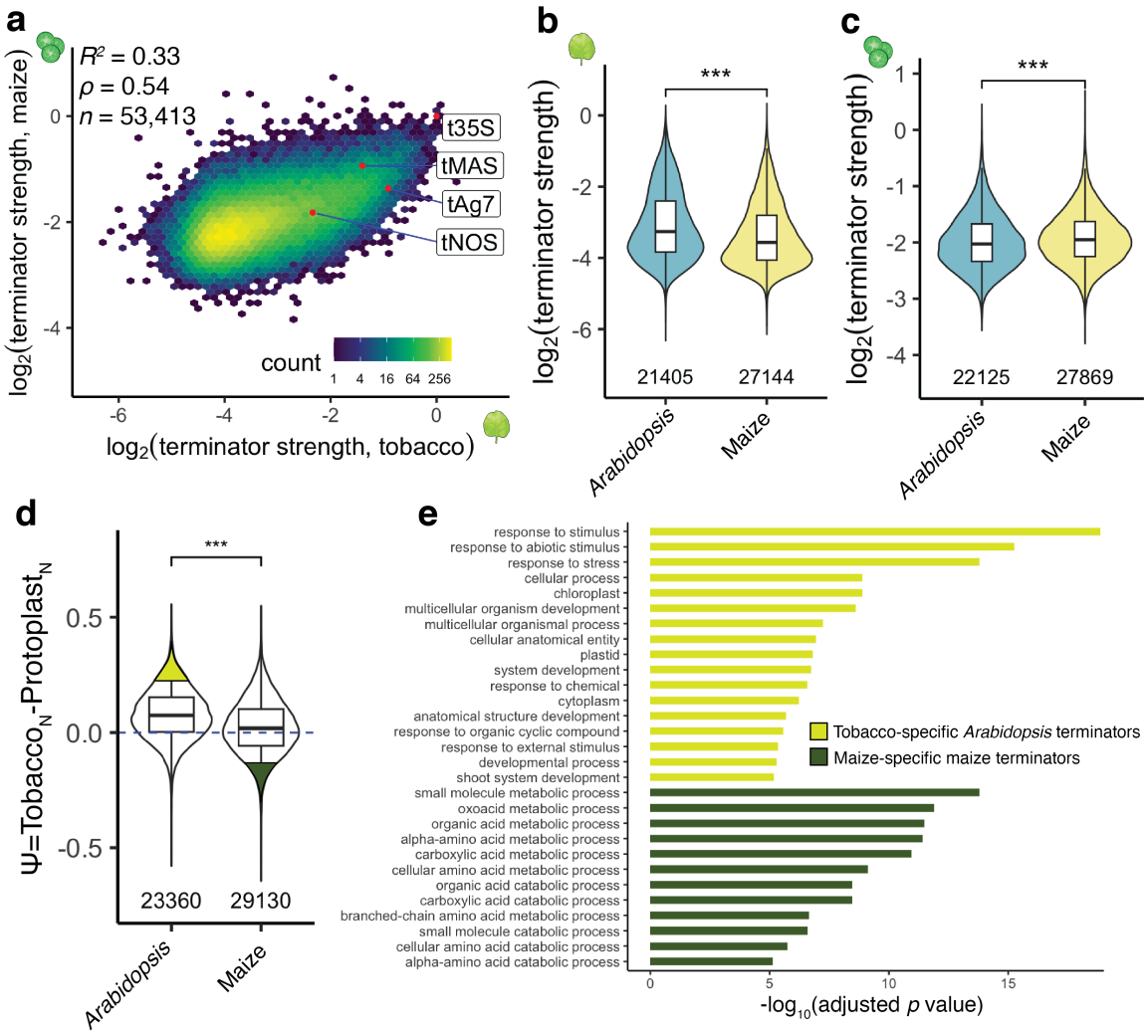
Most plant transgenes use viral or bacterial terminators, such as the 35S, Ag7, NOS, and MAS terminators [10]. We included these terminators in our library in their full form (not limited to 170 nucleotides) as a reference for the activity of the plant terminators. The 35S terminator, the only viral sequence in our library, outperformed almost all plant terminators. However, some plant terminators (8 in tobacco leaves and 20 in maize protoplasts) exceeded the activity of the 35S terminator. The bacterial Ag7, NOS, and MAS terminators were far weaker than the viral 35S terminator. In tobacco leaves, we found 2,224 plant terminators that were stronger than the MAS terminator and 9,389 plant terminators that were stronger

than the NOS terminator. In maize protoplasts, we found 1,369 plant terminators that were stronger than the MAS terminator and 19,585 terminators that were stronger than the NOS terminator. The strongest plant terminators are attractive alternatives to the bacterial terminators commonly used in plant transgenes.

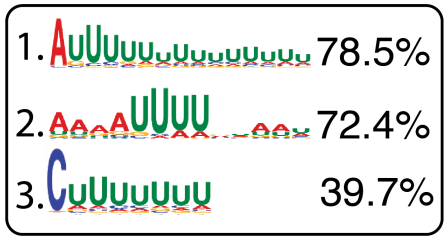
2.2.2 Terminator strength correlates not with gene expression but with gene function

Next, we asked if our data could give insights into the importance of terminator strength for gene expression. We compared terminator strength determined by Plant STARR-seq with metrics related to mRNA levels, stability and degradation. We found only weak correlation between terminator strength and gene expression [295], mRNA half-life [300], or nascent transcription [184] **(Supplemental Figure 2.4)**. These findings corroborate reports that upstream cis-regulatory elements, like enhancers, promoters, and the 5' UTR, drive large-scale gene expression changes while the 3' UTR and terminators fine-tune expression levels [329, 374]. However, the choice of terminator can have a significant impact on the expression of transgenes [78, 324].

While the correlation between terminator strength and gene expression was low, we wondered whether the function of a gene and the strength of its terminator were more correlated. To address this question, we performed gene ontology (GO) term enrichment analysis on the genes associated with the strongest 10% of terminators from *Arabidopsis* or maize. For both species, we found a significant enrichment (adjusted p value < 0.05) for genes related to metabolism and response to stimulus **(Supplemental Figure 2.5a-d)**. Terminators derived from oxidoreductase-related and stress-responsive genes in *Arabidopsis* were overrepresented in the top 10% of terminators ranked by strength in tobacco leaves and maize protoplasts. Maize terminators derived from genes involved in small molecule metabolic processes were overrepresented in the strongest 10% of terminators in both systems. However, we found many more significant GO terms associated with metabolism for the strongest maize terminators in maize protoplasts than in tobacco leaves **(Supplemental Figure 2.5e, f)**. The latter finding pointed to possible species-specific differences in terminator strength, prompting us to investigate this possibility further.



f Tobacco-specific *Arabidopsis* terminators



g Maize-specific maize terminators

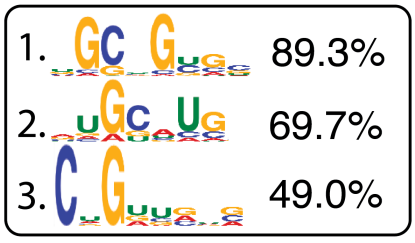


Figure 2.2: **Plant terminator strength is species-specific.** a Hexbin plot comparing terminator strength in tobacco leaves and maize protoplasts. b, c Violin plots of the strength of terminators derived from the indicated species in tobacco leaves (b) or maize protoplasts (c). d Violin plot of Ψ, the difference between normalized (0, weakest; 1, strongest) $\log_2$(terminator strength) in tobacco leaves (Tobacco_N) and maize protoplasts (Protoplasts_N). Terminators are grouped by their species of origin. The top 10% of Arabidopsis terminators with highest Ψ values (tobacco-specific terminators) and the top 10% of maize terminator with the lowest Ψ values (maize-specific terminators) are highlighted in yellow and green, respectively. e GO terms enriched in genes associated with the tobacco-specific Arabidopsis terminators or the maize-specific maize terminators highlighted in d. Only GO terms with adjusted p value ≤ 0.0001 are shown. The p values were determined using the gprofiler2 library in R with gSCS correction for multiple testing. All enriched GO terms and exact p values are listed in **Supplementary Table 6**. f, g Top three motifs enriched in tobacco-specific Arabidopsis terminators relative to maize-specific maize terminators (f) and vice versa (g). The hexbin plot in a and the violin plots, box plots, and significance levels in b-d are as defined in Figure 2.1

2.2.3 Plant terminator strength is species-specific

mRNA 3' end processing may differ between monocotyledonous and dicotyledonous plants, as suggested from studies in the 1980s. For example, the gene encoding the small subunit of ribulose 1,5-bisphosphate carboxylase (rbcS) from the monocot wheat is improperly polyadenylated when tested in the dicot tobacco [149]. In contrast, the mRNA of the rbcS gene from the dicot pea is efficiently polyadenylated in tobacco [221]. Consistent with these earlier findings, we observed a relatively weak correlation ($R^2 = 0.33$) between the strength of a given terminator in the dicot model tobacco leaves versus the monocot model maize protoplasts **(Figure 2.2a)**. In tobacco leaves, terminators from the dicot *Arabidopsis* performed significantly better than terminators from the monocot maize. Similarly, in maize protoplasts, maize terminator sequences significantly outperformed *Arabidopsis* terminators **(Figure 2.2b, c)**. These observations are consistent with species-specific differences in plant terminator strength.

Based on these observations, we wanted to identify terminators with strong species-specific activity. Since our assay systems show different dynamic ranges, we normalized terminator strength within each assay system to a scale from 0 (weakest) to 1 (strongest).

We defined a variable Ψ as the difference between the normalized terminator strength in tobacco leaves and in maize protoplasts, such that a high Ψ value denotes a terminator that was considerably stronger in tobacco leaves than in maize protoplasts. We found that *Arabidopsis* terminators show a higher average Ψ value than maize terminators (Figure 2.2d). We selected the top 10% of *Arabidopsis* terminators (n = 1,923) with highest Ψ values, i.e. tobacco-specific strength, and the top 10% of maize terminators (n = 2,439) with the lowest Ψ values, i.e. maize-specific strength, for GO term enrichment analysis. *Arabidopsis* terminators with highly tobacco-specific strength tended to be derived from stimulus- and stress-responsive genes, while maize terminators with highly maize-specific strength tended to be derived from metabolic genes **(Figure 2.2e)**.

To understand how species-specific terminator strength is mediated, we used STREME [20] to search for RNA motifs that were enriched in the tobacco-specific Arabidopsis terminators (high Ψ) relative to the maize-specific maize terminators (low Ψ), and vice versa. The most enriched motifs in the tobacco-specific Arabidopsis terminators were dominated by long stretches of U nucleotides, while the most enriched motifs found in maize-specific maize terminators favored G and C nucleotides **(Figure 2.2f, g)**. These findings motivated us to further analyze the effect of nucleotide composition on terminator strength.

2.2.4 Nucleotide composition affects terminator strength in a position-specific manner

Dicot genomes have a lower GC content than monocot genomes [290]. This bias also holds true for *Arabidopsis* and maize terminator sequences **(Figure 2.3a)**. Assessing the impact of GC content on terminator strength, we found that terminators with either high or low GC content were weaker than those with intermediate GC content in both tobacco leaves and maize protoplasts **(Figure 2.3b, c)**. There are, however, subtle differences between the two assay systems. In tobacco leaves, terminators with a GC content around 30-35% were the strongest, while in maize protoplasts, terminators with a GC content of approximately 35-40% were the strongest. To substantiate these subtle differences, we measured the terminator strength of random sequences with a GC content of 30%, 40%, 50%, 60%, or 70%. In tobacco leaves, the random sequences with a GC content of 30% performed best,

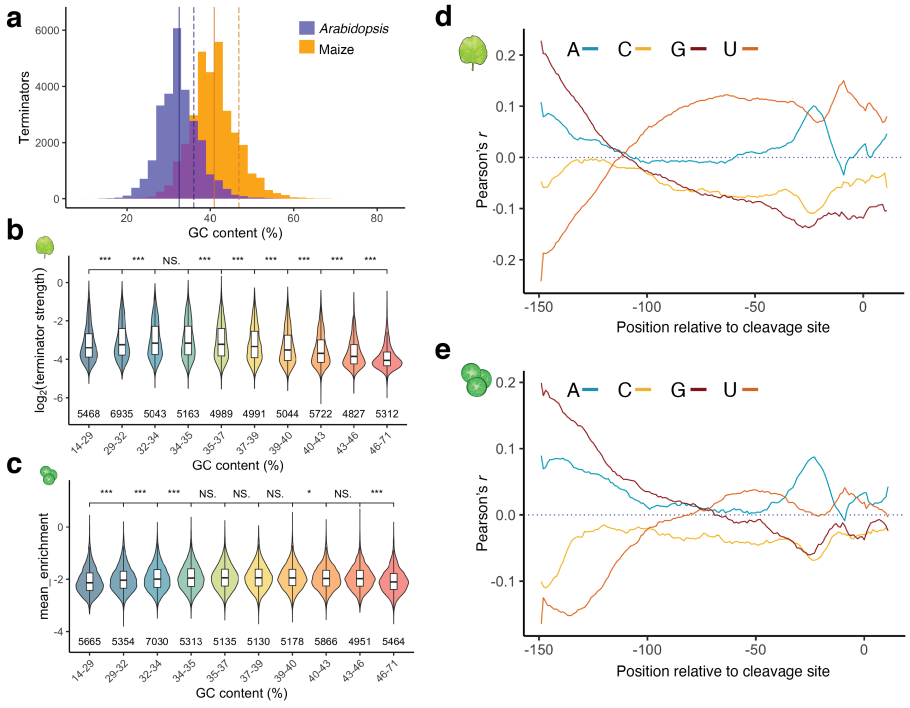


Figure 2.3: **Nucleotide composition affects terminator strength in a species- and position-specific manner.** a Histogram of terminator GC content. The solid line indicates the mean GC content of terminators (*Arabidopsis* = 32.53%, maize = 40.89%) and the dashed line indicates the GC content of the genome (*Arabidopsis* = 36.06%, maize = 46.86%). b, c Violin plots, box plots, and significance levels (as defined in Figure 2.1) of terminator strength in tobacco leaves (b) or maize protoplasts (c). Terminators were binned by GC content to yield groups of approximately the same size. d, e Correlation (Pearson's R) between terminator strength in tobacco leaves (d) or maize protoplasts (e) and the A, C, G, or U content of a ten-base window starting at the indicated position in the plant terminators.

while in maize protoplasts, the random sequences with a GC content of 40% showed the highest strength **(Supplemental Figure 2.6)**. These values coincide with the average GC content of *Arabidopsis* terminators (32.5%; the average GC content of tobacco terminators is likely in a similar range) and maize terminators (40.9%). Thus, we conclude that the cleavage and polyadenylation machinery is likely attuned to the species-specific GC content of plant terminators.

Since terminator strength is governed at the RNA level, G and C or A and U nucleotides are not necessarily represented equally. Therefore, we teased apart the positional effect on terminator strength for each of the four nucleotides **(Figure 2.3d, e)**. We found that C nucleotides have little to no effect on terminator strength in tobacco leaves or maize protoplasts. High G content was beneficial at the 5' end of our tested sequences but not tolerated near the cleavage site, especially in tobacco leaves. Conversely, high U content near the 5' end was associated with lower terminator strength. The system preferences diverged, however, for sequences with high U content starting from 100 nucleotides upstream from the cleavage site: U-rich sequences in this region were associated with increased terminator strength in tobacco leaves but not in maize protoplasts. High A content around 20 nucleotides upstream of the cleavage site was correlated with terminator strength in both systems, probably due to the canonical polyadenylation signal — a short sequence often represented by the hexamer AAUAAA — typically found at this location.

2.2.5 Canonical cleavage and polyadenylation motifs contribute to terminator strength

Since the strength of a terminator is only partially explained by its nucleotide composition, we searched for sequence motifs that contribute to terminator strength. To this end, we conducted a motif enrichment analysis on the strongest 25% of terminators in both assay systems. The most significantly enriched motifs match canonical cleavage and polyadenylation signals: the UGUA sequence motif from the Far Upstream Element (FUE) and the AAUAAA polyadenylation signal **(Figure 2.4a)**. While these motifs are known to affect cleavage and polyadenylation, our assay allowed us to determine the quantitative contribution of each motif to terminator strength.

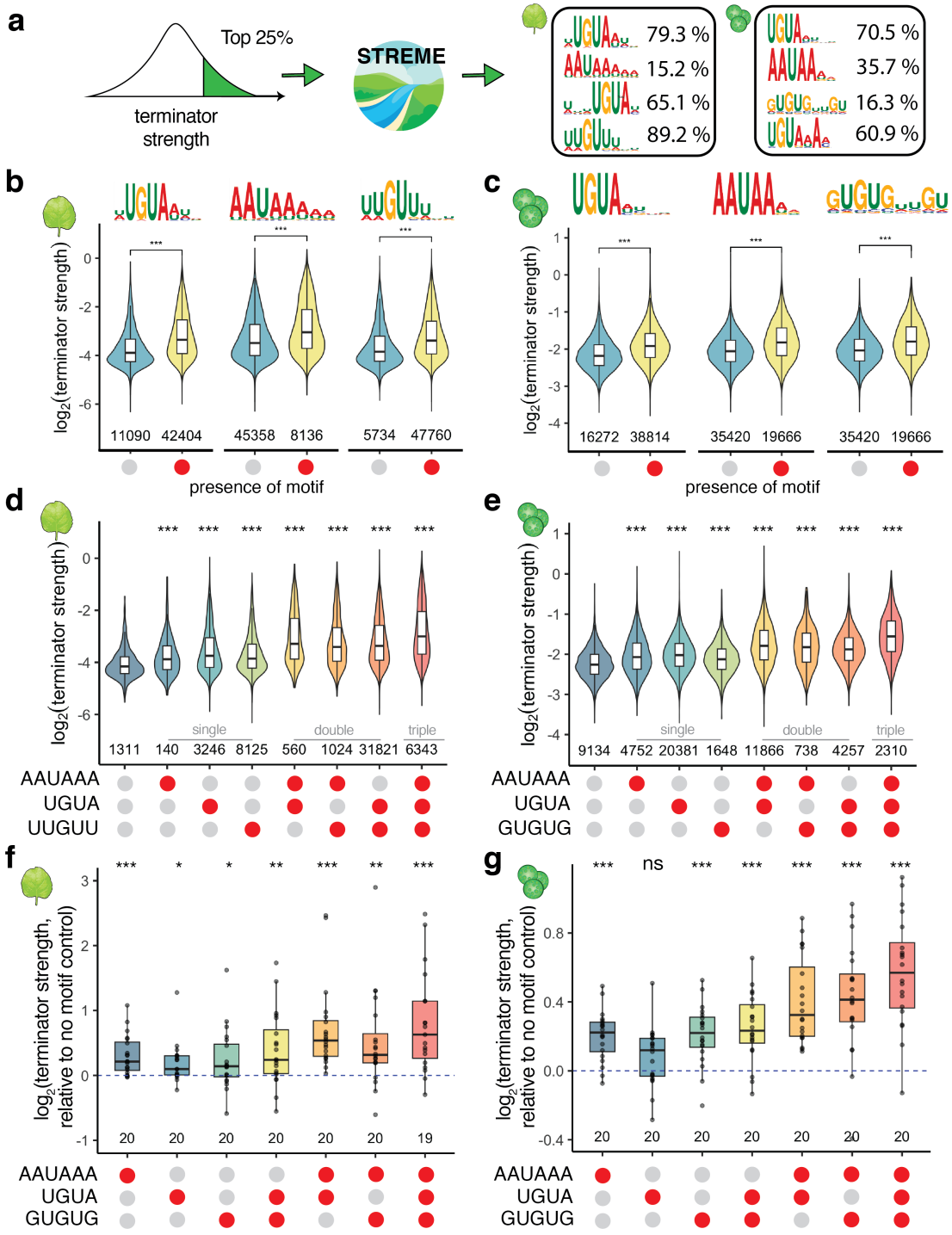

a
Top 25%
terminator strength
STREME
79.3 %
15.2 %
65.1 %
89.2 %
70.5 %
35.7 %
16.3 %
60.9 %
b
c
log2(terminator strength)
presence of motif
11090 42404 45358 8136 5734 47760
16272 38814 35420 19666 35420 19666
d
e
single
double
triple
1311 140 3246 8125 560 1024 31821 6343
9134 4752 20381 1648 11866 738 4257 2310
AAUAAA
UGUA
UUGUU
GUGUG
f
g
log2(terminator strength, relative to no motif control)
ns
20 20 20 20 20 20 19
20 20 20 20 20 20 20

Figure 2.4: **Cleavage and polyadenylation motifs increase terminator strength.** a Sequence logo plots of the top four motifs enriched in the strongest terminators in tobacco leaves (left) or maize protoplasts (right). For each motif, the percentage of *Arabidopsis* and maize terminators harboring it is indicated. b-e Violin plots of terminator strength in tobacco leaves (b, d) or maize protoplasts (c, e) with or without AAUAAA, UGUA, or U/G-rich motifs individually (b, c) or in combination (d, e). f, g Box plots (centre line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range),significance levels (compared to no motif), and underlying data points for the strength of terminators supplemented with the indicated motifs relative to the corresponding terminator without any motif (set to 0). In b-g, red circles indicate presence of a motif and gray circle indicate its absence. Violin plots, box plots, and significance levels in b-e are as defined in Figure 2.1.

The UGUA motif is highly prevalent in our library, with 60% to 80% (small differences between the UGUA-containing motifs identified by STREME lead to different numbers of positive terminators) of all terminators containing this motif. The UGUA motif is predominantly located 30 to 40 nucleotides upstream of the cleavage site in both Arabidopsis and maize terminators **(Supplemental Figure 2.7a, b)**. In tobacco leaves and maize protoplasts respectively, terminators with a UGUA motif were on average 50% and 20% stronger than those without this motif **(Figure 2.4b, c)**. Consistent with prior reports [263], the number of UGUA motifs also correlates with terminator strength and cleavage probability **(Supplemental Figure 2.8a, b)**.

The polyadenylation signal is less prevalent in our library (15% to 35% of all terminators) but shows a distinct localization profile, with a sharp peak about 20 nucleotides upstream of the cleavage site **(Supplemental Figure 2.7c, d)**. The polyadenylation signal contributes significantly to terminator strength. In tobacco leaves and maize protoplasts respectively, terminators with a polyadenylation motif were 40% and 20% stronger than those without this motif **(Figure 2.4b, c)**. Since fewer than half of the terminators in our library contain a canonical polyadenylation signal, we asked if similar sequences are functionally equivalent. Terminators with a perfect AAUAAA sequence were stronger than terminators with a sequence that differs by a single nucleotide (carrying a single-nucleotide variant of this motif), although AAUAAG and AAUAAU motifs performed nearly as well. Terminators

without any AAUAAA-like sequence were considerably weaker **(Supplemental Figure 2.8a, b)**.

In addition to the AAUAAA and UGUA motifs [124, 62], we found novel U- and G-rich motifs enriched in the strongest terminators **(Figure 2.4a)**. These U/G-rich motifs are broadly distributed upstream of the polyadenylation signal and at slightly higher frequency just upstream or downstream of the cleavage site **(Supplemental Figure 2.7e, f)**. While the U/G-rich motifs share a similar localization pattern, there are striking differences between the motifs discovered in the tobacco leaf and the maize protoplast system. The motif detected in the tobacco leaf data is found in nearly 90% of all terminators and consists of a G nucleotide surrounded by 2 to 3 U nucleotides. In contrast, the motif identified in the maize protoplast data is much less prevalent (16% of all terminators) and consists of alternating U and G nucleotides **(Figure 2.4a)**. Despite these differences, both motifs have an influence on terminator strength, leading to a 40% and 20% increase in strength for terminators with the motif as compared to those without, in tobacco and maize respectively **(Figure 2.4b, c)**.

While each individual motif contributes positively to terminator strength, we observed additive effects when multiple motifs are present within the same terminator. Independent of the assay system, the strongest terminators contain all three motifs and were on average 60% more active than terminators without any motifs **(Figure 2.4d, e)**.

To validate the effect of the identified AAUAAA and UGUA motifs on terminator strength, we selected 20 terminators each with a single strong AAUAAA or UGUA motif and generated single or double nucleotide mutations to break the motif. The terminator strength of the original and mutated sequences was then measured in tobacco leaves and maize protoplasts. Most motif mutations lowered terminator strength **(Supplemental Figure 2.8c, d)**. Mutations in the AAUAAA motif caused, on average, a 24% decrease in terminator strength in tobacco leaves and a 16% decrease in maize protoplasts. Mutations in the UGUA motif decreased terminator strength by approximately 13% and 7% in tobacco leaves and maize protoplasts, respectively.

Next, we wondered if we could use the enriched motifs to improve the strength of weak terminators. We selected randomized sequences from our initial data that showed low

terminator strength and did not contain an AAUAAA, a UGUA, or a U/G-rich motif. We then inserted a CA dinucleotide at the predicted cleavage site (position 150) and added all possible single, double, and triple combinations of the AAUAAA, UGUA, and U/G-rich motifs. All sequences were then subjected to Plant STARR-seq in tobacco leaves and maize protoplasts. In most cases, adding motifs increased terminator strength and additive effects were observed when multiple motifs were inserted into a terminator **(Figure 2.4f, g)**. On average, single motifs led to a 5% to 20% increase in terminator strength. Terminators with all three motifs were approximately 60% stronger than motif-less terminators in tobacco leaves and maize protoplasts. Although the GUGUG motif was originally identified in data from the maize protoplast system, it also increased terminator activity in tobacco leaves. Taken together, these findings demonstrate that the motifs enriched in strong terminators are indeed contributing to terminator strength.

2.2.6 Computational models accurately predict terminator strength

Computational models can successfully predict the activity of regulatory elements by learning key sequence features [56, 100, 142]. To develop computational models that can predict terminator strength, we initially focused on using k-mer counts as a proxy for terminator activity. To test the validity of this approach, we counted the occurrence of all possible 6-mers (4,096 sequences) in the terminator sequences, and calculated the correlation between terminator strength and how often a given 6-mer was represented in a terminator sequence. As expected, we found that the 6-mers most correlated with terminator strength in tobacco leaves and maize protoplasts are variations of the canonical cleavage and polyadenylation signals AAUAAA and UGUA **(Figure 2.5a, b)**. However, we observed only a moderate conservation of the 6-mer rank orders between the two assay systems (Spearman's ρ = 0.29; **Figure 2.5c**). 6-mers that were highly correlated with terminator strength in tobacco leaves but not in maize protoplasts were U-rich. Conversely, 6-mers with a high correlation to terminator strength in maize protoplasts but not in tobacco leaves were rich in G and C nucleotides **(Figure 2.5c, d)**. These findings are consistent with our prior observations on sequence motifs associated with high terminator activity, both within and across the two

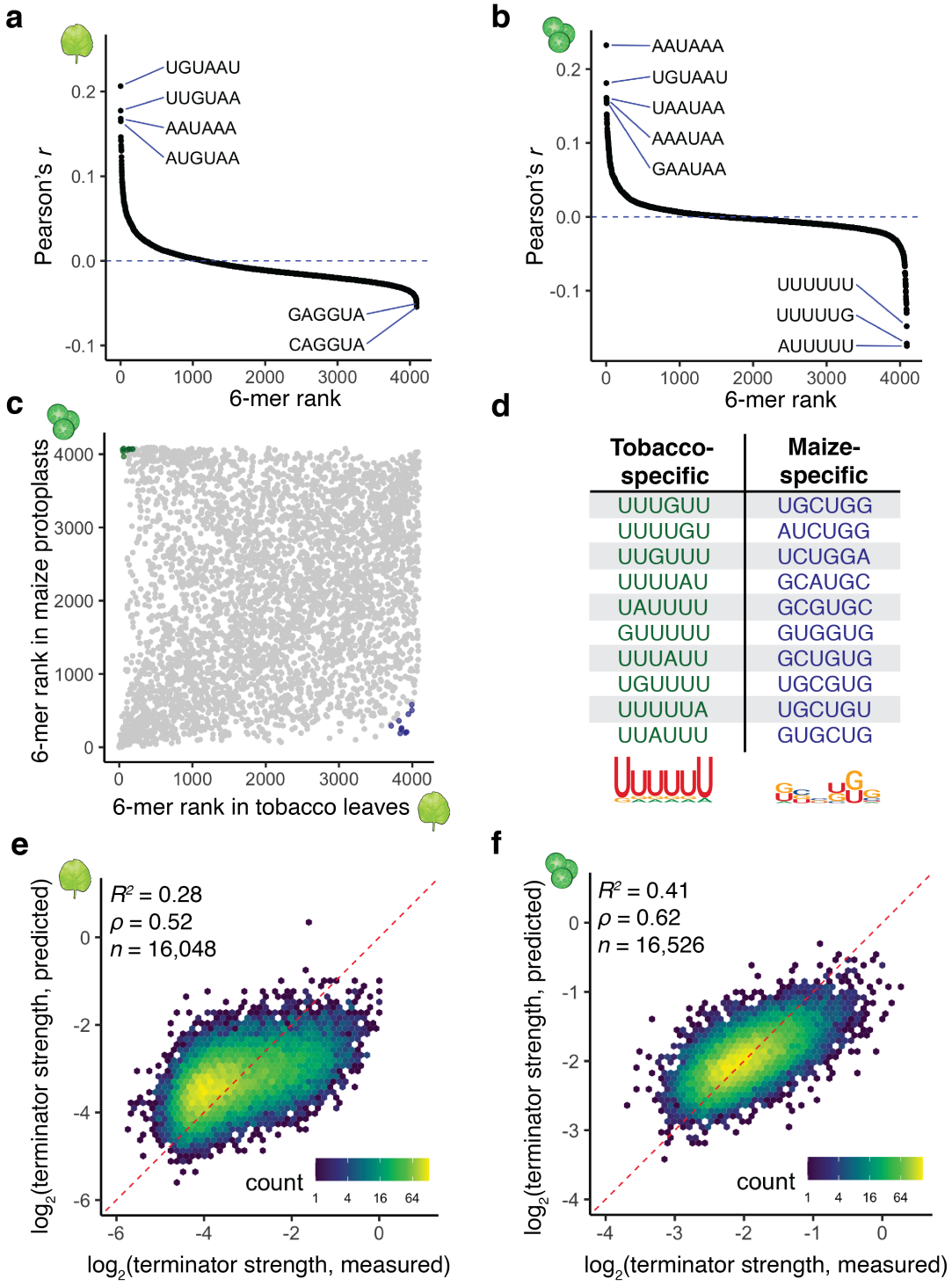

a
UGUAAU
UUGUAA
AAUAAA
AUGUAA
GAGGUA
CAGGUA
Pearson's r
6-mer rank
b
AAUAAA
UGUAAU
UAAUAA
AAAUAA
GAAUAA
UUUUUU
UUUUUG
AUUUUU
Pearson's r
6-mer rank
c
6-mer rank in maize protoplasts
6-mer rank in tobacco leaves
d
Tobacco-specific
Maize-specific
UUUGUU
UUUUGU
UUGUUU
UUUUAU
UAUUUU
GUUUUU
UUUAUU
UGUUUU
UUUUUA
UUAUUU
UGCUGG
AUCUGG
UCUGGA
GCAUGC
GCGUGC
GUGGUG
GCUGUG
UGCGUG
UGCUGU
GUGCUG
e
R² = 0.28
ρ = 0.52
n = 16,048
count
log2(terminator strength, predicted)
log2(terminator strength, measured)
f
R² = 0.41
ρ = 0.62
n = 16,526
count
log2(terminator strength, predicted)
log2(terminator strength, measured)

Figure 2.5: **K-mers can be used for species-specific terminator strength predictions.** a, b Correlation (Pearson's R) between terminator strength in tobacco leaves (a) or maize protoplasts (b) and how often a given 6-mer is present in the terminators. The horizontal axis displays the 6-mer "rank" based on Pearson's correlation. The sequences of the highest and lowest ranked 6-mers are indicated. c Comparison of the 6-mer ranks in tobacco leaves and maize protoplasts. 6-mers highlighted in green and blue had the biggest differences in rank order between the two assay systems. d Top 10 tobacco- or maize-specific 6-mers (highlighted in c) and sequence logo plots generated from them. e, f Hexbin plot (as defined in Figure 2.1) of the correlation between terminator strength measured by STARR-seq in tobacco leaves (e) or maize protoplasts (f) and terminator strength predictions from a lasso regression model based on 6-mer counts. Only the terminators not used for model training are shown (30% of all terminators).

assay systems. Therefore, we used 6-mer counts as input features to train a lasso regression model to predict terminator strength. The model was trained on 70% of our terminator data and tested on 30% of the data. The lasso regression model had moderate predictive power, explaining 28% of the variance in terminator strength in tobacco leaves and 41% in maize protoplasts **(Figure 2.5e, f)**.

To build a model with increased predictive power, we turned to a convolutional neural network with a DenseNet architecture [121], because this approach had worked well with Plant STARR-seq data previously [65]. Our DenseNet model uses the sequence of a terminator as an input and predicts the strength of this sequence in tobacco leaves and maize protoplasts. After the model was trained on 90% of the terminator data, it could accurately predict terminator strength for the remaining 10% of the data. The features learned by the model explained 76% and 67% of the variance in terminator strength in tobacco leaves and maize protoplasts, respectively **(Figure 2.6a, b)**. To understand what features the DenseNet model had learned, we used DeepLIFT and TF-MoDISco to extract consolidated sequence motifs that positively or negatively impact terminator strength [283, 284] prediction **(Figure 2.6c-f)**. According to this analysis, AAUAAA and UGUA motifs are associated with increased terminator strength in both assay systems. U-rich sequences decrease terminator strength in maize protoplasts, especially if they surround a UGUA motif. In contrast, U-rich sequences increase terminator strength in tobacco leaves, although this

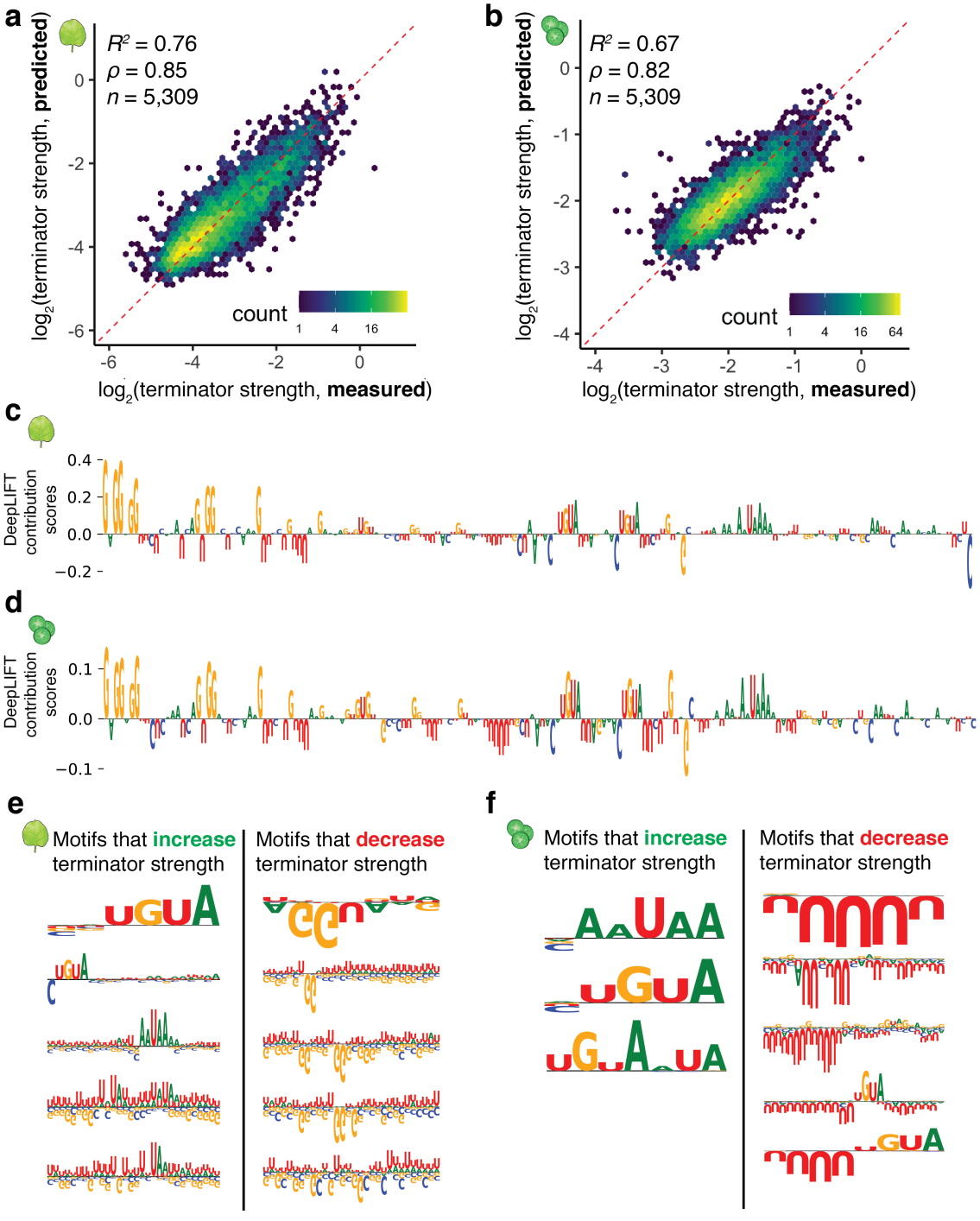

a
R^2 = 0.76
ρ = 0.85
n = 5,309
count
1 4 16
log2(terminator strength, predicted)
log2(terminator strength, measured)
b
R^2 = 0.67
ρ = 0.82
n = 5,309
count
1 4 16 64
log2(terminator strength, predicted)
log2(terminator strength, measured)
c
DeepLIFT contribution scores
d
DeepLIFT contribution scores
e
Motifs that increase terminator strength
Motifs that decrease terminator strength
f
Motifs that increase terminator strength
Motifs that decrease terminator strength

Figure 2.6: **A convolutional neural network accurately predicts terminator strength.** a, b Hexbin plot (as defined in Figure 2.1) of the correlation between terminator strength measured by STARR-seq in tobacco leaves (a) or maize protoplasts (b) and terminator strength predictions from a DenseNet convolutional neural network trained on terminator sequences. Only the terminators not used for model training are shown (10% of all terminators). c, d DeepLIFT importance scores based on our DenseNet model predictions of terminator strength in tobacco leaves (c) or maize protoplasts (d). The terminator of AT1G31180 is shown as an example. e, f Motifs identified by TF-MoDISco that positively or negatively contribute to the DenseNet model predictions of terminator strength in tobacco leaves (e) or maize protoplasts (f).

effect can be reversed by GG dinucleotides **(Figure 2.6e, f)**. These findings are consistent with the results from our motif enrichment and k-mer analyses and indicate that the DenseNet model has learned biologically relevant terminator features.

2.2.7 In silico evolution of terminator sequences

While our previous attempts to improve terminator strength by adding polyadenylation and cleavage motifs were successful, the gain in terminator strength was modest **(Figure 2.4f, g).** *In silico* evolution using convolutional neural network models can drastically improve the activity of regulatory elements [56, 142]. Therefore, we used our DenseNet model for *in silico* evolution of 222 terminators,111 each from *Arabidopsis* and maize. As starting sequences, we randomly selected terminators from the held-out test set used to validate the DenseNet model. For each terminator, we generated every possible single-nucleotide substitution variant and scored these variants with the DenseNet model. We kept the variant with the highest predicted terminator strength and subjected it to another round of evolution for a maximum of ten rounds **(Figure 2.7a)**. We performed this process three times using the prediction for either tobacco leaves or maize protoplasts, or the sum of both predictions to identify the strongest variant. To test the evolved elements experimentally, we synthesized the starting sequences and those obtained after three and ten rounds of evolution, and experimentally assayed their terminator strength in tobacco leaves and maize protoplasts.

After three rounds of evolution (i.e., after changing only three nucleotides), terminator

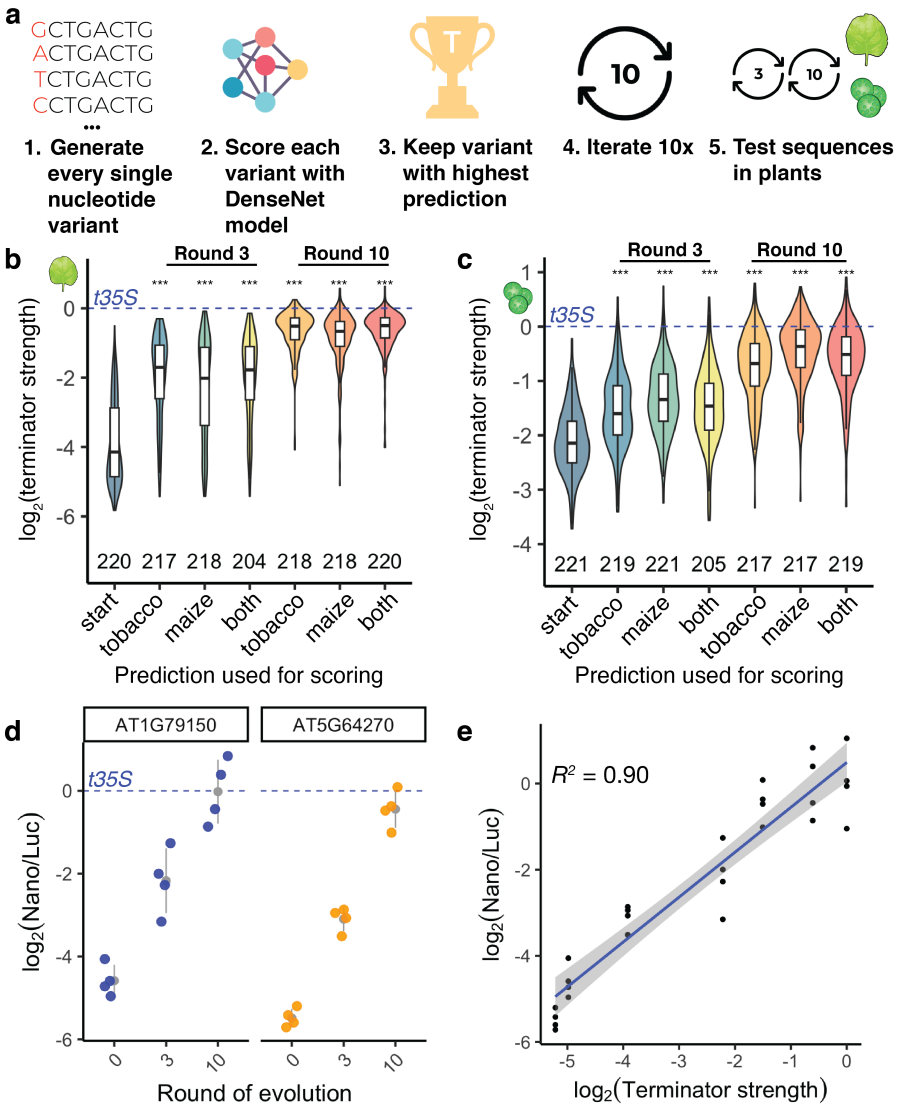


Figure 2.7: ***In silico* evolution of plant terminators.** a Scheme of the *in silico* evolution. b, c Violin plots, box plots, and significance levels (as defined in Figure 2.1) of terminator strength measured in tobacco leaves (b) or maize protoplasts (c) for unmodified terminators (start) and terminators after three or ten rounds of *in silico* evolution. The DenseNet model prediction for tobacco leaves (tobacco) or maize protoplasts (maize), or the sum of both predictions (both) was used as a score during *in silico* evolution. The dashed blue line indicates the strength of the 35S terminator (t35S). d Jitter plot of the nanoluciferase activity of the in silico evolution of AT1G79150 and AT5G64270. The dashed blue line indicates the average nanoluciferase activity of the 35S terminator (t35S), set to 0. The gray dot denotes the mean and the gray line denotes the variance. e Dot plot and Pearson's R^2 between terminator strength and nanoluciferase activity of evolved terminators in (d).

strength increased, on average, 5-fold in tobacco leaves and 1.6-fold in maize protoplasts **(Figure 2.7b, c)**. For sequences obtained after ten rounds of evolution, a further increase in terminator strength was observed. Even though the original sequences and those obtained after *in silico* evolution differed by only ten nucleotides, the difference in their terminator strength was 12-fold in tobacco leaves and 4-fold in maize protoplasts **(Figure 2.7b, c)**. As expected, the increase in terminator strength was most pronounced when the prediction used to score variants during *in silico* evolution matched the experimental assay system (i.e. using the tobacco leaf prediction for evolution, and testing the resulting terminators in tobacco leaves). Although all starting sequences were weaker than the 35S terminator, our *in silico* evolution approach generated several terminators of greater strength than this highly active viral terminator **(Figure 2.7b, c)**. We tested selected evolved terminators with the dual-luciferase assay and found strong correspondence with the Plant STARR-seq results **(Figure 2.7d, e)**. Furthermore, we observed that the model favored the insertion of multiple UGUA motifs into the evolved, stronger terminators **(Supplemental Figure 2.8g)**.

2.2.8 Cleavage and polyadenylation efficiency affects terminator strength

Over 70% of the genes in *Arabidopsis* and maize contain more than one possible polyadenylation site and can therefore give rise to transcript isoforms that differ in 3' UTR length [133, 343]. Alternative polyadenylation can have a strong effect on transcript levels through the inclusion or exclusion of regulatory features such as binding sites for microRNAs or RNA-binding proteins [28]. We wondered whether the terminator strength of putative polyadenylation sites affects the choice of which site is used. To address this question, we measured the terminator strength of primary and secondary cleavage and polyadenylation sites from over 5000 genes. However, there was no difference between the terminator strength of primary and secondary polyadenylation sites for *Arabidopsis* or maize terminators **(Supplemental Figure 2.9)**.

While terminator strength could not predict which potential polyadenylation site is more likely to be used in a genomic context, we wondered if the reverse were true: Is the strength

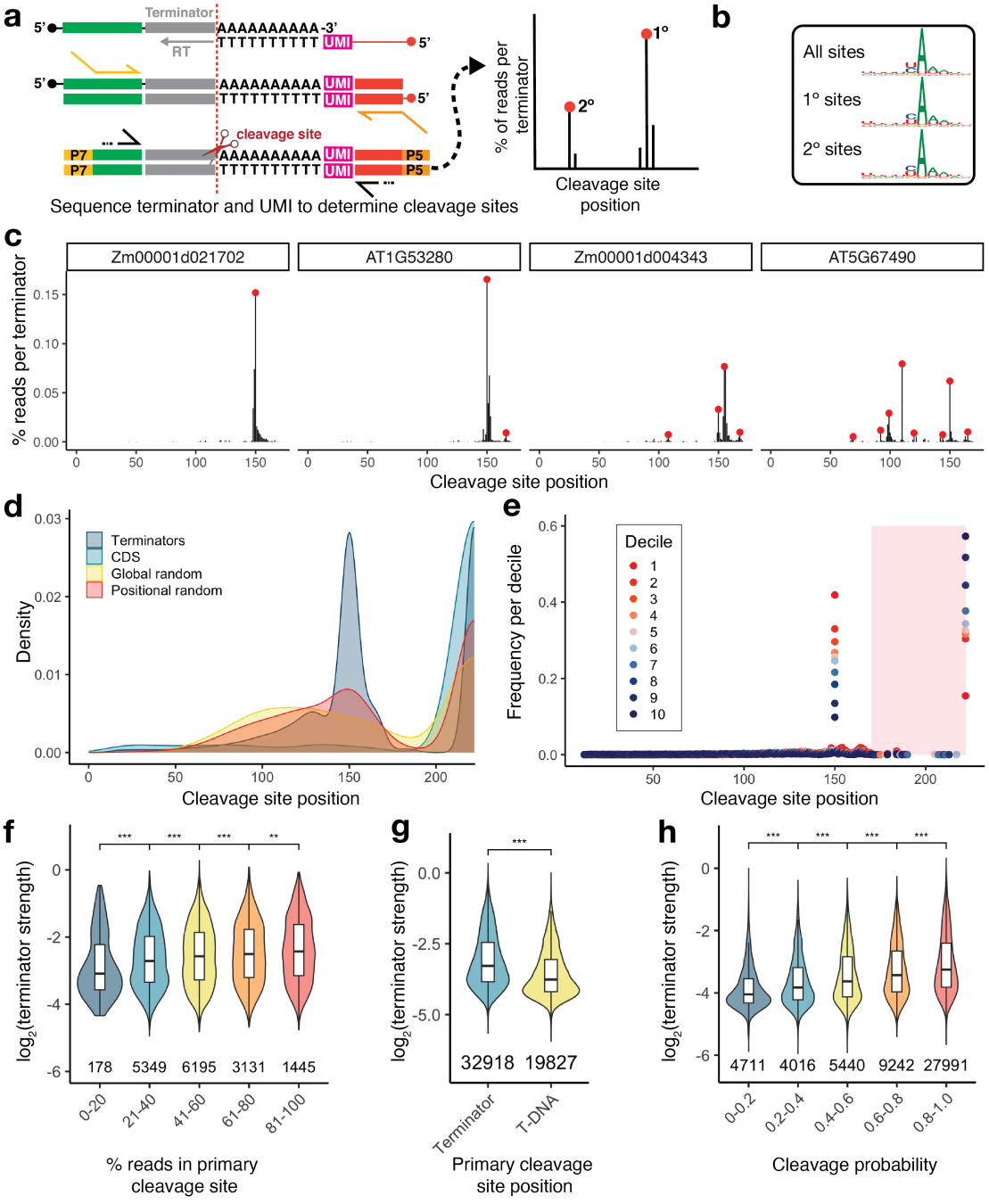

a
Terminator
5'
AAAAAAAAAA-3'
TTTTTTTTTT
UMI
RT
cleavage site
P7
P5
Sequence terminator and UMI to determine cleavage sites
% of reads per terminator
1°
2°
Cleavage site position
b
All sites
1° sites
2° sites
c
Zm00001d021702
AT1G53280
Zm00001d004343
AT5G67490
% reads per terminator
Cleavage site position
d
Terminators
CDS
Global random
Positional random
Density
Cleavage site position
e
Decile
Frequency per decile
Cleavage site position
f
log2(terminator strength)
178
5349
6195
3131
1445
0-20
21-40
41-60
61-80
81-100
% reads in primary cleavage site
g
32918
19827
Terminator
T-DNA
Primary cleavage site position
h
4711
4016
5440
9242
27991
0-0.2
0.2-0.4
0.4-0.6
0.6-0.8
0.8-1.0
Cleavage probability

Figure 2.8: **Polyadenylation and cleavage affect terminator strength.** a Polyadenylation and cleavage sites for terminators in tobacco leaves were determined by 3' end sequencing. Using an oligo-dT primer with a unique molecular identifier (UMI), cDNA was generated from polyadenylated terminators. The cDNA was subjected to paired-end sequencing of the terminator and UMI. Cleavage sites were defined as local maxima in a histogram of terminator read length after trimming of the poly-A tail. b Sequence logo plots generated from the 10 bp window around all cleavage sites (All sites), all primary sites (1° sites), or all secondary sites (2° sites). c Histograms of cleavage site position for four representative terminators. Red dots denote primary and alternative cleavage sites. d Kernel density distribution of the primary cleavage site position for plant terminators (Terminators), sequences from coding regions (CDS), and randomized sequences with an overall (Global random) or per-position (Positional random) nucleotide frequency similar to an average *Arabidopsis* or maize terminator. e Histogram of the primary cleavage site position for terminators grouped into deciles based on terminator strength in tobacco leaves (decile 1 contains the strongest 10% of the terminators and decile 10 the weakest 10%). Each decile contains approximately 5,300 terminators. The shaded red area corresponds to cleavage in the T-DNA backbone (i.e. downstream of the terminator sequence). f Violin plots of terminator strength. Terminators were grouped by the percentage of reads that coincide with the primary cleavage site. g Violin plots of the strength of terminators with a primary cleavage site within the terminator or in the T-DNA. h Violin plots of terminator strength for terminators grouped by cleavage probability (percent of unique reads showing cleavage within the terminator sequence). Violin plots, box plots, and significance levels in f-h are as defined in Figure 2.1.

of a terminator associated with the efficiency of its cleavage and polyadenylation? To determine the cleavage and polyadenylation sites in our terminators, we subjected the output RNA from a Plant STARR-seq experiment in tobacco leaves to 3' end sequencing **(Figure 2.8a)**. We defined the cleavage and polyadenylation sites for each terminator as local maxima in the distribution of the cleavage positions along the terminator sequence for all unique reads. As expected, the vast majority of the cleavage and polyadenylation sites identified by this approach coincided with a CA or UA dinucleotide **(Figure 2.8b)**, matching the known cleavage element [124, 186]. Consistent with prior reports [263, 215, 265, 214], we observed near complete cleavage of the 35S terminator with one major cleavage site at the same position as previously shown **(Supplemental Figure 2.10b)**. We confirmed the cleavage pattern of several terminators with RT-PCR and gel electrophoresis **(Supplemental Figure 2.10c)**.

The terminators in our library were designed to contain a cleavage site at position 150 (Fig. 1a). About 61% of the terminators showed a cleavage site within 5 nucleotides upstream or downstream of this position, and for approximately half of these (32% of all terminators), this was the primary cleavage site **(Table 2.1)**. Of the 54,825 terminators for which we could call cleavage sites, 68% had more than one cleavage sites, and of those, 20% (about 13% of all terminators) had a prominent secondary cleavage site indicated by at least 30% of the reads **(Table 2.1; see Figure 2.8c for examples)**. However, 17% of the tested sequences were not cleaved and polyadenylated but showed read-through into the downstream T-DNA sequence.

Terminator group	**Cleavage site at position 150 ± 5 bp**	**Primary cleavage site at position 150 ± 5 bp**	**Multiple cleavage sites**	**Strong secondary cleavage site (≥ 30% of reads)**	**Cleavage in T-DNA backbone only**
All terminators (n=54,825)	33,700 (61%)	17,986 (32%)	37,301 (68%)	7,393 (13%)	9,525 (17%)
Arabidopsis terminators (n=22,587)	15,331 (68%)	8,826 (39%)	16,827 (74%)	4,421 (20%)	1,082 (5%)
Maize terminators (n = 27,733)	16,931 (61%)	8,540(31%)	18,008 (65%)	2,474 (9%)	6,879 (25%)
CDS (n = 936)	36 (4%)	6 (1%)	234 (25%)	17 (2%)	663 (71%)
Global random sequences (n = 381)	68 (18%)	15 (4%)	227 (60%)	53 (14%)	107 (28.1%)
Positional random terminators (n = 1,865)	773 (41%)	285 (15%)	1,250 (67%)	266 (14%)	388 (21%)

Table 2.1: Terminator polyadenylation and cleavage statistics.

We observed differences in the cleavage pattern between *bona fide* terminators and control sequences **(Figure 2.8d and Table 2.1)**. Plant terminators were predominantly cleaved at position 150. In contrast, randomized sequences showed a broad distribution of cleavage sites throughout their sequence. However, for the positional random sequences which resemble terminators more closely than the global random controls, the distribution was shifted towards an enrichment of cleavage at or near position 150. Finally, control se-

quences derived from coding regions were rarely cleaved and polyadenylated, consistent with a selection against potential cleavage sites. These general trends held true when considering all cleavage sites, in addition to primary sites **(Supplemental Figure 2.10a)**.

Next, we tested the association of terminator strength and cleavage probability (i.e., the likelihood of a terminator being cleaved and polyadenylated instead of allowing read-through transcription). Cleavage probability was strongly correlated with terminator strength **(Figure 2.8e, f)**, and terminators with a primary cleavage site within their sequence were approximately 40% stronger than those without **(Figure 2.8g)**. Furthermore, terminators with a dominant cleavage site (i.e., most transcripts are cleaved at this site) were stronger terminators than those with multiple weak cleavage sites **(Figure 2.8f)**. Taken together, the efficiency and accuracy with which a terminator is cleaved and polyadenylated determines its strength. Strong terminators show a dominant cleavage site and prevent read-through into downstream sequences.

2.2.9 Polyadenylation motifs and GC content control cleavage probability

Since cleavage probability and terminator strength were positively correlated, we asked if the same features that influence terminator strength also affect cleavage probability. First, we tested if biological terminator function was associated with a high cleavage probability. Indeed, plant terminators showed a much higher cleavage probability than control sequences from coding regions **(Figure 2.9a)**. Randomized control sequences with a nucleotide composition similar to an average terminator were also significantly more likely to be cleaved and polyadenylated than the coding region controls, but their average cleavage probability was still lower than that of *bona fide* plant terminators. The cleavage probability for sequences derived from *Arabidopsis* was higher than the one for maize sequences **(Figure 2.9a)**.

As GC content is one of the key differences between terminators from *Arabidopsis* and maize, we asked if cleavage probability was affected by GC content. We observed a strong negative correlation (Pearson's R = -0.65) between these two factors. While a high AT content was associated with a high cleavage probability, increasing GC content led to decreased

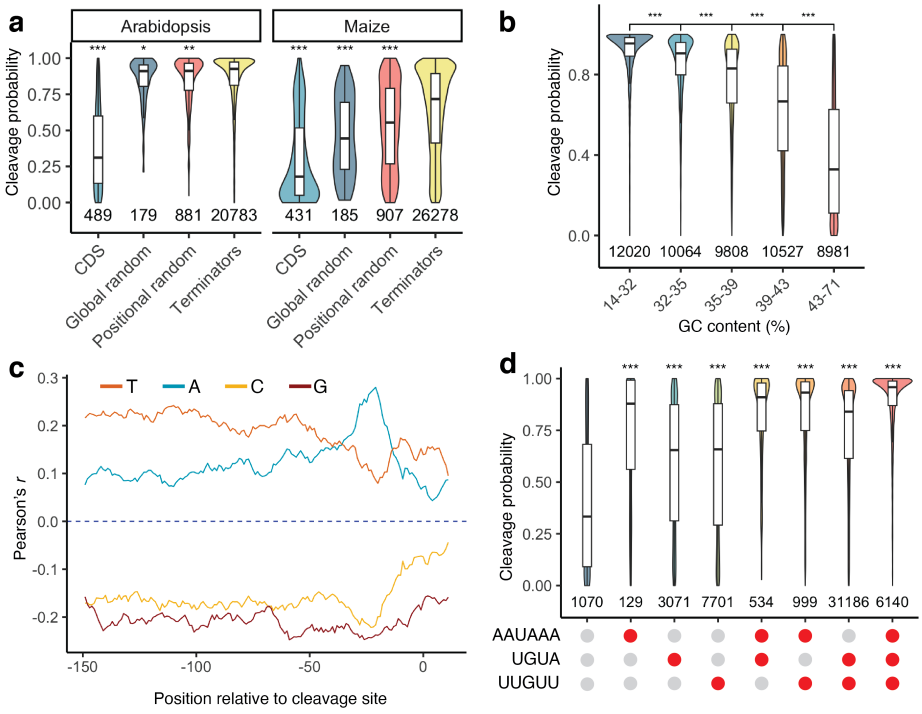


Figure 2.9: **GC content and polyadenylation motifs influence cleavage probability.** a Violin plots, box plots, and significance levels cleavage probability (percent of unique reads showing cleavage within the terminator sequence) for plant terminators (Terminators) compared to sequences from coding regions (CDS) and randomized sequences with an overall (Global random) or per-position (Positional random) nucleotide frequency similar to an average *Arabidopsis* or maize terminator. Each control group is correlated to the terminator group. b Violin plots of cleavage probability. Terminators were binned by GC content to yield groups of approximately the same size. c Correlation (Pearson's R) between cleavage probability and the A, C, G, or U content of a ten-base window starting at the indicated position in the terminators. d Violin plots of cleavage probability for terminators with (red dot) or without (gray dot) the indicated motifs. Violin plots, box plots, and significance levels in a, b, and d are as defined in Figure 2.1.

cleavage probability **(Figure 2.9b)**. This finding likely explains why maize sequences, which show on average an approximately 10% higher GC content than *Arabidopsis* sequences, were less likely to be cleaved than *Arabidopsis* sequences. Our observation that AU-rich sequences show high cleavage probability is consistent with previous reports that the insertion of AU-rich sequences into transcripts can lead to their cleavage and polyadenylation in maize [190]. Although both terminator strength and cleavage probability were affected by GC content, they followed different trends. Cleavage probability decreased monotonically with increasing GC content, while terminator strength was highest at an intermediate GC content and lowest in very AT- or GC-rich sequences **(Figure 2.3b, c)**. Similarly, we observed differing and position-specific effects of local nucleotide composition on terminator strength, whereas for cleavage probability, positional effects were much less pronounced, and A and T nucleotides behaved mostly similar to each other as did G and C nucleotides **(Figure 2.9c)**.

Finally, we tested if the AAUAAA, UGUA and U/G-rich motifs affected cleavage probability. While all three motifs increased cleavage probability, the AAUAAA motif clearly had the largest effect **(Figure 2.9d)**. As for terminator strength, we observed additive effects of the motifs on cleavage probability, and terminators with all three motifs were most likely to be cleaved and polyadenylated.

2.3 Discussion

Here, we adapted Plant STARR-seq to characterize and optimize plant terminator sequences. We find that thousands of *Arabidopsis* and maize terminators outperformed the commonly used bacterial NOS and MAS terminators; a handful even outperformed the viral 35S terminator. These results provide a large arsenal of diverse and strong plant terminators to use with transgenes, which should alleviate transgene silencing [70]. The optimal terminators to employ in constructing transgenes may be those with high cleavage probability, as these prevent the read-through transcription implicated in silencing [78]. Thus, for use in dicots, good terminators to choose would be those from the *Arabidopsis* genes AT3G46230 (HEAT SHOCK PROTEIN 17.4), AT2G05530, and AT4G39730 (PLAT DOMAIN PROTEIN 1); for use in monocots, good ones would be those from the maize genes

Zm00001d016542 (anthranilate 1,2-dioxygenase), Zm00001d047961, and Zm00001d017119 (glucose-6-phosphate dehydrogenase 5).

In addition to these naturally occurring terminators, *in silico* evolved, synthetic terminator sequences — several of which showed greater strength than the viral 35S terminator — resulted from this work. Use of the DenseNet prediction that best matched the assay system (the tobacco leaf prediction for dicots, and the maize protoplast prediction for monocots) led to the best results; however, a combination of both predictions can generate terminators that are active across a broad range of plant species, tissues, and applications. The iterative *in silico* evolution and testing of evolved elements in Plant STARR-seq appears to be a promising generalizable path toward improving the features of plant regulatory elements, as for core promoters [141, 142].

Beyond the practical benefits of this study, we identified novel features of terminator biology in plants. We found that GC content profoundly influences terminator strength and cleavage probability in different ways. The strongest terminators had a GC content of approximately 30 – 40% (depending on the assay system), with terminator strength lowest at either higher or lower GC percentage in both assay systems. In contrast, cleavage probability monotonically decreased with increasing GC content in tobacco leaves (and was not measured in maize protoplasts due to technical challenges). This negative correlation is similar to the effects of GC content on core promoter strength, which shows a strong negative correlation with GC content in tobacco leaves but not in maize protoplasts [142]. Taken together, GC content appears to affect different regulatory elements in different ways, and these effects vary across species, reflecting genomic GC content.

In addition to known polyadenylation motifs [28, 62, 124], we discovered novel motifs that increased terminator strength and cleavage probability. These motifs are rich in U and G nucleotides and reside broadly across terminator sequences, with a moderate enrichment immediately upstream and downstream of the cleavage site. The precise composition of these U/G-rich motifs differed between the tobacco leaf and the maize protoplast systems; however, the maize-specific motif also contributed to terminator strength in tobacco leaves. Human terminators contain a GU-rich element downstream of the cleavage site that is required for efficient polyadenylation and that is bound by the cleavage-stimulating factor

(CSTF) complex [29, 301, 305]. Plant homologs for members of the CSTF complex can form a similar complex and bind to RNA [26, 130, 354]. We speculate that the novel U/G-rich motifs identified here serve as binding sites for the plant CSTF complex.

The assay design employed here provided insight into cleavage probability and alternative polyadenylation. Alternative polyadenylation is a crucial mechanism to create diverse transcripts in response to environmental change and in particular tissues [28, 343, 358]. Alternative polyadenylation produces transcripts with altered 3' UTR length, resulting in different mRNAs and translation levels. Long 3' UTRs decrease transcript stability by triggering nonsense-mediated mRNA decay [150], while short 3' UTRs make translation more efficient by promoting polysome formation [44]. When we compared the 170 nucleotide-terminator sequences of *Arabidopsis* and maize genes derived from primary vs. secondary polyadenylation sites, we observed no difference in terminator strength among sites derived from *Arabidopsis* or maize. These comparisons suggest that the sequence surrounding a given polyadenylation site does not fully explain alternative polyadenylation. Trans-acting factors such as the components of the cleavage and polyadenylation complexes, RNA-binding proteins and non-coding RNAs are involved in the choice of polyadenylation sites, presumably by recognizing sequence context and secondary structure of the entire terminator sequence instead of the shorter terminator sequences tested here [28, 130, 365]. A comparison of all tested terminator sequences indicates that strong terminators tended to contain a dominant cleavage site rather than multiple weaker ones.

2.4 Methods

2.4.1 Library Design and Construction

We defined terminators as the sequence from 150 nucleotides upstream to 20 nucleotides downstream of a cleavage and polyadenylation site. For *Arabidopsis*, we used experimentally determined cleavage and polyadenylation sites described by Thomas et al. [304]. We selected the primary cleavage and polyadenylation sites for each gene (n = 18,450), as well as prominent secondary cleavage sites that had at least 30% of total reads per gene (n = 2,325). For the 3,754 genes without an experimentally defined cleavage site, we utilized the

end of their 3' UTR annotation in the *Arabidopsis* TAIR10 annotation as the cleavage site. This yielded a total of 24,529 *Arabidopsis* terminator sequences **(Supplementary Table 1)**. For maize, we used cleavage and polyadenylation sites defined by Jafar et al [133]. We selected primary cleavage sites for 25,685 maize genes, and included 4,407 secondary cleavage sites (at least 30% of total reads per gene) for a total of 30,092 maize terminator sequences **(Supplementary Table 1)**.

In addition to *bona fide* terminators, several control sequences were added to the library **(Supplementary Table 1)**. Since protein coding sequences should not contain cleavage and polyadenylation motifs, 170-nucleotide long sequences (589 each from *Arabidopsis* and maize) from coding regions with a similar GC content as the terminator library were included as negative controls. Furthermore, we generated random sequences with global (200 sequences per species) or position-specific (1,000 sequences per species) nucleotide frequencies derived from the average *Arabidopsis* or maize terminator. Randomized sequences with a GC content of 30%, 40%, 50%, 60%, or 70% (200 sequences of 170 nucleotides each) were added to test effects of GC content on terminator strength. Finally, four commonly used non-plant terminators (t35S, tAg7, tNOS, and tMAS) were added as a frame of reference in their full length (35S=204 nt, Ag7=208 nt, MAS= 253 nt, NOS=253 nt). The final library of sequences was ordered as an oligo pool from Twist Biosciences **(Supplementary Table 2)**.

To validate and extend our findings from the large-scale terminator library, we created a second, smaller library. This library included 187 terminators from the original library to test the correlation between results from the two libraries. Additionally, we added sequences to validate the importance of the discovered motifs. We picked 20 sequences from the randomized control sequences in our original library that had neither a CA dinucleotide at the predicted cleavage site, nor an AAUAAA, a UGUA, or a GUGUG motif. For each validation sequence, we added a CA dinucleotide at position 150 and all possible combinations of these three motifs at the positions where they are found most often in our library. To test the importance of the AAUAAA motif, we picked twenty terminators that had only one AAUAAA sequence and generated three variants with a mutation in this sequence (ACTCAA, ACTCAA, ACTAAA). Similarly, we took twenty sequences that had only one

UGUA sequence and replaced it with mutated variants (UGCA, CGUA, UCUA). Finally, we included the sequences from the *in silico* evolution (see below) in this validation library **(Supplementary Table 3)**.

The Plant STARR-seq plasmid pPSt (https://www.addgene.org/203590/) used in this study is based on the pPSup plasmid (https://www.addgene.org/149416/) [141] with adaptations to enable the characterization of terminators. The base plasmid pPSt contains three consecutive Golden Gate cloning sites. To generate the terminator libraries, we first inserted the 35S promoter together with a 5' UTR derived from the maize histone H3 gene Zm00001d041672, an ATG start codon, and a 18-nucleotide random barcode (VNNVNNVNNVNNVNNVNN, where V=A, C or G) into the upstream Golden Gate site using the restriction enzyme BbsI-HF (NEB). Next, we cloned the array-synthesized terminator library into the downstream Golden Gate site using the restriction enzyme Esp3I (NEB). In this step, the plasmid library was bottlenecked to approximately 1,000,000 variants (approximately 16 barcodes per terminator) or 60,000 variants (approximately 30 barcodes per terminator) for the large-scale or the validation library, respectively. Finally, we used BsaI-HFv2 (NEB) to replace the central Golden Gate site with the coding sequence of GFP lacking the start codon. Two versions of the library were created in this step: one with full-length GFP for use in Plant STARR-seq (see https://www.addgene.org/203592/ for an example of a fully assembled Plant STARR-seq plasmid with the 35S terminator.) and one with a truncated version of GFP lacking the central 188 amino acids for subassembly (see below). The sequences of all primers used for cloning and sequencing are included in **Supplementary Table 4**.

2.4.2 Tobacco cultivation and transformation

Tobacco (*Nicotiana benthamiana*) was grown in soil (Sunshine Mix no. 4) at 25°C with a long day photoperiod (16h light and 8h dark; cool-white fluorescent lights Philips TL-D 58W/840; intensity 300 $\mu molm^{-2}s^{-1}$). Plants were transformed approximately 3 weeks after germination. The *Agrobacterium* transformation and induction method was previously described [142]. The terminator libraries were introduced into *Agrobacterium tumefaciens*

strain GV3101 (harboring the virulence plasmid pMP90 and the helper plasmid pMisoG) by electroporation. *Agrobacterium* cultures were infiltrated into the first two mature leaves of tobacco plants (12 plants/24 leaves per replicate for the large-scale library; 3 plants/6 leaves for the smaller validation library). The plants were further grown for 48 hours under normal light conditions before mRNA extraction. The input sample for Plant STARR-seq was obtained from 20 mL of the *Agrobacterium* solution that was used to infiltrate tobacco leaves. To this end, the QIAprep Spin Miniprep Kit was used according to the manufacturer's instructions.

2.4.3 Maize mesophyll protoplast generation and PEG transformation

We used the PEG transformation method of maize mesophyll protoplasts as described in [310]. Maize (*Zea mays* L. cultivar B73) seeds were soaked in water overnight at 25°C. The seeds were germinated in soil for 3 days under long day conditions (16 hours light, 8 hours dark) at 25°C, then moved to complete darkness at 25°C for 10-11 days. From each seedling, 10 cm sections from the second and third leaf were cut into thin 0.5 mm strips perpendicular to veins and immediately submerged in 10 ml of protoplasting enzyme solution (0.6 M mannitol, 10 mM MES ph 5.7, 15 mg/ml cellulase R10, 3 mg/ml macerozyme, 1 mM CaCl2, 0.1% [w/v] BSA, and 5 mM beta-mercaptoethanol). The mixture was covered in foil to keep out light, vacuum infiltrated for 3 min at room temperature (RT), and incubated on a shaker at 40 rpm for 2.5 hours at RT. Protoplasts were released by incubating an extra 10 min at 80 rpm. To quench the reaction, 10 mL ice-cold MMG (0.6 M Mannitol, 4 mM MES ph 5.7, 15 mM MgCl2) was added to the enzyme solution and the whole solution was filtered through a 40 µM cell strainer. To pellet protoplasts, the filtrate was split into equal volumes of no more than 10 mL in chilled round-bottom glass centrifuge vials and centrifuged at 100 x g for 4 min at RT. Pellets were resuspended in 1 mL cold MMG each and combined into a single round-bottom vial. To wash, MMG was added to make a total volume of 5 mL and the solution was centrifuged at 100 x g for 3 min at RT. This wash step was repeated two more times. The final pellet was resuspended in 1-2 mL of MMG. A sample of the resuspended protoplasts was diluted 1:20 in MMG and used to count the

number of viable cells using Fluorescein Diacetate as a dye.

For each replicate, 20 million protoplasts were mixed with 200 µg of the terminator plasmid library in a fresh tube, topped with MMG to a volume of 2,288 µL, and incubated on ice for 30 min. For PEG transformation, 2,112 µL of PEG solution (0.6 M Mannitol, 0.1 M CaCl2, 25% [w/v] poly-ethylene glycol MW 4000) was added to reach a final concentration of 12% (w/v) PEG. The mixture was incubated for 10 min in the dark at RT. After incubation, the transformation solution was diluted with 22 mL (5 x 4.4 mL) incubation solution (0.6 M Mannitol, 4 mM MES pH 5.7, 4 mM KCl), and centrifuged at 100 x g for 4 min at RT. For the PEG transformation with the validation library, we used 4 million protoplasts, 40 µg of the plasmid library, and one fifth of the buffer volumes used for the large-scale library.

After transformation, the protoplast pellet was washed with 5 mL of incubation solution, centrifuged at 100 x g for 3 min at RT, and resuspended in incubation solution to a concentration of 500 cells/µL. Protoplasts were incubated overnight in the dark at RT to allow for transcription of the plasmid library and then pelleted (4 min, 100 x g, RT). The pellet was washed with 5 mL incubation solution and centrifuged (3 min, 100 x g, RT). The pellet was finally resuspended in 5 mL incubation solution. An aliquot of the solution was used to check transformation efficiency under a microscope (12.4% [replicate 1] and 49.5% [replicate 2] transformation efficiency for the large-scale library; 62% [replicate 1] and 45% [replicate 2] for the validation library). Cells were pelleted (4 min, 100 x g, RT) and resuspended in 2 mL Trizol for subsequent mRNA extraction. An aliquot of the plasmid library used for PEG transformation was used as the input sample for Plant STARR-seq.

2.4.4 Plant STARR-seq assay

For all Plant STARR-seq experiments, two independent biological replicates were performed. Different plants and fresh *Agrobacterium* cultures were used for each biological replicate. Tobacco leaves were harvested 2 days after infiltration and partitioned into batches of 4 (large-scale library) or 3 (validation library) leaves. The leaf batches were frozen in liquid nitrogen, finely ground with mortar and pestle, and immediately resuspended in 12 mL Trizol. The suspensions were cleared by centrifugation (5 min, 4,000 x g, 4°C) and

each supernatant was mixed with 2.5 mL chloroform. After centrifugation (15 min, 4,000 x g, 4°C), 7 mL of the upper, aqueous phase was transferred to a new tube, and mixed by inversion with 3.5 mL high salt buffer (0.8 M sodium citrate, 1.2 M NaCl) and 3.5 mL isopropanol. The solution was incubated for 15 min at RT to precipitate the RNA and centrifuged (30 min, 4,000 x g, 4°C). The pellet was washed in 10 mL ice-cold 70% ethanol, centrifuged (5 min, 4000 x g, 4°C), and air-dried. The pellet was resuspended in 180 µL of warm (65°C) nuclease-free water and transferred to a new tube. The solution was supplemented with 10 µL 20X DNase I buffer (1 mM CaCl2, 100 mM Tris pH 7.4), 10 µL 200 mM MnCl2, 2µL DNase I (ThermoFisher Scientific), and 1 µL RNaseOUT (ThermoFisher Scientific), and incubated for 30 min at 37°C. To precipitate the RNA, 20 µL 8M ice-cold LiCl and 500 µL ice-cold 100% ethanol was added. After incubation for 15 min at -80°C, the RNA was pelleted by centrifugation (20 min, 20,000 x g, 4°C). The pellet was washed with 500 µL ice-cold 70% ethanol, centrifuged (5 min, 20,000 x g, 4°C), air-dried for 10 min, and resuspended in 200 µL nuclease-free water supplemented with 0.5 µL RNaseOUT. For cDNA synthesis, eight reactions with 11 µL mRNA solution, 1 µL 2 µM GFP-specific reverse transcription primer, and 1 µL 10 mM dNTPs were incubated at 65°C for 5 min then immediately placed on ice. The reactions were supplemented with 4 µL 5X SuperScript IV buffer, 1 µL 100 mM DTT, 1 µL RNaseOUT, and 1 µL SuperScript IV reverse transcriptase (ThermoFisher Scientific). To ensure that the samples were largely free of DNA contamination, four reactions were used as controls, where the reverse transcriptase and RNaseOUT were replaced with water. Reactions were incubated for 10 min at 55°C, followed by 10 min at 80°C. Sets of 4 reactions each were pooled. The cNDA was purified with the Zymo Clean&Concentrate-5 kit, and eluted in 20 µL 10 mM Tris. The barcode was amplified with 10-20 cycles of polymerase chain reaction (PCR) and read out by next generation sequencing.

For Plant STARR-seq in maize protoplasts, the protoplast-containing Trizol solution from PEG transformation was transferred to 2 mL Phasemaker tubes (1 mL per tube; ThermoFisher Scientific), mixed thoroughly with 300 µL chloroform, and centrifuged (5 min, 15,000 x g, 4°C). RNA was extracted using the RNeasy Plant Mini Kit (QIAGEN). The supernatant was transferred to a QIAshrededer column and centrifuged (2 min, 20,000

x g, RT). The flowthrough was transferred to a new 1.5 mL tube and mixed with 300 µL 100% ethanol. Up to 500 µL of the solution was loaded on an RNeasy mini spin column. After centrifugation (10 seconds, 16,100 x g, RT) the flowthrough was discarded. This was repeated until all the solution had been added to the column. The column was washed once with 700 µL RW1 buffer and twice with 500 µL RPE buffer. After each wash step, the column was centrifuged (30 sec, 16,100 x g, RT) and the flowthrough was discarded. The column was dried with an extra centrifugation step (30 sec, 16,100 x g, RT) and transferred to a 1.5 mL collection tube. For elution, 50 µL of RNase-free water was added, and the column was incubated for 1 minute, and centrifuged (1 min, 16,100 x g, RT). This elution step was repeated with an additional 40 µL of RNase-free water. The eluate was treated with DNase I (5 µL of 20x DNaseI buffer, 5 µL 200 mM MnCL2, 1 µL RNaseOUT, and 2 µL DNase I) for 1 h at 37°C. The solution was supplemented with 20 µL 500 mM EDTA, 1 µL 20 mg/mL glycogen, 12 µL ice-cold 8M LiCl, and 300 µL ice-cold 100% ethanol. The solution was incubated 15 min at -80°C, centrifuged (20 min, 20,000 x g, 4°C). The pellet was washed with 500 µL ice-cold 70% ethanol, and centrifuged (3 min, 20,000 x g, 4°C). The pellet was air-dried for 10 min and resuspended in 100 µL RNase-free water. Reverse transcription, purification, PCR amplification and sequencing were performed as for the tobacco samples.

2.4.5 Subassembly and barcode sequencing

Paired-end sequencing on an Illumina NextSeq 2000 platform was used to link terminators to their respective barcodes. To facilitate sequencing, we created a shortened version of the terminator library plasmid with a large deletion in the GFP gene. The terminator region was sequenced using paired 151-nucleotide reads, and two 18-nucleotide indexing reads were used to sequence the barcodes. The paired terminator and barcode reads were assembled using PANDAseq (version 2.11) [201] and the terminator reads were aligned to the designed terminator library using BowTie2 (version 2.4.1) [168]. Terminator-barcode pairs with less than 5 reads and terminators with a mutation or truncation were discarded. For each Plant STARR-seq experiment, barcodes were sequenced using paired-end reads on an Illumina

NextSeq 2000 system. The paired barcode reads were assembled using PANDAseq.

2.4.6 Computational methods

For calculating terminator strength, the reads for each barcode were counted in the input and cDNA (output) samples. Barcodes with less than five reads were discarded. For each terminator, the sum of the reads for all associated barcodes was calculated in the input and output samples. The input and output counts were normalized by dividing each by the sum of all counts in the output and input sample, respectively. Terminator strength was calculated as the normalized output counts divided by the normalized input counts. Terminator strength was normalized to the strength of the 35S terminator. We used the average terminator strength across two replicates for all analyses **(Supplementary Tables 2 and 3)**. Spearman and Pearson's correlation were calculated using base R (version 4.3.0). Significance was calculated using the two-sided Wilcoxon rank-sum test and the ggsignif library (version 0.6.4) in R **(Supplementary Table 5)**. GO term enrichment analysis was performed using the ggprofiler2 package (version 0.2.1) in R **(Supplementary Table 6)**.

2.4.7 Discovery of terminator motifs

To find motifs enriched in strong terminators, we ranked terminators according to their strength in tobacco leaves or maize protoplasts. The sequences of the top 25% terminators were analyzed by STREME (version 5.5.1) [20] to find ungapped RNA motifs that are enriched in this set relative to the sequences of the bottom 25% terminators. To find tobacco-specific terminator motifs, we used STREME with the same parameters but using the top 10% Arabidopsis of terminators ranked by Ψ as positive and the bottom 10% of maize terminators ranked by Ψ as negative sequences. For maize-specific terminator motifs, we repeated the analysis but switched the positive and negative sequences. Meme files with all discovered motifs are available on GitHub (https://github.com/lampoona/Terminators-Plant-STARR-seq).

All terminator sequences were analyzed using the universalmotif package (version 1.18.0) in R to find the position and frequency of the discovered motifs. For each sequence, the

maximum motif score was identified 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 to contain a motif match.

2.4.8 Computational modeling of terminator strength

In order to predict terminator strength, we first build a lasso regression model using the glmnet package (version 4.1.7) in R to predict terminator strength based on the counts of all possible, overlapping 6-mer counts. The regression model was trained on 70% of the data and tested on the remaining 30%.

For our second model, we built a convolutional neural net model using EUGENe (version 0.0.6) [158] and PyTorch (version 1.11.0) [236] in Python (version 3.8.10). We used a 'DenseNet' [121] architecture adapted from iCREPCP [65]. The model takes one-hot encoded DNA as an input which is fed to a convolutional layer with 128 filters and a kernel size of 5. This layer is followed by four dense blocks consisting of 6, 12, 24, and 16 convolutional layers with 12 filters each and a kernel size of 3. In each dense block, the output of a convolutional layer is appended to its input and the combined output is used as the input for the subsequent layer. Between each dense block, the output feature map is reduced in size through convolution (using half as many filters as the input and a kernel size of 1) and average pooling (with a kernel size of 2). The output of the final dense block is fed into a fully connected layer with two outputs corresponding to the terminator strength in tobacco leaves and maize protoplasts, respectively. Terminator sequences from our original library that were detected in tobacco leaves and maize protoplast STARR-seq experiments (n = 53,409) were used for model training and evaluation. Of this data, 81% were used to train the model, 9% were used as a validation set during training and the remaining 10% were used to test the generalizability and accuracy of the trained model. Feature attributions for sequences in the test set were calculated using the DeepLIFT method [283] implemented in EUGENe. These feature attributions were used as input for TF-MoDISco-lite (version 2.0.6) [284] to extract motifs that increase or decrease terminator strength. Since we were looking for RNA motifs, we modified the TF-MoDISco-lite algorithm to not

consider reverse complemented seqlets for clustering, alignments, or pattern generation. The code for model training and evaluation, and the trained model are available on GitHub (https://github.com/lampoona/Terminators-Plant-STARR-seq).

2.4.9 In silico evolution of terminator sequence

We used the DenseNet model to iteratively improve terminator strength. We randomly selected 222 sequences (111 each from *Arabidopsis* and maize) from the test set of our model to ensure that the model had not seen these sequences during training. In each iteration, we generated every possible point mutation for each sequence, scored them with the DenseNet model, and kept the sequence variant with the highest predicted strength in tobacco leaves, maize protoplasts, or both (using the sum of the activity in the individual systems) systems for the next iteration. We then experimentally determined the terminator strength of the sequences after three and ten rounds of evolution using Plant STARR-seq in both tobacco leaves and maize protoplasts.

2.4.10 3' end sequencing

We used 3' end sequencing to determine the cleavage and polyadenylation site for each terminator. For cDNA synthesis, we incubated 5 µL of the extracted mRNA from the second Plant STARR-seq replicate in tobacco leaves with 5 µL RNase free water, 1 µL 10 mM DNTPs, and 2 µL 25 mM UMI-containing oligo-dT primer for 5 min at 55°C. The solution was supplemented with 4 µL 5X SuperScript IV buffer, 1 µL 100 mM DTT, 1 µL RNaseOUT, and 1 µL Superscript IV reverse transcriptase. The reaction was incubated for 2 min each at 4°C, 10°C, 20°C, 30°C, 40°C and 50°C, and finally for 15 min at 55°C. The sample was supplemented with 1 µL RNase H (NEB) and incubated for 30 min at 37°C. The cDNA was purified using the Zymo Clean&Concentrate-5 kit and PCR-amplified using indexed sequencing primers specific to the end of the GFP construct and the adapter added using the RT primer **(Fig. 8a)**. Terminators were sequenced using an Illumina NextSeq 2000 platform with a 222-nucleotide read from the 5' end of the amplicon. The paired read was used to sequence the 8-nucleotide UMI. UMIs and terminator reads were linked using UMI-

Tools (version 1.1.2) [292]. Poly-A tails of each read were removed using CutAdapt (version 2.5) [147]. Terminators were aligned to the designed terminator library plus 52 nucleotide of plasmid backbone sequence using BowTie2 (version 2.4.1) [168] and reads that had a map quality of 0 or 1 were removed using Samtools (version 1.9) [174]. Duplicate terminator-UMI pairs were removed and terminators with fewer than 20 supporting reads were discarded. Bam files for each terminator were generated, and cleavage sites were determined using a custom R script available on GitHub (https://github.com/lampoona/Terminators-Plant-STARR-seq). Briefly, the algorithm identifies local maxima in the distribution of read lengths per terminator. Cleavage probability was calculated on a per terminator basis as the percentage of reads shorter than 171 nucleotides **(Supplementary Table 2).**

2.4.11 Dual-luciferase assay

For the dual-luciferase assay, terminators were cloned downstream of a nanoluciferase gene under control of the 35S enhancer and minimal promoter in the plasmid pDLterm (https://www.addgene.org/211903/). The plasmid also harbors a luciferase gene under control of the Arabidopsis UBQ10 promoter. For the dual-luciferase assay in tobacco, two independent biological replicates, with two technical replicates each, were performed. The dual-luciferase reporter plasmids were introduced into *Agrobacterium tumefaciens* strain GV3101 (harboring the virulence plasmid pMP90 and the helper plasmid pSoup) by electroporation and used for transient transformation of tobacco leaves. Two days after the transformation, a total of 4 leaf discs from the third and fourth leaf of the tobacco plants was collected using a cork borer (4 mm diameter). The leaf discs were transferred to 1.5 mL tubes filled with approximately 10 glass beads (1 mm diameter), snap-frozen in liquid nitrogen, and disrupted by shaking twice for 5 sec in a Silamat S6 (Ivoclar) homogenizer. The leaf disc debris was resuspended in 100 µL 1X Passive Lysis Buffer (Promega). The solution was cleared by centrifugation (5 min, 20,000 x g, RT) and an aliquot of the supernatant was diluted 1:10 with 1X passive lysis buffer. Luciferase and nanoluciferase activities were measured on a Biotek Synergy H1 plate reader using the Promega Nano-Glo Dual-Luciferase Reporter Assay System according to the manufacturer's instructions. Specifically, 10 µL of

the diluted leaf extracts were combined with 75 µL ONE-Glo EX Reagent, mixed for 3 min at 425 rpm, and incubated for 2 min before measuring luciferase activity. Subsequently, 75 µL NanoDLR Stop&Glo Reagent were added to the sample. After 3 min mixing at 425 rpm and 12 min incubation, nanoluciferase activity was measured. For the dual-luciferase assay in maize protoplasts, two independent biological replicates, with a technical replicate for one of them, were performed. One million protoplasts were transformed with 15 µg of the dual-luciferase reporter plasmid following the same transformation protocol as used for Plant STARR-seq (see above). After incubation overnight in the dark, the transformed protoplasts were washed twice with 1 mL incubation solution and lysed by adding 100 µL 1X passive lysis buffer. The protoplast lysate was diluted 1:10 with 1X passive lysis buffer and used to measure luciferase and nanoluciferase activities with the same protocol as used for the tobacco leaf extracts. Raw values and calculated scores from assay are provided in **(Supplementary Table 7)**.

2.4.12 RT-PCR and gel electrophoresis

To visualize terminator cleavage patterns, RNA was extracted from tobacco leaves transiently transformed with dual-luciferase reporter constructs. A total of 8 leaf discs (4 mm diameter) were collected two days after transformation. The fresh leaf discs were transferred to a 2 mL tube containing a 5 mm stainless steel bead and 500 µL buffer RLT (QIAGEN). The leaf discs were disrupted using a TissueLyser LT (QIAGEN) bead mill for 3 min at 50 Hz. Debris was pelleted by centrifugation (2 mi, 20,000 x g, RT) and the supernatant was subjected to RNA extraction using the QIAGEN RNease Plant Mini Kit according to the manufacturer's instructions. Reverse transcription of the extracted RNA was performed in the same way as for the 3' end sequencing. The resulting cDNA was PCR-amplified with primers binding to the 3' end of the nanoluciferase gene and the 5' end of the reverse transcription primer. The PCR products were separated by gel electrophoresis and visualized with SYBR green.

2.5 Data Availability

All sequencing results are deposited in the NCBI Sequence Read Archive under the BioProject accession PRJNA991151.

2.6 Code availability

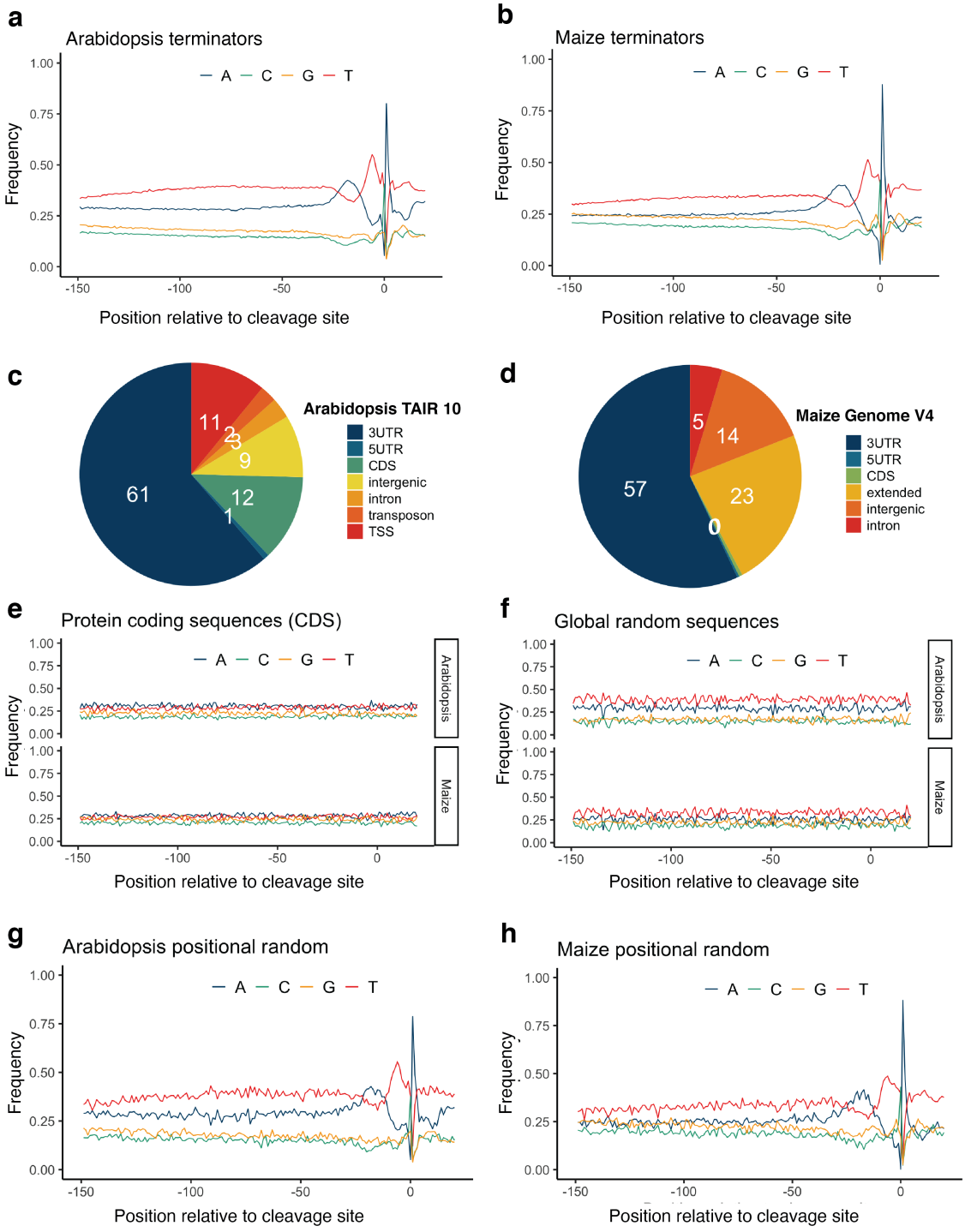
The code used in this study is available on GitHub (https://github.com/lampoona/Terminators-Plant-STARR-seq).

2.7 Author contributions

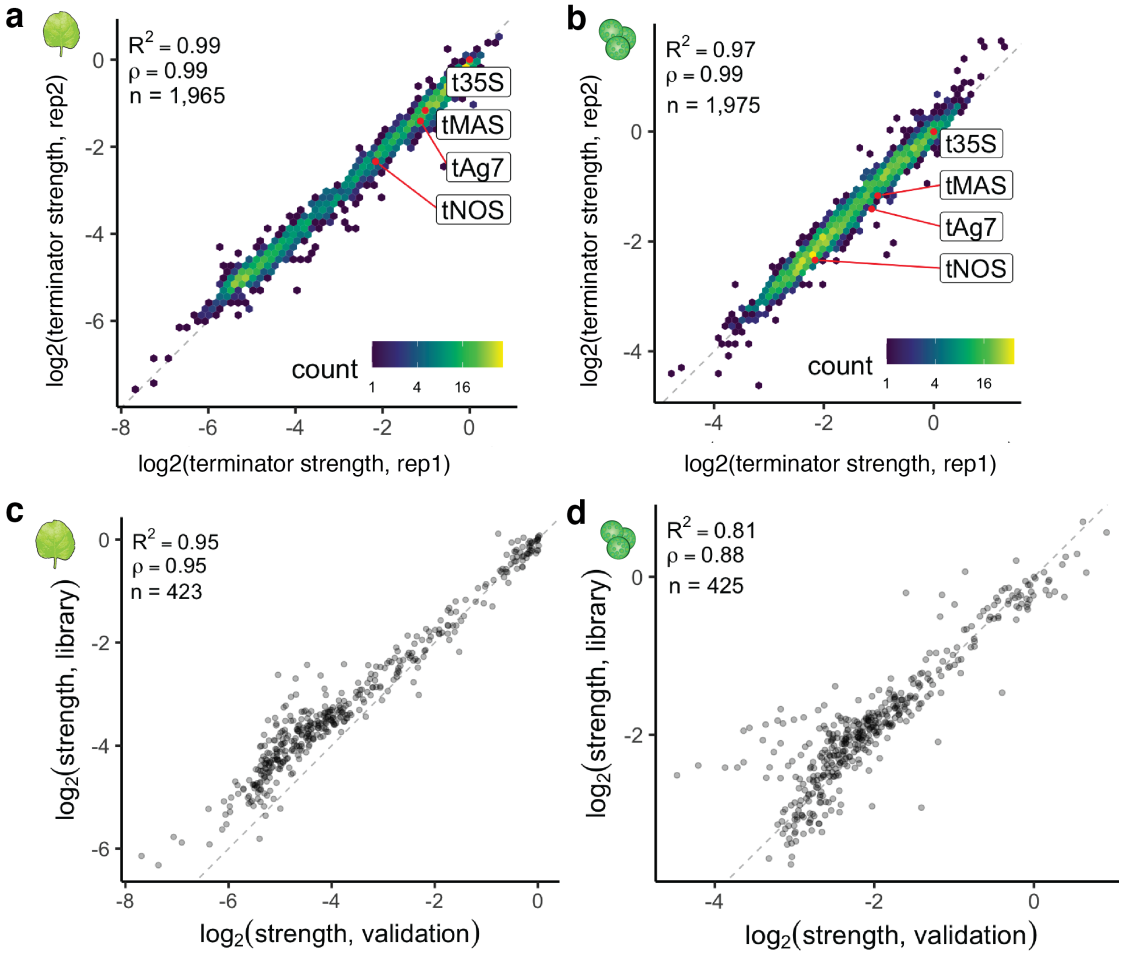
S.G., T.J., and C.Q. constructed the experimental design. S.G. designed terminator libraries with guidance from T.J.. S.G. conducted nearly all the experiments, with the exception of J.T. and N.A.M. who conducted the protoplasting experiments, and analyzed nearly all the data. T.J. was responsible for the tobacco husbandry, CNN analysis, and *in silico* modeling. K.B. wrote a script to call major PAC sites from 3' end sequencing. S.G. wrote the main manuscript and drafted all figures. T.J., C.Q., and S.F. were critical in the manuscript shaping and editing process.

2.8 Supplemental Figures

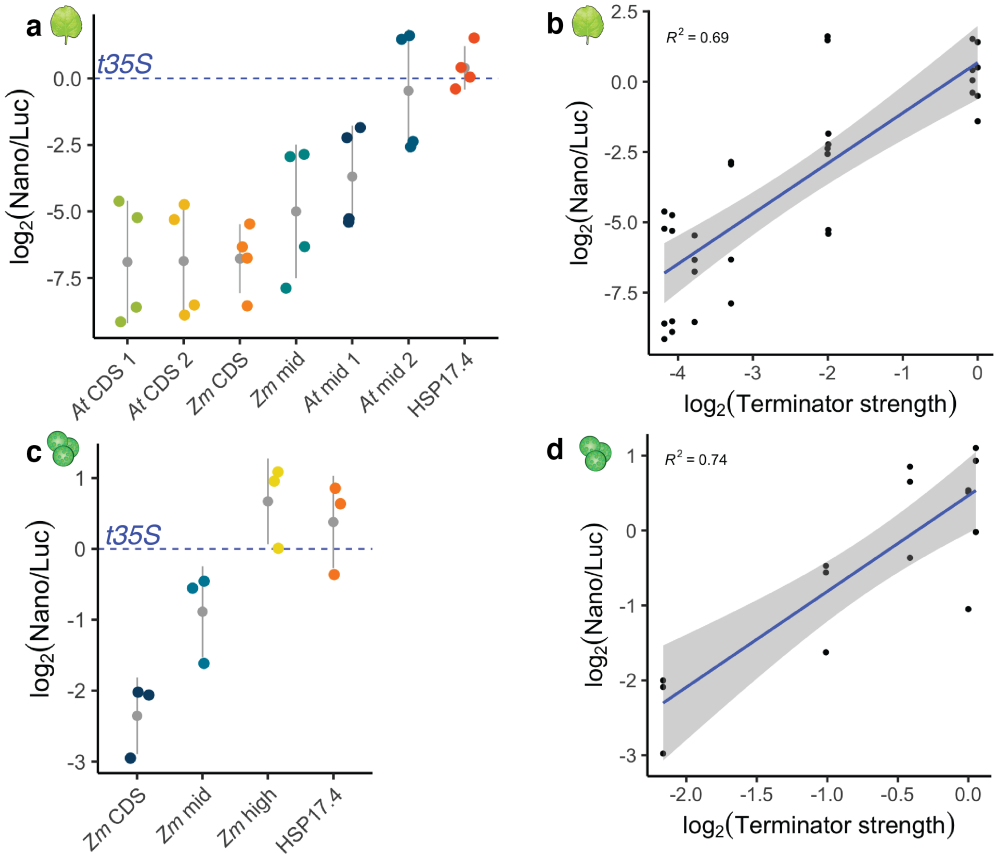
2.9 Supplemental Tables


a
Arabidopsis terminators
A C G T
Frequency
Position relative to cleavage site
b
Maize terminators
A C G T
Frequency
Position relative to cleavage site
c
Arabidopsis TAIR 10
3UTR
5UTR
CDS
intergenic
intron
transposon
TSS
61
1
12
9
3
2
11
d
Maize Genome V4
3UTR
5UTR
CDS
extended
intergenic
intron
57
0
23
14
5
e
Protein coding sequences (CDS)
A C G T
Arabidopsis
Maize
Frequency
Position relative to cleavage site
f
Global random sequences
A C G T
Arabidopsis
Maize
Frequency
Position relative to cleavage site
g
Arabidopsis positional random
A C G T
Frequency
Position relative to cleavage site
h
Maize positional random
A C G T
Frequency
Position relative to cleavage site

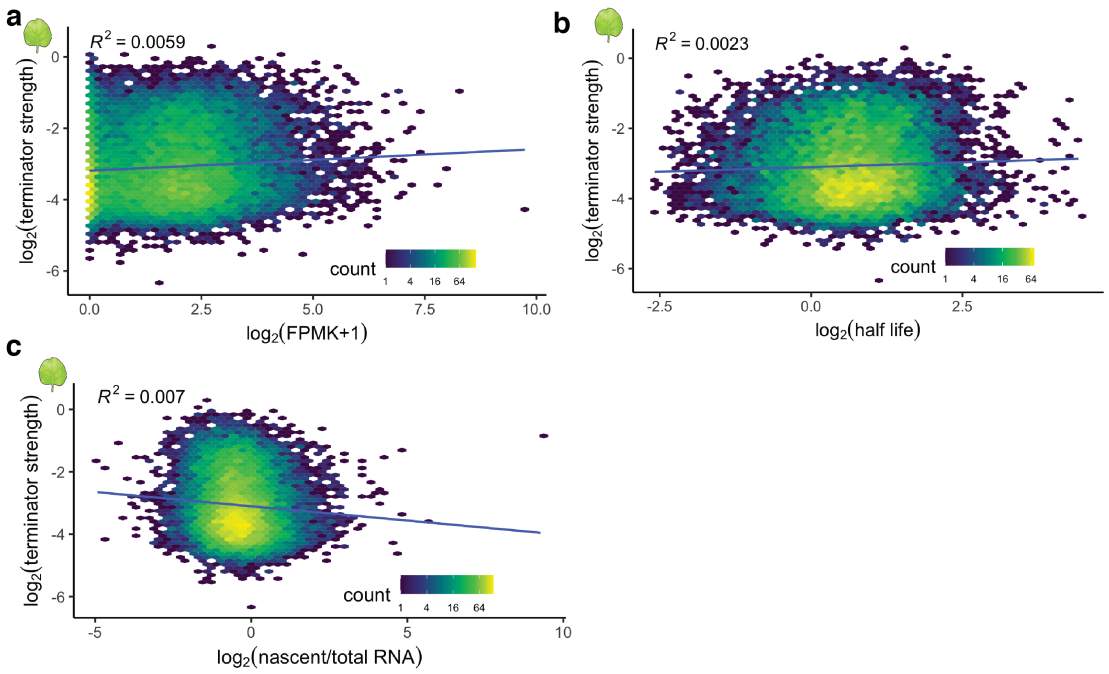
Supplemental Figure 2.1: **Nucleotide composition for terminators in our library**. a, b Per-position nucleotide frequencies of *Arabidopsis* (a) or maize (b) terminators. c, d Pie charts showing the distribution of the cleavage site location (annotated according to the A*Arabidopsis* TAIR10 and the maize B73v4 genome annotations) for the terminators from *Arabidopsis* (c) or maize (d) in our library. e Per-position nucleotide frequencies for controls derived from coding sequences (CDS) in *Arabidopsis* or maize. f Per-position nucleotide frequencies of randomized sequences with an overall (Global random) nucleotide composition similar to average *Arabidopsis* or maize terminator. g, h Per-position nucleotide frequencies for randomized sequences with a per-position (Positional random) nucleotide composition similar to an average *Arabidopsis* (g) or maize (h) terminator.



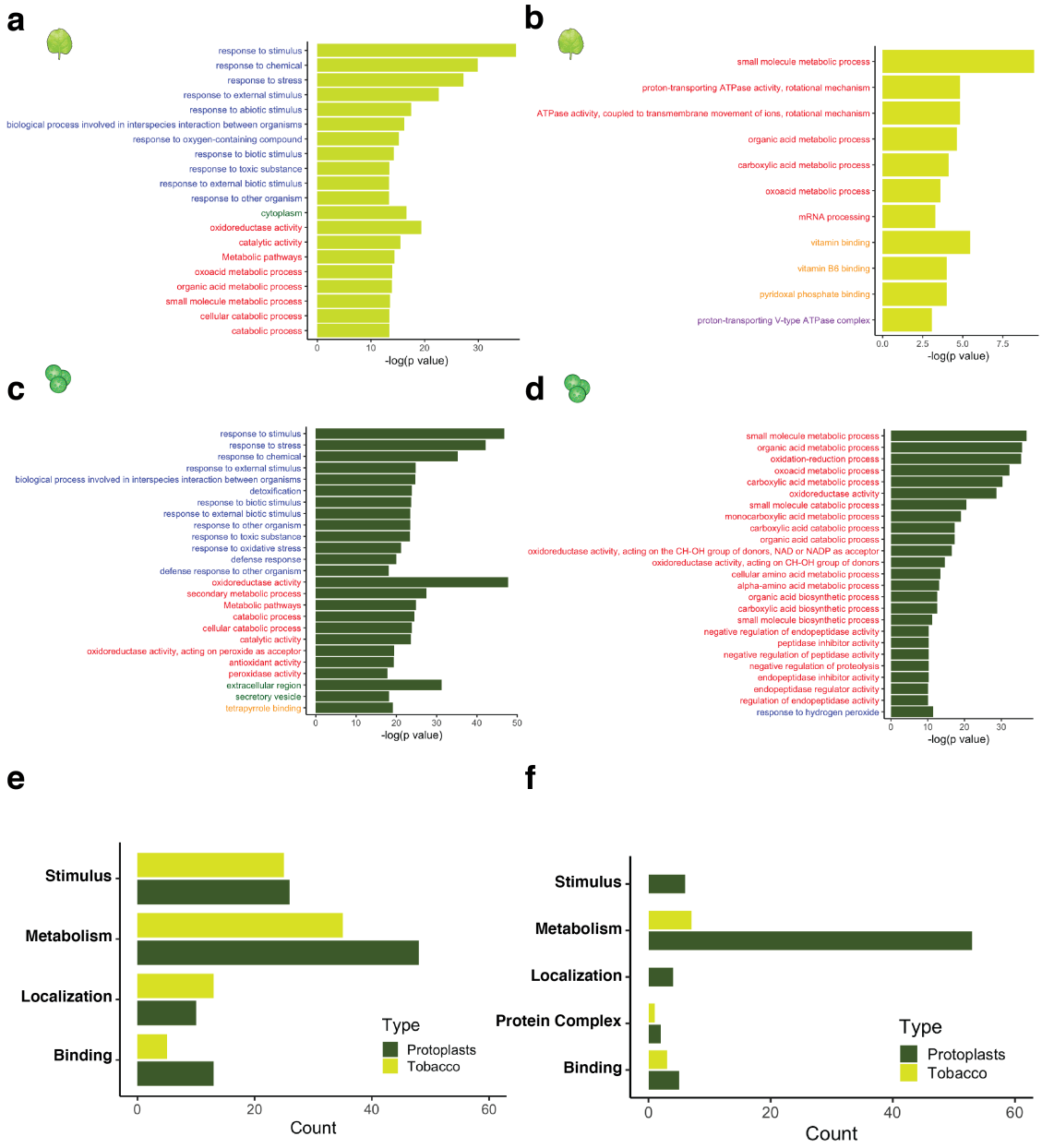
Supplemental Figure 2.2: **Plant STARR-seq yields highly reproducible results across libraries.** a, b Hexbin plots (as defined in Fig. 1) of the correlation between two biological replicates of Plant STARR-seq with the validation library in tobacco leaves (a) or maize protoplasts (b). Commonly used terminators are highlighted in red. c, d Correlation between terminator strength as measured in the large-scale library and the validation library in tobacco leaves (c) or maize protoplasts (d). Pearson's R^2, Spearman's ρ, and number (n) of terminators are indicated in all plots.



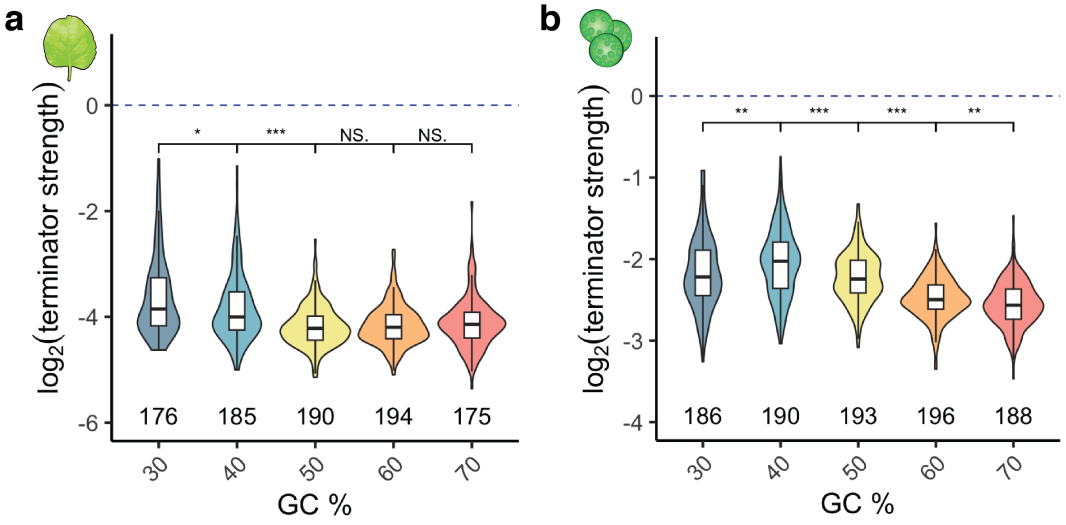
Supplemental Figure 2.3: **Nanoluciferase activity (protein abundance) reflects terminator strength.** Select weak (CDS), intermediate, and strong terminators are cloned immediately downstream of nanoluciferase. The nanoluciferase/luciferase ratio is normalized to a mean of the construct with the 35S terminator per experiment (35S average; log2 set to 0, dashed blue line). a,c Jitter plot of nanoluciferase activity for selected terminators in tobacco leaves (a) and maize protoplasts (c). The gray dot denotes the mean and the gray line denotes the variance. b,d Dot plot and Pearson's R^2 between terminator strength and nanoluciferase activity of tested terminators in (b) tobacco leaves and (d) maize protoplasts. Linear regression line is shown as a blue line, and the gray band around the regression line is the 95% confidence interval. Key: At mid 1= AT1G26300; AT mid 2= AT3G23110; Zm mid = Zm00001d012972; At CDS 1=AT3G22360-CDS; At CDS 2 = AT5G07380-CDS; Zm CDS = Zm00001d025717-CDS; Zm high = Zm00001d047961; HSP17.4 = AT3G46230.



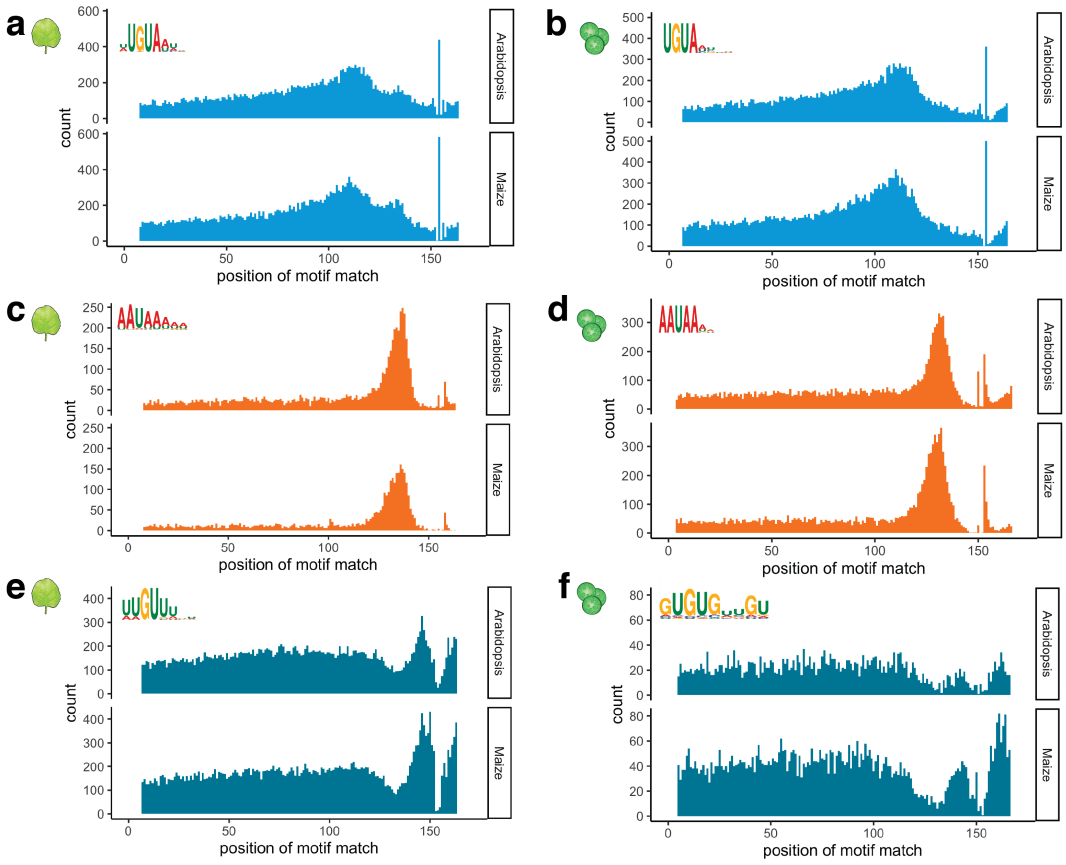
Supplemental Figure 2.4: **Terminator strength is not correlated to gene expression, mRNA half-life, and nascent transcription.** Hexbin plots (as defined in 2.1) of the correlation between the strength of *Arabidopsis* terminators and the expression (a), mRNA half-life (b), or nascent transcription (c) of the corresponding genes. Pearson's R^2 is indicated. See the main text for data sources.



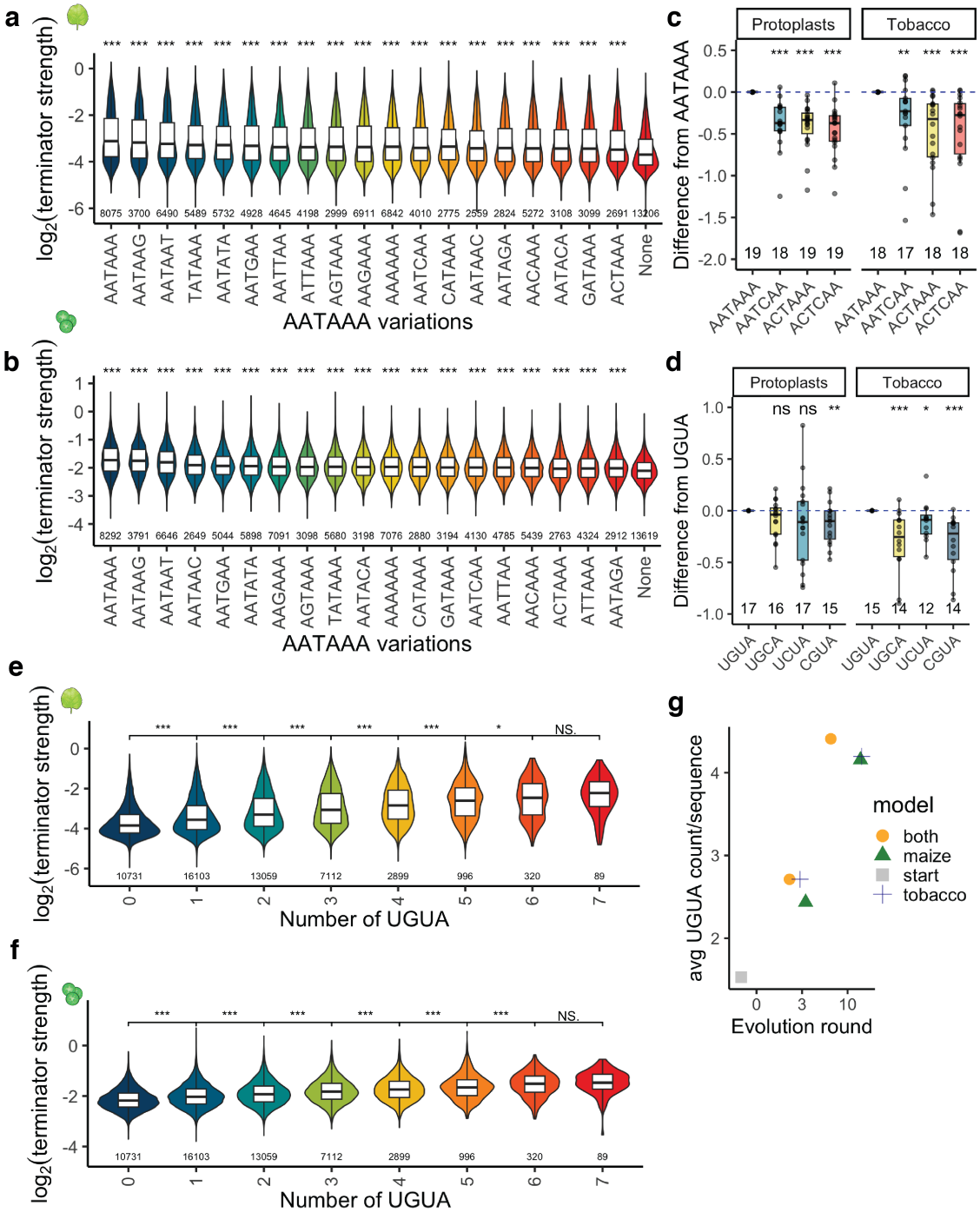
Supplemental Figure 2.5: **Metabolic and stimulus-responsive genes frequently use strong terminators.** a-d GO terms enriched in genes associated with the top 10% of *Arabidopsis* (a, c) or maize (b, d) terminators ranked by strength in tobacco leaves (a, b) or maize protoplasts (c, d). Only the most significant GO terms are shown. The p values were determined using the gprofiler2 library in R with gSCS correction for multiple testing. All enriched GO terms and exact p values are listed in Supplementary Table 6. e, f GO terms for *Arabidopsis* (e) or maize (f) terminators were collapsed into 5 major categories and counted for each assay system.



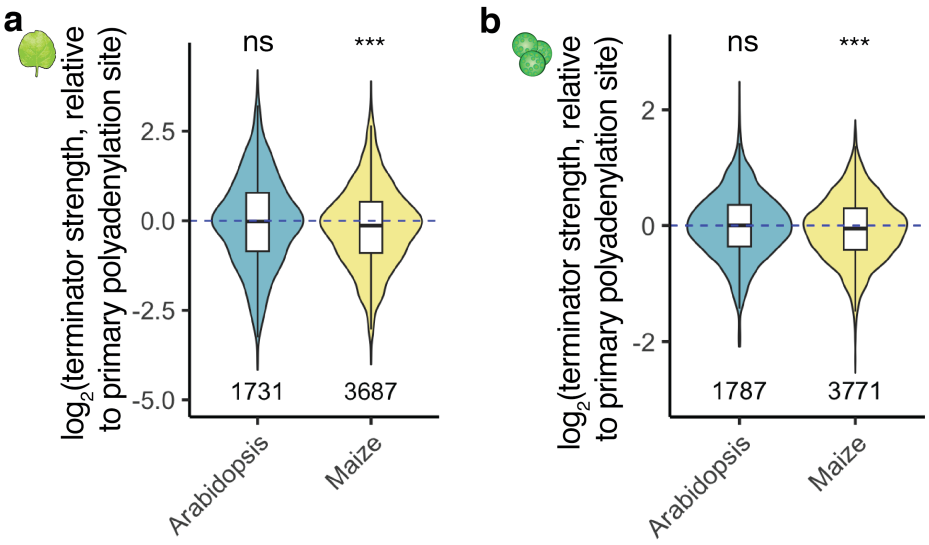
Supplemental Figure 2.6: **Optimal GC content for terminators is species-specific.** Violin plots, box plots, and significance levels (as defined in 2.1) of terminator strength in tobacco leaves (a) or maize protoplasts (b) for randomized sequences with the indicated GC content.



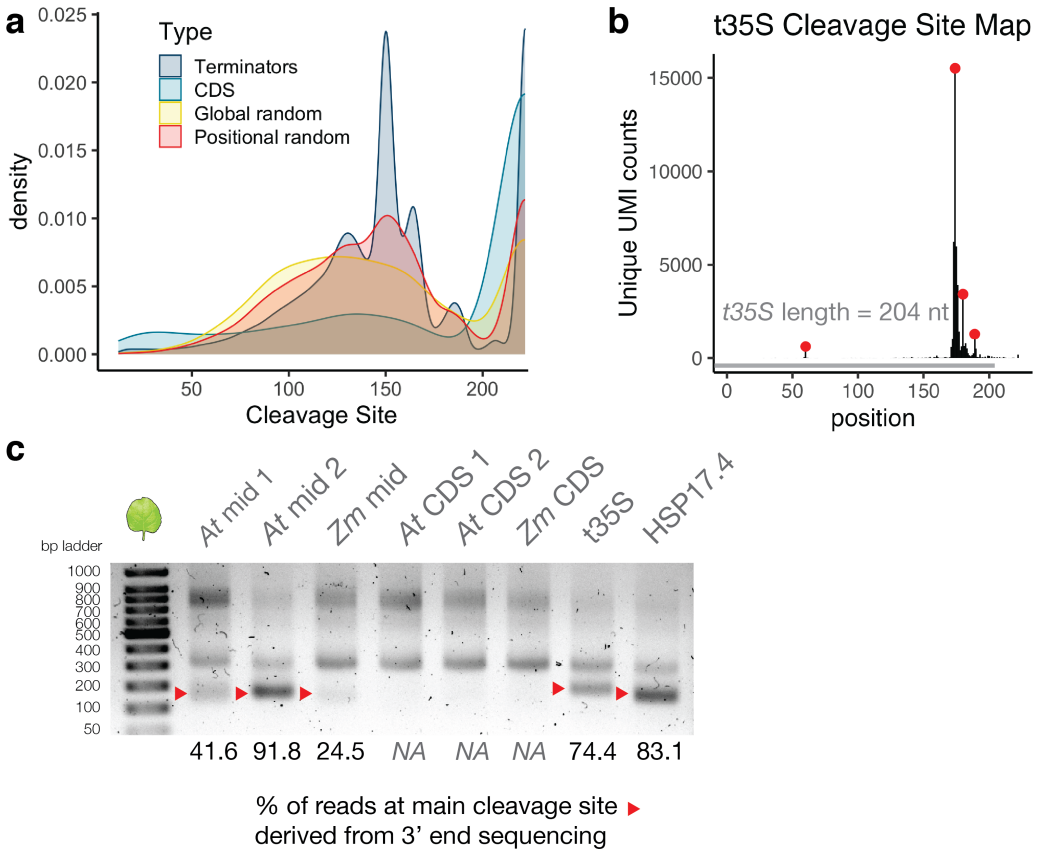
Supplemental Figure 2.7: **Polyadenylation motifs show distinct localization profiles.** Histograms showing the number of *Arabidopsis* and maize terminators with a UGUA motif (a, b), an AAUAAA motif (c, d), or a U/G-rich motif (e, f) at the indicated position. The motifs were discovered in terminators with high strength in tobacco leaves (a, c, e) or maize protoplasts (b, d, f).


a
log2(terminator strength)
AATAAA variations
AATAAA AATAAG AATAAT TATAAA AATATA AATGAA AATTAA ATTAAA AGTAAA AAGAAA AAAAAA AATCAA CATAAA AATAAC AATAGA AACAAA AATACA GATAAA ACTAAA None
b
log2(terminator strength)
AATAAA variations
AATAAA AATAAG AATAAT AATAAC AATGAA AATATA AAGAAA AGTAAA TATAAA AATACA AAAAAA CATAAA GATAAA AATCAA AATTAA AACAAA ACTAAA ATTAAA AATAGA None
c
Protoplasts
Tobacco
Difference from AATAAA
AATAAA AATCAA ACTAAA ACTCAA
d
Protoplasts
Tobacco
Difference from UGUA
UGUA UGCA UCUA CGUA
e
log2(terminator strength)
Number of UGUA
f
log2(terminator strength)
Number of UGUA
g
avg UGUA count/sequence
Evolution round
model
both
maize
start
tobacco

Supplemental Figure 2.8: **Cleavage and polyadenylation motifs are sensitive to mutations.** a, b Violin plots and box plots (as defined in 2.1) of terminator strength in tobacco leaves (a) or maize protoplasts (b) for terminators with the indicated variants of the AAUAAA motif. Terminators without any AAUAAA motif variant (None) are also shown. c, d Boxplots (as defined in 2.4) of the strength of terminators with the indicated variants of the AAUAAA (c) or UGUA (d) motif relative to the strength of the corresponding wild type terminator (set to 0). e-f Violin plots, box plots, and significance levels (as defined in 2.1) of terminator strength in (e) tobacco leaves and (f) maize protoplasts of terminators with varying numbers of UGUA motifs. g Jitter plots of the average number of UGUA per terminator through 0, 3, and 10 rounds of *in silico* evolution.



Supplemental Figure 2.9: **Terminators derived from primary or secondary polyadenylation sites are indistinguishable by terminator strength.** Violin plots and box plots (as defined in 2.1) of the terminator strength in tobacco leaves (a) or maize protoplasts (b) of the experimentally determined secondary polyadenylation site of Arabidopsis and maize genes relative to the primary polyadenylation site of the same gene (set to 0).



Supplemental Figure 2.10: **Cleavage site positions differ between *bona fide* terminators and control sequences.** a Kernel density distribution of all cleavage site positions for plant terminators (Terminators), sequences from coding regions (CDS), and randomized sequences with an overall (Global random) or per-position (Positional random) nucleotide frequency similar to an average *Arabidopsis* or maize terminator. b Cleavage site map of the CaMV 35S terminator (length=204). c RNA was extracted from replicate 1 of the nanoluciferase assay (shown in Supplementary Figure 2.3a) and reverse transcribed using the same oligo(DT) primer for the 3' end sequencing method. cDNA was amplified and run on a 1.0% agarose gel to resolve cleavage. Red triangles denote the site of primary cleavage determined by 3' end sequencing. Key: At mid 1= AT1G26300 ; AT mid 2= AT3G23110; Zm mid = Zm00001d012972; At CDS 1=AT3G22360-CDS, At CDS 2= AT5G07380-CDS; Zm CDS = Zm00001d025717-CDS, HSP17.4 = AT3G46230.

Supplemental Table 2.1: Terminator library composition

Type	Count	Description
Arabidopsis terminators (main cleavage site)	22,204	Sequences surrounding the main polyadenylation and cleavage site of Arabidopsis genes (−150 to +20 relative to cleavage site)
Arabidopsis terminators (secondary cleavage site)	2,325	Sequences surrounding the secondary polyadenylation and cleavage site (30% of total reads) of Arabidopsis genes (−150 to +20 relative to cleavage site)
Maize terminators (main cleavage site)	25,685	Sequences surrounding the main polyadenylation and cleavage site of maize genes (−150 to +20 relative to cleavage site)
Maize terminators (secondary cleavage site)	4,407	Sequences surrounding the secondary polyadenylation and cleavage site (30% of total reads) of maize genes (−150 to +20 relative to cleavage site)
CDS	1,178	Coding sequences (170 bp) with similar overall GC content as Arabidopsis or maize terminators (589 from each species)
Fixed GC content	1,000	Randomized sequences (170 bp) with a GC content of: 30%, 40%, 50%, 60%, or 70% (200 each)
Global random sequences	400	Randomized sequences (170 bp) with overall nucleotide frequency similar to an average Arabidopsis or maize terminator (200 per species)
Positional random terminators	2,000	Randomized sequences (170 bp) with per-position nucleotide frequency similar to an average Arabidopsis or maize terminator (1,000 per species)
Commonly used terminators	4	35S, Ag7, NOS, and MAS terminators

Chapter 3

FUTURE DIRECTIONS AND DISCUSSION

3.1 Filling in the gaps

In an old Indian parable, a group of blind men come across an elephant, whom they have never seen, and have to imagine what an elephant is by touching it. The story has many forms, but the simplest version goes like so:

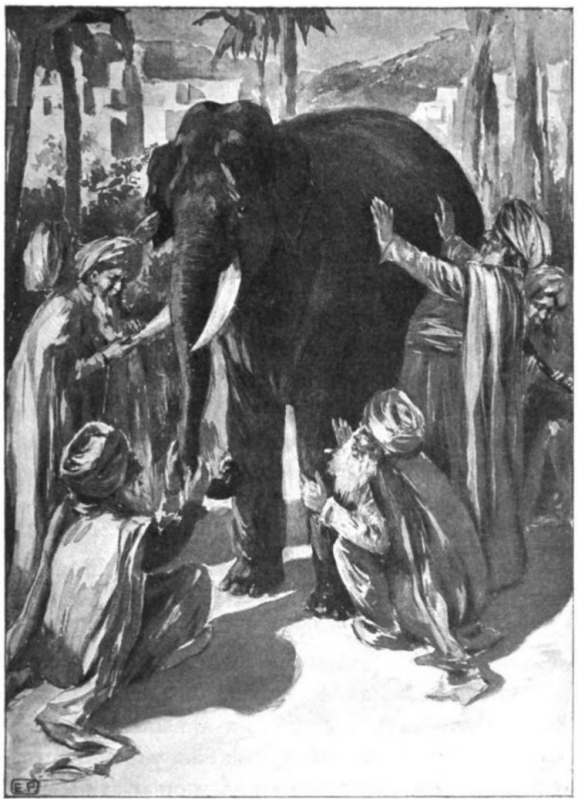


Figure 3.1: **The Blind Men and the Elephant** Illustrator unknown. From *The Heath Readers by Grades*, D.C. Heath and Company (Boston) pg. 69.

Six blind men are brought to examine an elephant that has come to their village.
The first man touches the trunk and says that the elephant is like a thick snake.
The second man touches the tusk and says that the elephant is like a spear.
The third man touches the ear and says that the elephant is like a fan.
The fourth man touches the leg and says that the elephant is like a tree.
The fifth man touches the side and says the elephant is like a wall.
The sixth man touches the tail and says the elephant is like a rope.
Each of the blind men is convinced that he is right, and that everyone else is wrong.

Of course, each was partly in the right, and all were in the wrong. The parable of the blind men is more than the Gestalt theory that "the whole is something else than the sum of its parts," which I will get to later. When we talk about data, we must accept that every standard of measurement is a proxy for reality. Picking the best proxies to shed light on what you care about is an art, not a science. Every metric you come up with will have shortcomings in conveying the complete truth. Even in conducting massively parallel reporter assays, what we learn is defined entirely by what we measure. In measuring the effects of plant terminators on transgene expression, for example, I did not capture all the binding sites for RNA binding proteins and microRNA target sites that might have influenced the output. I did not measure protein abundance. I barely scratched the surface of how terminators might affect localization within the cytoplasm. Yet I know from literature research that all of these features play a significant role in terminator processing or stability and thus gene expression and protein abundance.

But what if we had infinite blind men touch all individual parts of the elephant, would they come closer to the truth? What does it take for blind men to see the whole elephant? Trying to understand what we cannot see is a regular Tuesday in molecular biology. If the ultimate goal for plant transgenics is to predict the expression and production of any gene introduced into plants, then we need more massively parallel reporter assays that functionally characterize all parts of the plant transgene on gene expression and protein production. In this next section, I will discuss a few ways to glean more information about

the mRNA stability of each transgene tested. I will discuss two new variations of the plant STARR-seq assay that will uncover more about the regulatory grammar of the 5' UTR and intronic sequences and how they might impact expression. Finally, I will go back to the elephant in the room. How can one take all the information learned from MPRAs and "omic" studies and synthesize them into a complete mechanistic model of gene expression, or better, a complete and accurate *in silico* simulation of the plant post-transcriptional machinery?

3.1.1 Reanalyzing terminator assays for mRNA differential degradation

Poly(A)-tails play a critical role in multiple aspects of the transcript's life. Poly(A) tails mediate mRNA translocation, translation efficiency, and mRNA quality control and degradation [47, 84, 167, 225]. When mRNA is no longer used for translation, poly(A) tail shortening is a key step initiating degradation. The length of the poly(A) tail is dynamically regulated by poly(A) polymerase and deadenylase, such that shortening poly(A) tails to a certain threshold releases poly(A)-binding protein (PABP) and triggers decay. Next-generation-based sequencing methods like PAL-seq (poly(A)-tail length profiling by sequencing), TAIL-seq, mTAIL-seq, PAT-seq (Poly(A)-Test RNA-sequencing), and TED-seq (tail-end displacement sequencing) have all been used to characterize poly(A) tails [42, 107, 179, 294, 338]. Long read platforms like PacBio and Nanopore have led to techniques that detect full length mRNA and poly(A) tail information, including FLAM-seq, PAIso-seq, Nanopore direct RNA sequencing, and FLEP-seq [171, 185, 187, 234]. Recently, scientists adapted FLEP-seq (full-length elongating and polyadenylated RNA sequencing) to sequence the full length mRNA and poly(A) tail in plants, analyzing 120 million polyadenylated mRNAs from seven different *Arabidopsis* tissues and shoot tissue from maize, soybean and rice [139]. The study reports that poly(A) tail lengths are different among different genes but are highly correlated among different tissues for the same gene. In other words, poly(A) tail length is regulated in a gene-specific manner. Additionally, mRNAs with short half-lives have longer poly(A) tails, while mRNA with long-half-lives have relatively short poly(A) tails. These long-read sequencing results beget testing how the length of the poly(A) for

each terminator might correlate with measured terminator strength in both tobacco leaves and maize protoplasts. While long-read sequencing would be preferred (to get both the polyA tail and the entire terminator), sequencing poly(A) tails can be done with a cheaper short-read approach, like TAIL-seq, especially since we have already determined cleavage sites with 3' end sequencing. Interestingly, Jia et al. [139] also found that compared to *Arabidopsis*, monocots like maize and soybean have fewer transcripts with extremely short tails but had more transcripts with longer tails. Comparing TAIL-seq results from both tobacco and maize protoplast terminator experiments will elucidate species-specific terminator strength differences as a function of poly(A) length. I hypothesize that stronger terminators have longer poly(A) tails in both species. But since the studies have shown poly(A) tail length might be regulated in a gene-specific manner, I predict that terminators from the same gene families will have similar length profiles.

3.1.2 Incorporate RNA structural insight into terminator screens

One great challenge in RNA biology has been to accurately describe mRNA structures–especially of the noncoding portion of the mRNA–that modulate function. Free from the constraint to encode proteins, UTRs can form considerable Watson–Crick and non-canonical base pairing that can impact mRNA regulation. Accurately describing RNA secondary structures will uncover mechanistic insight about post-transcriptional regulation. Determining RNA tertiary structure using experimental techniques, like NMR and X-ray crystallography, poses significant challenges due to the high cost and resolution limits on RNA measurement. Cryo-EM techniques have improved resolution relative to X-ray and NMR, but still face serious limitations [146]. The most popular approach for predicting RNA structure is based on thermodynamic models, in which a secondary structure is decomposed into several substructures, called nearest neighbor loops (i.e. hairpin loops, internal loops, bulge loops, base-pair stackings, multi-branch loops, and external loops). Adding up the free energy parameters characterizing each loop calculates the free energy of each nearest-neighbor loop, and summing the free energy of the decomposed nearest neighbor loops determines the overall free energy of the RNA structure. The optimal RNA secondary structure is the one

that has the least free energy. A number of tools like Mfold/UNAfold [197, 375], RNAfold [117, 188], and RNAstructure [254] use this thermodynamic approach, although there are serious experimental constraints on testing the accuracy of the predictions **(Figure 3.2)**.

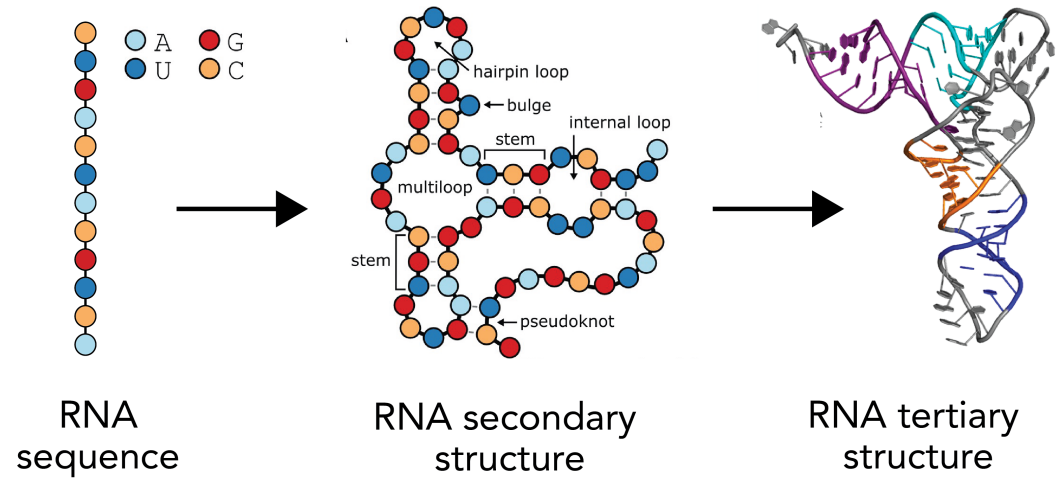


Figure 3.2: **How RNAs fold** RNA structure An example of RNA primary (left), secondary (middle), and tertiary structures (right). The RNA folding process is hierarchical: The RNA secondary structure forms rapidly from linear RNA (primary structure) due to hydrogen bonding between complementary bases on the same strand. An RNA secondary structure can be decomposed into several types of nearest-neighbor loops. The formation of a complex tertiary structure is usually much slower. RNA can adopt a variety of tertiary structures due to the enormous rotational freedom in the backbone of its non-base paired regions.

RNA structure is not as rigid as that of proteins and there are nearly infinite parameters of motion to consider. Still, deep learning techniques that are trained on some reference RNA structures have also been developed, such as SPOT-RNA [289], E2Efold [50], and MXfold2 [269]. E2fold and SPOT-RNA formulate RNA secondary structure prediction as multiple binary classification problems that predict whether each pair of nucleotides forms a base pair or not. MXfold2 integrates thermodynamic parameters on top of base pair calculations for a more robust prediction. I propose testing the predicted mRNA structure from the 3' end sequencing results across these tools. If we employ TAIL-seq, we can include in the prediction the entire poly(A) tail, which is known to base-pair with U-rich

regions elsewhere on the mRNA, and thus improve the accuracy of free energy predictions. Another alternative is to use Shape-seq (2'-hydroxyl acylation analyzed by primer extension sequencing), which works by modifying the 2'-OH of less structured RNA nucleotides so that reverse transcription is halted one nucleotide before a modification. NGS of the resulting cDNA fragments determines the location and frequency of the modifications across each RNA. A maximum-likelihood estimation strategy is then used to infer structure. The crux of the assay falls on the idea that the more modified the RNA, the more unstructured [189, 332].

3.1.3 MPRA for studying the effect of introns on transgene expression

In both *Arabidopsis* and rice, about 80% of the coding regions of these genes contain introns, with about 4 introns per gene on average [217]. Genome-wide analysis of mRNA decay rates in *Arabidopsis* found that genes possessing at least one intron produce mRNA transcripts significantly more stable than those without introns, and this was not due to overall length, sequence composition, or number of introns [224]. Studies have collectively shown that introns increase mRNA accumulation, and thus protein expression [88], but there are differences in the splicing machinery within the plant kingdom. In conventional genetic studies, the effect of an intron is usually quantified by dividing the amount of mRNA accumulation or reporter produced from an intron-containing gene by the amount made by an intron-less control [38]. When the first intron of the castor bean catalase gene (cat1) was placed inside the coding sequence of the beta-glucuronidase gene (gusA) and expressed in transgenic rice (monocot) calli/tissues and transgenic tobacco (dicot), the intron-containing version of gusA had 10-to-40 fold and 80-to-90 fold more expression compared to the intronless gusA in transgenic rice protoplasts and transgenic rice tissues, respectively. However, the presence of the intron made no difference in gene expression in tobacco leaves [303], due to errant splicing of the intron. Similar to findings in rice, introns increase gene expression in cultured maize cells [39]. Dissecting the monocot versus dicot differences in splicing mechanics will improve our ability to express species-specific transgenes that can avoid degradation.

Nucleotide composition of introns is critical for function. Plant introns harbor AU-rich

sequences and are on average 10-15% more AU rich than exons [97]. Monocot introns are on average 63% AU-rich, but about 20% of them have higher GC content (50%), while dicot introns are 67% AU-rich and less tolerant to variations. Many monocot introns, such as the maize Adh1 intron1 (57% AU-rich) are not efficiently spliced in dicot cells [98, 149]. Adding stretches of Ts can lead to efficient splicing of artificial and poorly spliced GC-rich introns in transfected tobacco protoplasts [96]. Similarly, a synthetic intron, 75% AU-rich with canonical 5' and 3' splice sites and a branchpoint consensus sequences, is efficiently spliced in tobacco protoplasts, but cannot tolerate a GC-rich sequence insertion that reduces the AU content below 59% [97]. The large impact nucleotide composition plays in intron splicing will make learning species-specific patterns much easier.

Despite the genome-wide and transcriptome-wide sequencing efforts to characterize plant intron space, what is still left unexplored is a massively parallel reporter assay measuring the *cis*-regulatory effect of introns on plant gene expression and splicing efficiency. One can measure the efficiency of intron splicing of individual variants in a large library by sequencing at splice junctions while also capturing the strength of each construct through mRNA accumulation. While vertebrates have introns spanning hundreds of thousands of base pairs, plant introns span from 1 to 1.5 kb, with a maximum of just 3 kilobases. Given the smaller space of plant introns, it's surprising there is still no functional characterization of introns for deep learning. However, testing native sequences in parallel is laborious and inefficient. Since it's been shown that the nucleotide composition and distribution of introns plays a huge role in splicing efficiency and ultimately expression, I propose testing a large library of short (either N50, N100, or N150) randomized sequences of varying GC content inside an intron of the GFP reporter. Experimentally testing the consequence of every possible genetic variant on endogenous alternative splicing is impractical, but we can develop predictive models of the splicing code by keeping the 3' and 5' splice sites and branch point, while varying the other nucleotides [238].

In transient tobacco leaves, exogenous transgenes containing introns were less vulnerable to post-transcriptional gene silencing [59]. When two introns of RBCS1A were added to a GFP reporter (making GiFiP), transient transgene expression improved in tobacco leaves [78]. In the pilot experiment testing a small terminator library with different enhancers, I

also tested how changing the GFP to GiFiP would affect the terminator strength. Having an intron should makes no difference when the reporter is driven by a a strong enhancer. Enhancers increase transcription but should have little effect on intron processing. However, when no enhancer or a weak enhancer drives expression, the stabilizing effect of the intron-containing GFP is significant **(Figure 3.3**.

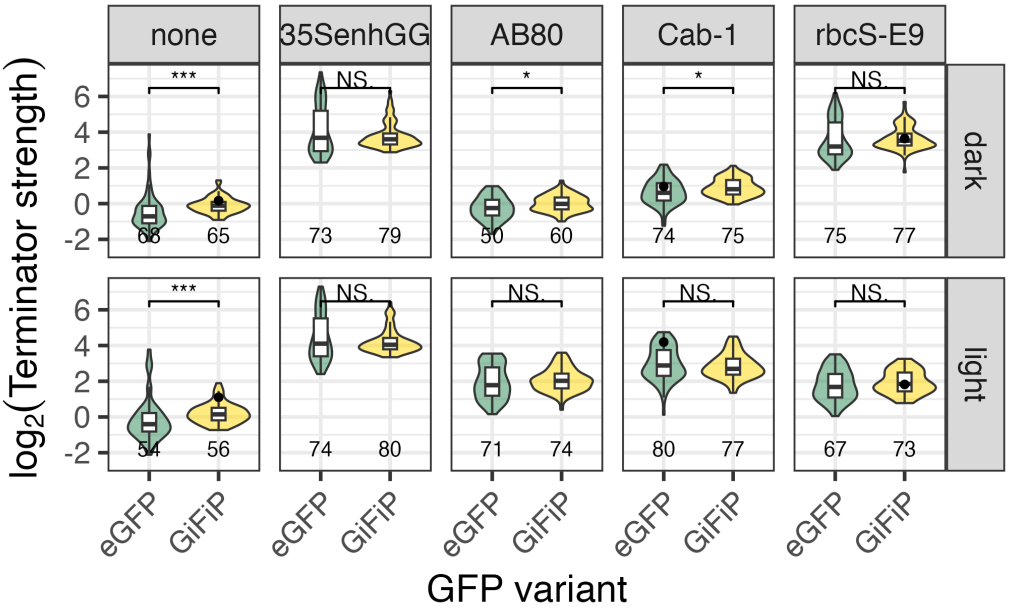


Figure 3.3: **Introns moderately increase transgene stability** A small terminator library (n=102) was cloned downstream of either a barcoded GPF or a GFP with two introns from RBCS1A (GiFiP). Each library was tested either with no enhancer, the CaMV 35S enhancer, the AB80 enhancer, Cab-1 enhancer, or the rbcS-E9 enhancer. Violin plots, box plots, and significant levels are described in 2.1.

I propose modifying Sort-seq, a splicing MPRA developed in mammalian cells, to test an intronic library for exon skipping in GiFiP. In the original Sort-seq, a red fluorophore (mCherry) is constitutively expressed, and a three-exon, two-intron minigene construct is cloned into a plasmid such that if the middle (tested) exon is skipped, a GFP protein is also expressed with the mCherry. One can input the variable intronic sequences on either side of the middle exon, keeping the native RBCS intron on the other side or place variable sequences on both sides. Cells can be then sorted into bins using GFP:mCherry ratios by

fluorescence-activated cell sorting (FACS), where a higher ratio indicates a greater intron excision. One issue with adapting this protocol for FACS using plants is that protoplasts do not take in one single construct at a time. A sequencing based readout to determine if the exon was skipped will bypass this issue and is applicable to both tobacco leaves and maize protoplasts systems [52]. A sequencing-based readout will determine how the variable sequences impact exon skipping and transgene expression and where aberrant splicing occurrs.

A loftier goal for studying intronic grammar is to be able to predict alternative splicing patterns based on environmental stress. In plants, alternative splicing is one the primary mechanisms of proteome diversity, generating numerous protein isoforms from a single gene. It is the major mechanism involved in plant stress and environmental response by allowing rapid adjustment of the abundance and function of key stress-response components. If splicing is repressed, the plant undergoes immediate abiotic stress [249]. High temperatures significantly impact the splicing profile of many genes, demonstrating how integral alternative splicing is to heat stress response [260]. Thus, teasing out the grammar of alternative splicing will unlock new control switches for fine tuning a desired phenotypic response to combat climate change.

Alternative splicing (AS) also generates alternate functional mRNAs isoforms with differences in subcellular localization, stability, or function by changing or completely removing functional domains, via the introduction of premature termination codons (PTC), intron retention, or alternative 3' or 5' splice site selection [299]. mRNA subcellular localization controls gene expression both spatially (by transporting mRNA into different subcellular structures) and quantitatively (by controlling the accessibility of mRNA to ribosomes) [177]. Exploring how introns govern subcellular localization will greatly benefit transgenic studies. Imagine an AS plant transgenic method where, if one knows the grammar of tissue specific intron splicing, one can introduce designed tissue-specific splice sites that allow the transcript to produce two different protein isomers in different tissues. Tissue specific splicing engineering would be critical for climate change research as many stress response genes in plants change splicing patterns under stress.

I propose that testing the same GiFiP library of introns under different stresses, like

heat, drought, and pathogen, will uncover environmental specific AS *cis*-regulatory signals. To find patterns of tissue specificity, we can test the same library for splicing efficiency in different tissues, like roots and leaf protoplasts. Exploring the grammar of alternative splicing holds the greatest potential in spatial and temporal expression of transgenes in plants.

3.1.4 MPRA for studying effects of 5' UTR on plant translation efficiency

The 5' UTR sequence plays a major role in translation efficiency and is a high priority target for plant transgenic technology. What good is a highly transcribed mRNA if it's poorly translated? To initiate translation, the ribosomal 43S preinitiation complex (PIC) scans the 5' UTR in the 5'-to-3' direction until a start codon is found. The 5' UTR can affect translation by capturing PICs prematurely with an upstream start codon (uAUGs) and ORFs (uORFs), which interfere with PIC scanning or directly recruit ribosomes via Internal Ribosome Entry Sites (IRESs) [113]. uORFs in the 5' UTR can also repress translation from the downstream ORF by stalling ribosome movement. In *Arabidopsis*, roughly 37% of genes have at least one uORF [319], of which 187 genes get translated [119]. Roughly 29% of maize genes have uORFs, where variation in uORFs contributes to allelic diversity in maize protein abundance [87]. In maize, uORF translation is globally enhanced under drought stress [172].

In 2018, it was shown that knocking out an endogenous uORF is an efficient and tunable method for upregulating protein expression [285, 349, 364]. To follow suit, scientists also showed that generation of *de novo* uORFs can tunably repress protein expression [351]. By using base editing or prime editing to generate *de novo* uORFs or to extend existing uORFs by mutating their stop codons, they generated a suite of uORFs that incrementally downregulate the translation of primary open reading frames (pORFs) to 2.5–84.9% of the WT level. The power of uORFs in modulating protein production without changing coding sequences highlights the need to explore the 5' UTR sequence space for optimal translation.

Whole genome sequencing of a large number of organisms including *Arabidopsis* revealed that the 5' UTRs have great diversity in their nucleotide sequence [222]. *Arabidopsis* has

an average 5' UTR length of 131, and studies have shown that the sequences with the greatest effect on translation lie immediately upstream of the AUG start site. A small scale assay testing 25 *Arabidopsis* 5' UTRs for translation efficiency found a 200-fold difference among them [153]. 5' UTR grammar differs in dicots and monocots, indicating a need to test the library across both types of angiosperms [353]. In dicots, the 5' UTR of certain dicot mRNAs act as translational enhancers, but they often offer no enhancement when expressed in monocots. Since most of the major food crops are monocots, dissecting the species-specific differences would behoove crop engineering.

Just as MPRAs can be used to study the function of 3' UTR sequences, they can be used to study the effect of sequence variation on the 5' UTR. In a mammalian 5' UTR massively parallel reporter assay, Sample et al. [264] measured the impact of randomized 5' UTR sequences on ribosome loading by having putative 5' UTR sequences inserted upstream of a GFP coding sequence [264]. After introducing the 5' UTR reporter library in cells, the 5' UTR sequences that are actively translated on ribosomes are directly sequenced after polysome profiling. Earlier MPRAs designed for 5' UTR learning relied on FACS sorting or growth selection to profile activity [56, 160], which are not that amenable to plant based MPRAs like STARR-seq. Polysome profiling is used to study the association of mRNAs with the ribosome, providing an assay to test how well 5' UTR sequences initiate translation. A major limitation to a plant 5' UTR screen is the efficiency of plant polysome profiling on a library scale. However, extensive optimization work has been done to improve polysome and ribosome profiling in plant tissue [169, 339]

I propose adapting the mammalian 5' UTR MPRA with polysome profiling for plants. After 5' UTR library transfection, plant lysates are prepared and centrifuged through a sucrose gradient, with larger, heavier complexes containing multiple ribosomes (polysomes) traveling further down the gradient than smaller, lighter complexes with single ribosomes (monosomes). When paired with qRT-PCR and next-generation sequencing, the identity of the mRNA variant present in different fractions can be determined. For a given 5' UTR, the relative counts per fraction are multiplied by the number of ribosomes associated with each fraction and then summed to obtain a measured Mean Ribosome Load (MRL), or 5' UTR strength [204]. A limitation of this approach, however, is that polysome profiling is

expensive and requires a special ultracentrifuge. Since there are labs that specialize in plant polysome profiling [119, 191], a collaboration would expedite the research.

The ultimate goal is to train a deep learning model that predicts species-specific 5' UTR strength [56, 105, 264]. The highly diverse but relatively short sequence space of 5' UTRs, coupled with a large dynamic range of activity seen in plants, is promising for training a deep learning model of 5' UTR grammar in plants. The data and model here will enable the quantitative assessment of secondary structure, uAUGs and uORFs, and other *cis*-regulatory sequence elements in the 5' UTR plant space. Despite all the plethora of DL models that predict transcription, accurately predicting translation will be the most salient to crop engineering.

3.1.5 MPRA to study m^6A RNA methylation in plants

While not discussed in the introduction, RNA methylation is an important post-transcriptional modification that influences gene regulation. In plants, over 200 different RNA modifications have been discovered, playing key roles in plant development processes like embryo development, shoot stem cell fate, floral transition, trichome morphogenesis, leaf initiation, and root development [279]. The most common and abundant RNA methylation in plants is N6-methyladenosine (m^6A). *In Arabidopsis*, mutations in writers, erasers, and RNA methylation readers have huge impacts on phenotype. Methylated RNA immunoprecipitation coupled with next-generation sequencing has allowed the transcriptome-wide RNA methylation profiles of *Arabidopsis*, rice, *Brassica*, and maize [282]. An earlier study on maize in 1980 discovered that most of the m^6A loci are present in the poly-A tail of mRNAs, indicating that m6A methylation might play a role in RNA stability [228]). In 2014, a study on the *Arabidopsis* m^6A methylome found a high enrichment of methylation around the stop codon, within the 3' UTR, and around the start codon [192]. Long read nanopore RNA sequencing of *Arabodopsis* also found that the loss of m^6A from 3' UTRs lead to decreased transcript accumulation and defective 3' end formation [233]. Sequencing-based methods like MeRIP-seq (Methylated RNA immunoprecipitation sequencing) have been the most successful to detect RNA methylation at a transcriptome-wide in *Arabidopsis*, rice,

Brassica, and maize [55, 120, 183, 208]. Yet these strategies require high input (over 500 µg) of total RNA, as m^6A is generally less than 0.1% of the total RNA. Despite the low levels of overall m^6A, scientists developed a massively parallel reporter assay for m6A called MPm^6A [108], in which thousands of endogenously-methylated m^6A sites along with 102 nucleotides (nt) of sequence surrounding each site were synthesized and cloned into the 3' UTR of a plasmid-based, intronless GFP transgene. In mammalian cells, the sequences were m^6A methylated through transfection into cells. The methylation status of each individual sequence was assessed by its enrichment after m^6A –immunoprecipitation (IP) of mRNA and then sequenced by NGS to determine enriched m^6A sequences. I hypothesize that this method can be adapted to analyze terminator plant-STARR seq for methylation and how it might impact terminator strength, although the extent of endogenous m^6A methylation might make it harder to have a large dynamic range for analysis.

3.2 One model to rule them all, one model to bind them

In the not too distant future, the world of synthetic biology will be a lot like the generative AI models ChatGPT and DALL-E3. If the user inputs an entirely new synthetic genome sequence, the model will not only generate an *in silico* dynamic model of the cell the genome encodes, but also predict the phenotypic effect of any variation in the genome the user wants to test on the model. The phenotypic effect of any transgene or gene variant will be accurately predicted. No one would need to touch a pipette again!

It sounds like science fiction, in that it is very challenging to achieve, but it is not impossible. Creating accurate *in silico* gene expression models would require a deep understanding of all the genes, their products, and their roles in cellular processes [200]. One day we will reach a nexus of understanding when all this is possible, but until then, we have to rethink how we incorporate what we learn from multivariate experiments and models to better recapitulate the whole system. As Einstein coined, "Everything should be made as simple as possible, but not simpler."

3.2.1 Single order MPRAs fail to capture higher order interactions

There are many reasons for the lag behind "GeneGPT." First, we are limited by data volume acquisition. Unlike fields like computer vision and natural language processing in which one could easily collect terabytes of data, biological data have to be generated from biological experiments. If a particular mechanism cannot be studied by established experimental techniques in order to generate a large enough dataset for training, it will be impossible to use a machine learning model to "learn" anything from it. The biological and technical variations across experimental conditions limits a model's generalization performance. In the last twenty years, there has been significant progress in generating larger scale and broader biological experiments across many samples, but the data are only a small percentage of what could be measured.

The multiple layers of gene regulation do not happen independently. Changes in the protein abundance of one protein may positively or negatively impact the transcription of another. Even for the multi-task models that predict multiple genomic features simultaneously, the interactions between those predicted events are not explicitly taken into consideration. Studies in plants have shown that some enhancers and promoters interact with intrinsic specificity [288], yet their effect is driven through specific proteins and cofactors that we cannot identify with plant STARR-seq. It is expensive to combinatorially test every promoter enhancer pair in plants, but one might learn specific enhancer-promoter interactions on a small scale of well chosen candidates. In an effort to find promoters with sequence-encoded preferences for certain enhancers in humans, Bergman et al [27] designed a high throughput assay called enhancer x promoter self-transcribing active regulatory region (ExP STARR-seq) and applied it to examine the combinatorial compatibilities of 1000 enhancers and 1000 promoters in human cell lines. They found that most enhancers activate all promoters by a similar amount, but intrinsic enhancers and promoter activities multiplicatively combine to determine RNA output. Plant ExP-STARR-seq would be ideal for studying *cis*-regualtory interactions without having to measure the proteins and cofactors involved. The results of this ExP-STARR-seq experiment compared to results of the individual enhancer and promoter STARR-seq experiments will add insight into synergistic

effects.

The practice of drawing mRNA as a straight line belies its structure. We know that 3' UTR and 5' UTR regions of mRNA interact to modulate degradation and translation. There is ample evidence that supports the existence and importance of 5'-3' communications in determining the fate of the mRNA, whether through closed loop conformations or protein-mediated crosstalk [318]. For example, changes in poly(A)-tail length are sufficient to change 5' end activities in eukaryotes [219] by triggering 5' decapping. Everything in biology is in motion and must be viewed through the lens of dynamics and kinetics. The inherent flexibility of the RNA backbone, the power of base stacking, and multiple possible tertiary interactions make RNA an inherently structured molecule that may not need to rely on external factors to achieve 5'-3' interactions. Unfortunately, it is challenging to test if the 5' and 3' regions interact in a massively parallel reporter assay since it is unclear how the interactions occurs and for how long, especially given the effects of molecular crowding, localization of mRNAs in subcellular compartments, and the unknowns of cellular organization [318]. RNA structural prediction algorithms can give us an estimate of how likely 5' and 3' regions might interact in our assay, but they are not dynamic predictions. Developing a quantitative framework for the relationship between gene regulation and UTR proximity might require single-molecular resolution to correlate a specific measurement of proximity of individual mRNAs to its structure.

No plant deep learning model accounts for the effect of terminator-promoter interactions in plants. The NOS terminator is a great case in point. It behaves as a weak terminator when paired with UBQ10 and lexA promoters, a strong terminator when paired with the NDUFA8 promoter, and a medium strength terminator when paired with alcSynth and pOp6 promoters [10]. The synergistic regulation between promoters and terminators can be also explained by the direct interaction of the terminator with the promoter (via chromatin looping).

It's critical to recognize that gene regulation is a holistic multifactorial process with many moving parts. There needs to be a push for high-throughput screens that uncover more about the dynamic interactions among genetic elements. The biochemical and physical interactions of the system should not be ignored either. Future deep-learning models for

gene regulation should not only incorporate various multi-omics data sources as inputs but also account for the connections between these multi-omics factors in their outputs.

3.2.2 Building predictive models of gene expression

Accurately predicting gene expression is a critical task for plant transgenics. Currently, no sequence-based model is capable of holistically accounting for all stages of gene expression from transcription initiation to protein degradation and can accurately predict the abundance of each processed protein isoform in any given cellular context. Predicting expression on a few elements is the same as asking two instead of six blind men to describe the elephant.

Promoter strengths as measured by STARR-seq did not correlate with expression data from *Arabidopsis*, maize, or sorghum [142]. Terminator strength did not correlate with expression, mRNA half-life, or nascent transcription. Nevertheless, I wondered how well the combined strengths of each regulatory component of the gene (accessibility, promoter activity, terminator activity) might inform the overall expression of that gene (agnostic of the gene's sequence and protein function). I lack information on how the 5' UTR and specific introns affect each gene's expression. I also lack measurements of transcription factor binding or histone modifications, both or which are known to significantly impact gene expression. Enhancer STARR-seq (unpublished data) assays the strength of accessible sites (derived from ATAC-seq profiles) in driving expression of a barcoded reporter gene in maize protoplasts and tobacco leaves. Since there are multiple peaks per gene, it is challenging to associate which promoter the accessible sites modulate and how each site affects another. For a baseline heuristic, I chose the closest accessible site with the strongest score for each gene. Taken together, all these limitations will inevitably deter accurate prediction of expression, but how much of the variance I can explain with what is already measured will provide a ground truth.

For genes in *Arabidopsis* and maize that had all values (no gene with any missing value was included), I correlated measured terminator strength, promoter strength, and enhancer strength to gene expression derived from *Arabidopsis* leaf tissue and maize leaf tissue **(Figure 3.4)**. In maize, most components had a weak positive correlation to bulk expression

in leaf and kern tissue. Predictably, the promoter strength from maize protoplasts had the highest correlation to maize leaf expression in B73 (the same cultivar used to make protoplasts). In *Arabidopsis*, promoter strength was slightly negatively correlated with expression, regardless of which promoter variation was tested (in light/dark, with/without 35S enhancer, tested in tobacco/maize). The one caveat with predicting expression for *Arabidopsis*, however, was that that all assays were done in tobacco leaves and lack the species-specific context of *Arabidopsis*. A gene's terminator strength had little correlation to its expression or promoter strength or enhancer strength, as one would expect since terminator strenth has a greater role in translation efficiency than transcription. The strength of the closest ACRs to the TSS start site had a slight correlation with the strength of the promoter testing in tobacco (with no 35S enhancer). This might be an artifact however, since the promoters were chosen as bases -165 to $+5$ relative to the TSS and could very well overlap with the accessible peaks nearest to the TSS.

Despite the low Pearson correlations across the board, I applied a multiple linear regression model to the data in order to estimate the relationships between the multiple STARR-seq assay scores and gene expression. The data could explain only 9% ($R^2 = 0.08985$) of expression in maize tissue and 3% ($R^2 = 0.03291$) of the expression in *Arabidopsis* tissue. However, the model might not accurately capture the true relationship if important variables were missing (like 5' UTR strength perhaps). Also, a linear model might not be the best choice since I know that the variables are not independent from each other. Multicollinearity can make it challenging to interpret the individual contributions of each variable to expression. A support vector machine, which is better at capturing non-linear relationships, did not have greater predictive power than the linear model, however, with an $R^2 = 0.073$.

The little correlation among the tested elements to each other and to gene expression is a sobering reminder that we might not be capturing the most important features or not understanding how relationships between features. We lack important features like transcription factor binding, long range elements, and 3D genome organization. In humans, most of the focus in the field is to resolve upstream long-range interactions between enhancers and promoters. ML techniques based on natural language processing have improved mod-

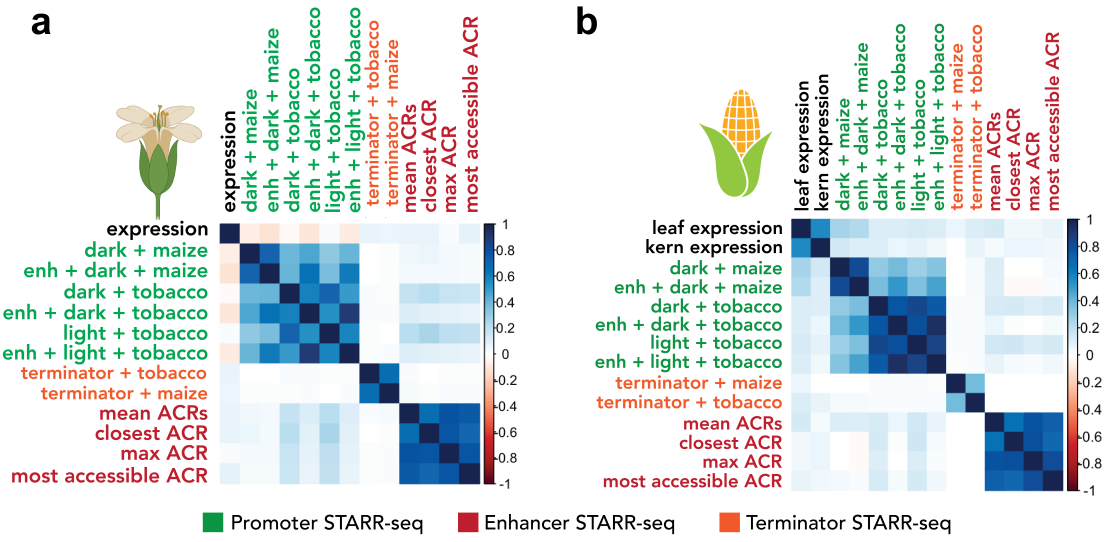


Figure 3.4: **Plant STARR-seq experiments show weak correlation to gene expression** Pearson's R correlation matrix between promoter STARR-seq experiments, enhancer STARR-seq, terminator STARR-seq, and gene expression for a) *Arabidopsis* genes and b) maize genes. For enhancer STARR-seq, since there were often multiple ACRs per gene, we correlated the average ACR strength (Mean ACRs), the ACR strength of the ACR closest to the gene TSS (closest ACR), the ACR strength of the strongest ACR (max ACR), and ACR strength of the ACR with the highest cut count from ATAC-seq (most accessible ACR).

eling interactions of long range *cis*-regulatory elements. One such model, called Enformer, can predict thousands of epigenetic and transcriptional profiles from human and mouse cell types using only DNA sequence as input [17]. Enformer employs a model architecture called a "transformer" to focus on different parts of the input sequence when making predictions. This architecture proved to be useful as Enformer substantially outperformed state-of-the-art models in predicting gene expression from cap-analysis gene expression (CAGE) experiments. Incorporating long range *cis*-regulatory information of plant genes, like long-read accessibility assays, will offer important training data as evinced by the success of Enformer in humans.

BIBLIOGRAPHY

[1] *A Survey of the Sorghum Transcriptome Using Single-molecule Long Reads*. United States. Department of Energy. Office of Science, 2016.

[2] Salah E Abdel-Ghany, Michael Hamilton, Jennifer L Jacobi, Peter Ngam, Nicholas Devitt, Faye Schilkey, Asa Ben-Hur, and Anireddy S N Reddy. A survey of the sorghum transcriptome using single-molecule long reads. *Nat. Commun.*, 7:11706, June 2016.

[3] J M Adams and S Cory. Untranslated nucleotide sequence at the 5'-end of R17 bacteriophage RNA. *Nature*, 227(5258):570–574, August 1970.

[4] M Adesnik, M Salditt, W Thomas, and J E Darnell. Evidence that all messenger RNA molecules (except histone messenger RNA) contain poly (a) sequences and that the Poly(A) has a nuclear function. *J. Mol. Biol.*, 71(1):21–30, October 1972.

[5] Vikram Agarwal, Sereno Lopez-Darwin, David R Kelley, and Jay Shendure. The landscape of alternative polyadenylation in single cells of the developing mouse embryo. *Nat. Commun.*, 12(1):5101, August 2021.

[6] Othman Al-Dossary, Agnelo Furtado, Ardashir KharabianMasouleh, Bader Alsubaie, Ibrahim Al-Mssallem, and Robert J Henry. Long read sequencing to reveal the full complexity of a plant transcriptome by targeting both standard and long workflows. *Plant Methods*, 19(1):112, October 2023.

[7] Edwards Allen and Miya D Howell. miRNAs in the biogenesis of trans-acting siRNAs in higher plants. *Semin. Cell Dev. Biol.*, 21(8):798–804, October 2010.

[8] Mohammed AlQuraishi. AlphaFold at CASP13. *Bioinformatics*, 35(22):4862–4865, November 2019.

[9] G An, A Mitra, H K Choi, M A Costa, K An, R W Thornburg, and C A Ryan. Functional analysis of the 3' control region of the potato wound-inducible proteinase inhibitor II gene. *Plant Cell*, 1(1):115–122, January 1989.

[10] Andreas I Andreou, Jessica Nirkko, Marisol Ochoa-Villarreal, and Naomi Nakayama. Mobius assembly for plant systems highlights promoter-terminator interaction in gene regulation. March 2021.

[11] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant arabidopsis thaliana. *Nature*, 408(6814):796–815, December 2000.

[12] Ashraful Arefeen, Juntao Liu, Xinshu Xiao, and Tao Jiang. TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics*, 34(15):2521–2529, August 2018.

[13] Ashraful Arefeen, Xinshu Xiao, and Tao Jiang. DeepPASTA: deep neural network based polyadenylation site analysis. *Bioinformatics*, 35(22):4577–4585, November 2019.

[14] Mohammad Shamsul Arefin, M Shamim Kaiser, Anirban Bandyopadhyay, Md Atiqur Rahman Ahad, and Kanad Ray. *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*. Springer Nature, December 2021.

[15] Cosmas D Arnold, Daniel Gerlach, Christoph Stelzer, Łukasz M Boryń, Martina Rath, and Alexander Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077, March 2013.

[16] H Aviv and P Leder. Purification of biologically active globin messenger RNA by chromatography on oligothymidylic acid-cellulose. *Proc. Natl. Acad. Sci. U. S. A.*, 69(6):1408–1412, June 1972.

[17] Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods*, 18(10):1196–1203, October 2021.

[18] Michael A Ayliffe, Martin Steinau, Robert F Park, Lee Rooke, Maria G Pacheco, Scot H Hulbert, Harold N Trick, and Anthony J Pryor. Aberrant mRNA processing of the maize Rp1-D rust resistance gene in wheat and barley. *Mol. Plant. Microbe. Interact.*, 17(8):853–864, August 2004.

[19] Kyungmin Baeg, Hiro-Oki Iwakawa, and Yukihide Tomari. The poly(a) tail blocks RDR6 from converting self mRNAs into substrates for gene silencing. *Nat Plants*, 3:17036, March 2017.

[20] Timothy L Bailey. STREME: accurate and versatile sequence motif discovery. *Bioinformatics*, 37(18):2834–2840, September 2021.

[21] Julia Bailey-Serres and Daniel R Gallie. *A Look Beyond Transcription: Mechanisms Determining MRNA Stability and Translation in Plants*. American Society of Plant Physiologists, 1998.

[22] Matthew N Bainbridge, René L Warren, Martin Hirst, Tammy Romanuik, Thomas Zeng, Anne Go, Allen Delaney, Malachi Griffith, Matthew Hickenbotham, Vincent Magrini, Elaine R Mardis, Marianne D Sadar, Asim S Siddiqui, Marco A Marra, and Steven J M Jones. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7:246, September 2006.

[23] Ruth E Baker, Jose-Maria Peña, Jayaratnam Jayamohan, and Antoine Jérusalem. Mechanistic models versus machine learning, a fight worth fighting for the biological community? *Biol. Lett.*, 14(5), May 2018.

[24] David Baulcombe. RNA silencing in plants. *Nature*, 431(7006):356–363, September 2004.

[25] E Beaudoing, S Freier, J R Wyatt, J M Claverie, and D Gautheret. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.*, 10(7):1001–1010, July 2000.

[26] Stephen A Bell and Arthur G Hunt. The arabidopsis ortholog of the 77 kda subunit of the cleavage stimulatory factor (AtCstF-77) involved in mRNA polyadenylation is an RNA-binding protein. *FEBS Lett.*, 584(8):1449–1454, April 2010.

[27] Drew T Bergman, Thouis R Jones, Vincent Liu, Judhajeet Ray, Evelyn Jagoda, Layla Siraj, Helen Y Kang, Joseph Nasser, Michael Kane, Antonio Rios, Tung H Nguyen, Sharon R Grossman, Charles P Fulco, Eric S Lander, and Jesse M Engreitz. Compatibility rules of human enhancer and promoter sequences. *Nature*, 607(7917):176–184, July 2022.

[28] Willian Souza Bernardes and Marcelo Menossi. Plant 3' regulatory regions from mRNA-Encoding genes and their uses to modulate expression, 2020.

[29] K Beyer, T Dandekar, and W Keller. RNA ligands selected by cleavage stimulation factor contain distinct sequence motifs that function as downstream elements in 3'-end processing of pre-mRNA. *J. Biol. Chem.*, 272(42):26769–26779, October 1997.

[30] C E Birse, L Minvielle-Sebastia, B A Lee, W Keller, and N J Proudfoot. Coupling termination of transcription to messenger RNA maturation in yeast. *Science*, 280(5361):298–301, April 1998.

[31] Benjamin J Blencowe. The relationship between alternative splicing and proteomic complexity. *Trends Biochem. Sci.*, 42(6):407–408, June 2017.

[32] Nicholas Bogard, Johannes Linder, Alexander B Rosenberg, and Georg Seelig. A deep neural network for predicting and engineering alternative polyadenylation. *Cell*, 178(1):91–106.e23, June 2019.

[33] S Brenner, F Jacob, and M Meselson. An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, 190:576–581, May 1961.

[34] Emily G Brooks, Estefania Elorriaga, Yang Liu, James R Duduit, Guoliang Yuan, Chung-Jui Tsai, Gerald A Tuskan, Thomas G Ranney, Xiaohan Yang, and Wusheng Liu. Plant promoters and terminators for High-Precision bioengineering. *Biodes Res*, 5:0013, July 2023.

[35] G G Brownlee and F Sanger. Chromatography of 32p-labelled oligonucleotides on thin layers of DEAE-cellulose. *Eur. J. Biochem.*, 11(2):395–399, December 1969.

[36] Quentin Bruggeman, Marie Garmier, Linda de Bont, Ludivine Soubigou-Taconnat, Christelle Mazubert, Moussa Benhamed, Cécile Raynaud, Catherine Bergounioux, and Marianne Delarue. The polyadenylation factor subunit CLEAVAGE AND POLYADENYLATION SPECIFICITY FACTOR30: A key factor of programmed cell death and a regulator of immunity in arabidopsis. *Plant Physiol.*, 165(2):732–746, June 2014.

[37] Christopher Buccitelli and Matthias Selbach. mRNAs, proteins and the emerging principles of gene expression control. *Nat. Rev. Genet.*, 21(10):630–644, October 2020.

[38] A R Buchman and P Berg. Comparison of intron-dependent and intron-independent gene expression. *Mol. Cell. Biol.*, 8(10):4395–4405, October 1988.

[39] J Callis, M Fromm, and V Walbot. Introns increase gene expression in cultured maize cells. *Genes Dev.*, 1(10):1183–1200, December 1987.

[40] S Carswell and J C Alwine. Efficiency of utilization of the simian virus 40 late polyadenylation site: effects of upstream sequences. *Mol. Cell. Biol.*, 9(10):4248–4258, October 1989.

[41] Serena Chan, Eun-A Choi, and Yongsheng Shi. Pre-mRNA 3'-end processing complex assembly and function. *Wiley Interdiscip. Rev. RNA*, 2(3):321–335, 2011.

[42] Hyeshik Chang, Jaechul Lim, Minju Ha, and V Narry Kim. TAIL-seq: genome-wide determination of poly(a) tail length and 3' end modifications. *Mol. Cell*, 53(6):1044–1052, March 2014.

[43] J C Chang, R Poon, K H Neumann, and Y W Kan. The nucleotide sequence of the 5' untranslated region of human gamma-globin mRNA. *Nucleic Acids Res.*, 5(10):3515–3522, October 1978.

[44] Jae-Woong Chang, Wei Zhang, Hsin-Sung Yeh, Ebbing P de Jong, Semo Jun, Kwan-Hyun Kim, Sun S Bae, Kenneth Beckman, Tae Hyun Hwang, Kye-Seong Kim, Do-Hyung Kim, Timothy J Griffin, Rui Kuang, and Jeongsik Yong. mRNA 3-UTR shortening is a molecular signature of mTORC1 activation, 2015.

[45] L C Chao, A Jamil, S J Kim, L Huang, and H G Martinson. Assembly of the cleavage and polyadenylation apparatus requires about 10 seconds in vivo and is faster for strong than for weak poly(a) sites. *Mol. Cell. Biol.*, 19(8):5588–5600, August 1999.

[46] Cho-Yi Chen, Shui-Tein Chen, Hsueh-Fen Juan, and Hsuan-Cheng Huang. Lengthening of 3UTR increases with morphological complexity in animal evolution. *Bioinformatics*, 28(24):3178–3181, October 2012.

[47] Chyi-Ying A Chen and Ann-Bin Shyu. Mechanisms of deadenylation-dependent decay. *Wiley Interdiscip. Rev. RNA*, 2(2):167–183, 2011.

[48] Fan Chen, Clinton C MacDonald, and Jeffrey Wilusf. Cleavage site determinants min the mammalian polydenylation signal. *Nucleic Acids Res.*, 23(14):2614–2620, July 1995.

[49] Moliang Chen, Guoli Ji, Hongjuan Fu, Qianmin Lin, Congting Ye, Wenbin Ye, Yaru Su, and Xiaohui Wu. A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Brief. Bioinform.*, 21(4):1261–1276, July 2020.

[50] Xinshi Chen, Yu Li, Ramzan Umarov, Xin Gao, and Le Song. RNA secondary structure prediction by learning unrolled algorithms. February 2020.

[51] Bing Cheng, Agnelo Furtado, and Robert J Henry. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *Gigascience*, 6(11):1–13, November 2017.

[52] Rocky Cheung, Kimberly D Insigne, David Yao, Christina P Burghard, Jeffrey Wang, Yun-Hua E Hsiao, Eric M Jones, Daniel B Goodman, Xinshu Xiao, and Sriram Kosuri. A multiplexed assay for exon recognition reveals that an unappreciated fraction of rare genetic variants cause Large-Effect splicing disruptions. *Mol. Cell*, 73(1):183–194.e8, January 2019.

[53] S N Cohen, A C Chang, H W Boyer, and R B Helling. Construction of biologically functional bacterial plasmids in vitro. *Proc. Natl. Acad. Sci. U. S. A.*, 70(11):3240–3244, November 1973.

[54] S Connelly and J L Manley. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.*, 2(4):440–452, April 1988.

[55] Xuean Cui, Zhe Liang, Lisha Shen, Qian Zhang, Shengjie Bao, Yuke Geng, Bin Zhang, Vonny Leo, Leah A Vardy, Tiegang Lu, Xiaofeng Gu, and Hao Yu. 5-methylcytosine RNA methylation in arabidopsis thaliana. *Mol. Plant*, 10(11):1387–1399, November 2017.

[56] Josh T Cuperus, Benjamin Groves, Anna Kuchina, Alexander B Rosenberg, Nebojsa Jojic, Stanley Fields, and Georg Seelig. Deep learning of the regulatory grammar of yeast 5' untranslated regions from 500,000 random sequences. *Genome Res.*, 27(12):2015–2024, December 2017.

[57] Malgorzata Cyrek, Halina Fedak, Arkadiusz Ciesielski, Yanwu Guo, Aleksandra Sliwa, Lien Brzezniak, Katarzyna Krzyczmonik, Zbigniew Pietras, Szymon Kaczanowski, Fuquan Liu, and Szymon Swiezewski. Seed dormancy in arabidopsis is controlled by alternative polyadenylation of DOG1. *Plant Physiol.*, 170(2):947–955, February 2016.

[58] Elena Dadami, Athanasios Dalakouras, Michele Zwiebel, Gabi Krczal, and Michael Wassenegger. An endogene-resembling transgene is resistant to DNA methylation and systemic silencing. *RNA Biol.*, 11(7):934–941, July 2014.

[59] Elena Dadami, Mirko Moser, Michele Zwiebel, Gabi Krczal, Michael Wassenegger, and Athanasios Dalakouras. An endogene-resembling transgene delays the onset of silencing and limits siRNA accumulation. *FEBS Lett.*, 587(6):706–710, March 2013.

[60] J E Darnell, L Philipson, R Wall, and M Adesnik. Polyadenylic acid sequences: role in conversion of nuclear RNA into messenger RNA. *Science*, 174(4008):507–510, October 1971.

[61] Carl G de Boer, Eeshit Dhaval Vaishnav, Ronen Sadeh, Esteban Luis Abeyta, Nir Friedman, and Aviv Regev. Author correction: Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nat. Biotechnol.*, 38(10):1211, October 2020.

[62] Felipe F de Felippes and Peter M Waterhouse. Plant terminators: the unsung heroes of gene expression. *J. Exp. Bot.*, 74(7):2239–2250, April 2023.

[63] Felipe Fenselau de Felippes, Kylie Shand, and Peter M Waterhouse. Identification of a transferrable terminator element that inhibits small RNA production and improves transgene expression levels. *Front. Plant Sci.*, 13:877793, May 2022.

[64] C Dean, S Tamaki, P Dunsmuir, M Favreau, C Katayama, H Dooner, and J Bedbrook. mRNA transcripts of several plant genes are polyadenylated at multiple sites in vivo. *Nucleic Acids Res.*, 14(5):2229–2240, March 1986.

[65] Kaixuan Deng, Qizhe Zhang, Yuxin Hong, Jianbing Yan, and Xuehai Hu. iCREPCP: A deep learning-based web server for identifying base-resolution cis-regulatory elements within plant core promoters. *Plant Commun*, 4(1):100455, January 2023.

[66] Zhi-Luo Deng, Philipp C Münch, René Mreches, and Alice C McHardy. Rapid and accurate identification of ribosomal RNA sequences via deep learning. *Nucleic Acids Res.*, 50(10):e60, June 2022.

[67] Department of Economic and Social Affairs. *World Economic Situation and Prospects 2022*. United Nations, January 2022.

[68] Adnan Derti, Philip Garrett-Engele, Kenzie D Macisaac, Richard C Stevens, Shreedharan Sriram, Ronghua Chen, Carol A Rohl, Jason M Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, 22(6):1173–1183, June 2012.

[69] P Dhaese, H De Greve, J Gielen, L Seurinck, M Van Montagu, and J Schell. Identification of sequences involved in the polyadenylation of higher plant nuclear transcripts using agrobacterium T-DNA genes as models. *EMBO J.*, 2(3):419–426, 1983.

[70] Andrew G Diamos and Hugh S Mason. Chimeric 3' flanking regions strongly enhance gene expression in plants. *Plant Biotechnol. J.*, 16(12):1971–1982, December 2018.

[71] S H Diehn, W L Chiu, E J De Rocher, and P J Green. Premature polyadenylation at multiple sites within a bacillus thuringiensis toxin gene-coding region. *Plant Physiol.*, 117(4):1433–1443, August 1998.

[72] P Early, J Rogers, M Davis, K Calame, M Bond, R Wall, and L Hood. Two mRNAs can be produced from a single immunoglobulin mu gene by alternative RNA processing pathways. *Cell*, 20(2):313–319, June 1980.

[73] Joshua D Eaton and Steven West. Termination of transcription by RNA polymerase II: BOOM! *Trends Genet.*, 36(9):664–675, September 2020.

[74] Christian R Eckmann, Christiane Rammelt, and Elmar Wahle. Control of poly(a) tail length. *Wiley Interdiscip. Rev. RNA*, 2(3):348–361, 2011.

[75] M Edmonds, M H Vaughan, Jr, and H Nakazato. Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship. *Proc. Natl. Acad. Sci. U. S. A.*, 68(6):1336–1340, June 1971.

[76] Timothy J Eisen, Stephen W Eichhorn, Alexander O Subtelny, Kathy S Lin, Sean E McGeary, Sumeet Gupta, and David P Bartel. The dynamics of cytoplasmic mRNA metabolism. *Mol. Cell*, 77(4):786–799.e10, February 2020.

[77] Yuval Eshed and Zachary B Lippman. Revolutions in agriculture chart a course for targeted breeding of old and new crops. *Science*, 366(6466), November 2019.

[78] Felipe F de Felippes, Marcus McHale, Rachel L Doran, Sally Roden, Andrew L Eamens, E Jean Finnegan, and Peter M Waterhouse. The key role of terminators on the expression and post-transcriptional gene silencing of transgenes. *Plant J.*, 104(1):96–112, September 2020.

[79] Jia-Wu Feng, Shanshan Huang, Yi-Xiong Guo, Dongxu Liu, Jia-Ming Song, Junxiang Gao, Huan Li, and Ling-Ling Chen. Plant ISOform sequencing database (PISO): a comprehensive repertory of full-length transcripts in plants. *Plant Biotechnol. J.*, 17(6):1001–1003, June 2019.

[80] M Fitzgerald. The sequence 5-AAUAAA-3 forms part of the recognition site for polyadenylation of late SV40 mRNAs. *Cell*, 24(1):251–260, April 1981.

[81] Kevin P Forbes, Balasubrahmanyam Addepalli, and Arthur G Hunt. An arabidopsis fip1 homolog interacts with RNA and provides conceptual links with a number of other polyadenylation factor subunits. *J. Biol. Chem.*, 281(1):176–186, January 2006.

[82] Haihui Fu, Dewei Yang, Wenyue Su, Liuyin Ma, Yingjia Shen, Guoli Ji, Xinfu Ye, Xiaohui Wu, and Qingshun Q Li. Genome-wide dynamics of alternative polyadenylation in rice. *Genome Res.*, 26(12):1753–1760, December 2016.

[83] Zong-Heng Fu, Si-Zhe He, Yi Wu, and Guang-Rong Zhao. Design and deep learning of synthetic b-cell-specific promoters. *Nucleic Acids Res.*, October 2023.

[84] Hiroyuki Fuke and Mutsuhito Ohno. Role of poly (a) tail as an identity element for mRNA nuclear export. *Nucleic Acids Res.*, 36(3):1037–1049, February 2008.

[85] Y Furuichi, M Morgan, S Muthukrishnan, and A J Shatkin. Reovirus messenger RNA contains a methylated, blocked 5'-terminal structure: m-7G(5')ppp(5')G-MpCp-. *Proc. Natl. Acad. Sci. U. S. A.*, 72(1):362–366, January 1975.

[86] Y Furuichi, M Morgan, A J Shatkin, W Jelinek, M Salditt-Georgieff, and J E Darnell. Methylated, blocked 5 termini in HeLa cell mRNA. *Proc. Natl. Acad. Sci. U. S. A.*, 72(5):1904–1908, May 1975.

[87] Joseph L Gage, Sujina Mali, Fionn McLoughlin, Merritt Khaipho-Burch, Brandon Monier, Julia Bailey-Serres, Richard D Vierstra, and Edward S Buckler. Variation in upstream open reading frames contributes to allelic diversity in maize protein abundance. *Proc. Natl. Acad. Sci. U. S. A.*, 119(14):e2112516119, April 2022.

[88] Jenna E Gallegos and Alan B Rose. The enduring mystery of intron-mediated enhancement. *Plant Sci.*, 237:8–15, August 2015.

[89] D R Gallie. The cap and poly(a) tail function synergistically to regulate mRNA translational efficiency. *Genes Dev.*, 5(11):2108–2116, November 1991.

[90] Xin Gao, Jie Zhang, Zhi Wei, and Hakon Hakonarson. DeepPolyA: A convolutional neural network approach for polyadenylation site prediction. *IEEE Access*, 6:24340–24349, 2018.

[91] Yipeng Gao, Lei Li, Christopher I Amos, and Wei Li. Analysis of alternative polyadenylation from single-cell RNA-seq using scDaPars reveals cell subpopulations invisible to gene expression. *Genome Res.*, 31(10):1856–1866, October 2021.

[92] D Gautheret, O Poirot, F Lopez, S Audic, and J M Claverie. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.*, 8(5):524–530, May 1998.

[93] Filippo Geraci, Indrajit Saha, and Monica Bianchini. *RNA-Seq Analysis: Methods, Applications and Challenges*. Frontiers Media SA, June 2020.

[94] A Gil and N J Proudfoot. A sequence downstream of AAUAAA is required for rabbit beta-globin mRNA 3'-end formation. *Nature*, 312(5993):473–474, 1984.

[95] A Gil and N J Proudfoot. Position-dependent sequence elements downstream of AAUAAA are required for efficient rabbit beta-globin mRNA 3' end formation. *Cell*, 49(3):399–406, May 1987.

[96] Marek Gniadkowski, Maja Hemmings-Mieszczak, Ulrich Klahre, Hong-Xiang Liu, and Witold Filipowicz. Characterisation of intronic Uridine-Rich sequence elements acting as possible targets for nuclear proteins during Pre-mRNA splicing in nicotiana plumbaginifolia. *Nucleic Acids Res.*, 24(4):619–627, February 1996.

[97] G J Goodall and W Filipowicz. The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell*, 58(3):473–483, August 1989.

[98] G J Goodall and W Filipowicz. Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. *EMBO J.*, 10(9):2635–2644, September 1991.

[99] J H Graber, C R Cantor, S C Mohr, and T F Smith. In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc. Natl. Acad. Sci. U. S. A.*, 96(24):14055–14060, November 1999.

[100] Dustin Griesemer, James R Xue, Steven K Reilly, Jacob C Ulirsch, Kalki Kukreja, Joe R Davis, Masahiro Kanai, David K Yang, John C Butts, Mehmet H Guney, Jeremy Luban, Stephen B Montgomery, Hilary K Finucane, Carl D Novina, Ryan Tewhey, and

Pardis C Sabeti. Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell*, 184(20):5247–5260.e19, September 2021.

[101] Sharon R Grossman, Xiaolan Zhang, Li Wang, Jesse Engreitz, Alexandre Melnikov, Peter Rogov, Ryan Tewhey, Alina Isakova, Bart Deplancke, Bradley E Bernstein, Tarjei S Mikkelsen, and Eric S Lander. Systematic dissection of genomic features determining transcription factor binding and enhancer function. *Proc. Natl. Acad. Sci. U. S. A.*, 114(7):E1291–E1300, February 2017.

[102] Aditya Grover, Ashish Kapoor, and Eric Horvitz. A deep hybrid model for weather forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 379–386, New York, NY, USA, August 2015. Association for Computing Machinery.

[103] Andreas J Gruber, Ralf Schmidt, Andreas R Gruber, Georges Martin, Souvik Ghosh, Manuel Belmadani, Walter Keller, and Mihaela Zavolan. A comprehensive analysis of 3' end sequencing data sets reveals novel polyadenylation signals and the repressive role of heterogeneous ribonucleoprotein C on cleavage and polyadenylation. *Genome Res.*, 26(8):1145–1159, August 2016.

[104] Kevin C H Ha, Benjamin J Blencowe, and Quaid Morris. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol.*, 19(1):45, March 2018.

[105] Sebastian Castillo Hair, Stephen Fedak, Ban Wang, Johannes Linder, Kyle Havens, Michael Certo, and Georg Seelig. Optimizing 5'UTRs for mRNA-delivered gene editing using deep learning. June 2023.

[106] Ashwin Hajarnavis, Ian Korf, and Richard Durbin. A probabilistic model of 3' end formation in caenorhabditis elegans. *Nucleic Acids Res.*, 32(11):3392–3399, June 2004.

[107] Paul F Harrison, David R Powell, Jennifer L Clancy, Thomas Preiss, Peter R Boag, Ana Traven, Torsten Seemann, and Traude H Beilharz. PAT-seq: a method to study the integration of 3'-UTR dynamics with gene expression in the eukaryotic transcriptome. *RNA*, 21(8):1502–1510, August 2015.

[108] P Cody He, Jiangbo Wei, Xiaoyang Dou, Bryan T Harada, Zijie Zhang, Ruiqi Ge, Chang Liu, Li-Sheng Zhang, Xianbin Yu, Shuai Wang, Ruitu Lyu, Zhongyu Zou, Mengjie Chen, and Chuan He. Exon architecture controls mRNA m6a suppression and gene expression. *Science*, 379(6633):677–682, February 2023.

[109] Carlos Hernández-Lucas, Joaquin Royo, Javier Paz-Ares, Fernando Ponz, Francisco García-Olmedo, and Pilar Carbonero. Polyadenylation site heterogeneity in mRNA

encoding the precursor of the barley toxin β-hordothionin. *FEBS Lett.*, 200(1):103–106, May 1986.

[110] Alan J Herr, Attila Molnàr, Alex Jones, and David C Baulcombe. Defective RNA processing enhances RNA silencing and influences flowering of arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, 103(41):14994–15001, October 2006.

[111] Lee T Hickey, Amber N Hafeez, Hannah Robinson, Scott A Jackson, Soraya C M Leal-Bertioli, Mark Tester, Caixia Gao, Ian D Godwin, Ben J Hayes, and Brande B H Wulff. Breeding crops to feed 10 billion. *Nat. Biotechnol.*, 37(7):744–754, July 2019.

[112] Jean-Michel Hily, Stacy D Singer, Yazhou Yang, and Zongrang Liu. A transformation booster sequence (TBS) from petunia hybrida functions as an enhancer-blocking insulator in arabidopsis thaliana. *Plant Cell Rep.*, 28(7):1095–1104, July 2009.

[113] Alan G Hinnebusch, Ivaylo P Ivanov, and Nahum Sonenberg. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science*, 352(6292):1413–1416, June 2016.

[114] Tadayoshi Hirai, Natsuko Kurokawa, Narendra Duhita, Kyoko Hiwasa-Tanase, Kazuhisa Kato, Ko Kato, and Hiroshi Ezura. The HSP terminator of arabidopsis thaliana induces a high level of miraculin accumulation in transgenic tomatoes. *J. Agric. Food Chem.*, 59(18):9942–9949, September 2011.

[115] Kyoko Hiwasa-Tanase, Mpanja Nyarubona, Tadayoshi Hirai, Kazuhisa Kato, Takanari Ichikawa, and Hiroshi Ezura. High-level accumulation of recombinant miraculin protein in transgenic tomatoes expressing a synthetic miraculin gene with optimized codon usage terminated by the native miraculin terminator. *Plant Cell Rep.*, 30(1):113–124, January 2011.

[116] Nam V Hoang, Agnelo Furtado, Patrick J Mason, Annelie Marquardt, Lakshmi Kasirajan, Prathima P Thirugnanasambandam, Frederik C Botha, and Robert J Henry. A survey of the complex transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing. *BMC Genomics*, 18(1):395, May 2017.

[117] Ivo L Hofacker. Vienna RNA secondary structure server. *Nucleic Acids Res.*, 31(13):3429–3431, July 2003.

[118] Mainul Hoque, Zhe Ji, Dinghai Zheng, Wenting Luo, Wencheng Li, Bei You, Ji Yeon Park, Ghassan Yehia, and Bin Tian. Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nat. Methods*, 10(2):133–139, February 2013.

[119] Polly Yingshan Hsu, Lorenzo Calviello, Hsin-Yen Larry Wu, Fay-Wei Li, Carl J Rothfels, Uwe Ohler, and Philip N Benfey. Super-resolution ribosome profiling reveals unannotated translation events in arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, 113(45):E7126–E7135, November 2016.

[120] Jianzhong Hu, Stefano Manduzio, and Hunseung Kang. Epitranscriptomic RNA methylation in plant development and abiotic stress responses. *Front. Plant Sci.*, 10:500, April 2019.

[121] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[122] X Huang, M D Adams, H Zhou, and A R Kerlavage. A tool for analyzing and annotating genomic sequences. *Genomics*, 46(1):37–45, November 1997.

[123] Y Huang and G G Carmichael. Role of polyadenylation in nucleocytoplasmic transport of mRNA. *Mol. Cell. Biol.*, 16(4):1534–1542, April 1996.

[124] A G Hunt. Messenger RNA3'end formation in plants. *Annu. Rev. Plant Biol.*, 1994.

[125] A G Hunt. Messenger RNA 3 end formation in plants. *Nuclear pre-mRNA Processing in Plants*, pages 151–177, 2008.

[126] A G Hunt, N M Chu, J T Odell, F Nagy, and N H Chua. Plant cells do not properly recognize animal gene polyadenylation signals. *Plant Mol. Biol.*, 8(1):23–35, January 1987.

[127] Arthur G Hunt. RNA regulatory elements and polyadenylation in plants. *Front. Plant Sci.*, 2:109, 2011.

[128] Arthur G Hunt. mRNA 3' end formation in plants: Novel connections to growth, development and environmental responses. *Wiley Interdiscip. Rev. RNA*, 11(3):e1575, May 2020.

[129] Arthur G Hunt. Review: Mechanisms underlying alternative polyadenylation in plants - looking in the right places. *Plant Sci.*, 324:111430, November 2022.

[130] Arthur G Hunt, Denghui Xing, and Qingshun Q Li. Plant polyadenylation factors: conservation and variety in the polyadenylation complex in plants. *BMC Genomics*, 13:641, November 2012.

[131] I L Ingelbrecht, L M Herman, R A Dekeyser, M C Van Montagu, and A G Depicker. Different 3' end regions strongly influence the level of gene expression in plant cells. *Plant Cell*, 1(7):671–680, July 1989.

[132] F Jacob and J Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, June 1961.

[133] Zehra Jafar, Salma Tariq, Irfan Sadiq, Tayyab Nawaz, and Malik Nadeem Akhtar. Genome-Wide profiling of polyadenylation events in maize using High-Throughput transcriptomic sequences. *G3*, 9(8):2749–2760, August 2019.

[134] Calvin H Jan, Robin C Friedman, J Graham Ruby, and David P Bartel. Formation, regulation and evolution of caenorhabditis elegans 3UTRs. *Nature*, 469(7328):97–101, November 2010.

[135] W Jelinek, M Adesnik, M Salditt, D Sheiness, R Wall, G Molloy, L Philipson, and J E Darnell. Further evidence on the nuclear origin and transfer to the cytoplasm of polyadenylic acid sequences in mammalian cell RNA. *J. Mol. Biol.*, 75(3):515–532, April 1973.

[136] Guoli Ji, Lei Li, Qingshun Q Li, Xiangdong Wu, Jingyi Fu, Gong Chen, and Xiaohui Wu. PASPA: a web server for mRNA poly(a) site predictions in plants and algae. *Bioinformatics*, 31(10):1671–1673, May 2015.

[137] Guoli Ji, Xiaohui Wu, Yingjia Shen, Jiangyin Huang, and Qingshun Quinn Li. A classification-based prediction model of messenger RNA polyadenylation sites. *J. Theor. Biol.*, 265(3):287–296, August 2010.

[138] Guoli Ji, Jianti Zheng, Yingjia Shen, Xiaohui Wu, Ronghan Jiang, Yun Lin, Johnny C Loke, Kimberly M Davis, Greg J Reese, and Qingshun Quinn Li. Predictive modeling of plant messenger RNA polyadenylation sites. *BMC Bioinformatics*, 8:43, February 2007.

[139] Jinbu Jia, Wenqin Lu, Bo Liu, Huihui Fang, Yiming Yu, Weipeng Mo, Hong Zhang, Xianhao Jin, Yi Shu, Yanping Long, Yanxi Pei, and Jixian Zhai. An atlas of plant full-length RNA reveals tissue-specific and monocots–dicots conserved regulation of poly(a) tail length. *Nature Plants*, 8(9):1118–1126, August 2022.

[140] T Jores, M Hamm, J T Cuperus, and C Queitsch. Frontiers and techniques in plant gene regulation. *Current Opinion in Plant Biology*, page (in press), 2023.

[141] Tobias Jores, Jackson Tonnies, Michael W Dorrity, Josh T Cuperus, Stanley Fields, and Christine Queitsch. Identification of plant enhancers and their constituent elements by STARR-seq in tobacco leaves. *Plant Cell*, 32(7):2120–2131, July 2020.

[142] Tobias Jores, Jackson Tonnies, Travis Wrightsman, Edward S Buckler, Josh T Cuperus, Stanley Fields, and Christine Queitsch. Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat Plants*, 7(6):842–855, June 2021.

[143] C P Joshi. Putative polyadenylation signals in nuclear genes of higher plants: a compilation and analysis. *Nucleic Acids Res.*, 15(23):9627–9640, December 1987.

[144] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[145] Manal Kalkatawi, Arturo Magana-Mora, Boris Jankovic, and Vladimir B Bajic. DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. *Bioinformatics*, 35(7):1125–1132, April 2019.

[146] Kalli Kappel, Kaiming Zhang, Zhaoming Su, Andrew M Watkins, Wipapat Kladwang, Shanshan Li, Grigore Pintilie, Ved V Topkar, Ramya Rangan, Ivan N Zheludev, Joseph D Yesselman, Wah Chiu, and Rhiju Das. Accelerated cryo-EM-guided determination of three-dimensional RNA-only structures. *Nat. Methods*, 17(7):699–707, July 2020.

[147] Andrey Kechin, Uljana Boyarskikh, Alexander Kel, and Maxim Filipenko. cutprimers: A new tool for accurate cutting of primers from reads of targeted next generation sequencing. *J. Comput. Biol.*, 24(11):1138–1143, November 2017.

[148] Jack D Keene. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.*, 8(7):533–543, July 2007.

[149] Brian Keith and Nam-Hai Chua. Monocot and dicot pre-mRNAs are processed with different efficiencies in transgenic tobacco, 1986.

[150] S Kertesz, Z Kerenyi, Z Merai, I Bartos, and others. Both introns and long 3-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic acids*, 2006.

[151] Pouya Kheradpour, Jason Ernst, Alexandre Melnikov, Peter Rogov, Li Wang, Xiaolan Zhang, Jessica Alston, Tarjei S Mikkelsen, and Manolis Kellis. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res.*, 23(5):800–811, May 2013.

[152] Sol Kim, Junichi Yamamoto, Yexi Chen, Masatoshi Aida, Tadashi Wada, Hiroshi Handa, and Yuki Yamaguchi. Evidence that cleavage factor im is a heterotetrameric

protein complex controlling alternative polyadenylation. *Genes Cells*, 15(9):1003–1013, September 2010.

[153] Younghyun Kim, Goeun Lee, Eunhyun Jeon, Eun Ju Sohn, Yongjik Lee, Hyangju Kang, Dong Wook Lee, Dae Heon Kim, and Inhwan Hwang. The immediate upstream region of the 5'-UTR from the AUG start codon has a pronounced effect on the translational efficiency in arabidopsis thaliana. *Nucleic Acids Res.*, 42(1):485–498, January 2014.

[154] J Kim-Ha, P J Webster, J L Smith, and P M Macdonald. Multiple RNA regulatory elements mediate distinct steps in localization of oskar mRNA. *Development*, 119(1):169–178, September 1993.

[155] Martin Kircher, Chenling Xiong, Beth Martin, Max Schubach, Fumitaka Inoue, Robert J A Bell, Joseph F Costello, Jay Shendure, and Nadav Ahituv. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat. Commun.*, 10(1):3583, August 2019.

[156] E H Kislauskis, X Zhu, and R H Singer. Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. *J. Cell Biol.*, 127(2):441–451, October 1994.

[157] Anna V Klepikova, Artem S Kasianov, Evgeny S Gerasimov, Maria D Logacheva, and Aleksey A Penin. A high resolution map of the arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. *Plant J.*, 88(6):1058–1070, December 2016.

[158] Adam Klie, Hayden Stites, Tobias Jores, Joe J Solvason, Emma K Farley, and Hannah Carter. EUGENe: A python toolkit for predictive analyses of regulatory sequences. November 2022.

[159] Chuan Hock Koh and Limsoon Wong. Recognition of polyadenylation sites from arabidopsis genomic sequences. *Genome Inform.*, 19:73–82, 2007.

[160] Sriram Kosuri, Daniel B Goodman, Guillaume Cambray, Vivek K Mutalik, Yuan Gao, Adam P Arkin, Drew Endy, and George M Church. Composability of regulatory sequences controlling transcription and translation in escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.*, 110(34):14024–14029, August 2013.

[161] Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins*, 89(12):1607–1617, December 2021.

[162] Michal Krzyszton, Monika Zakrzewska-Placzek, Aleksandra Kwasnik, Norbert Dojer, Wojciech Karlowski, and Joanna Kufel. Defective XRN3-mediated transcription

termination in arabidopsis affects the expression of protein-coding genes. *Plant J.*, 93(6):1017–1031, March 2018.

[163] Tomohiro Kubo, Tadashi Wada, Yuki Yamaguchi, Akira Shimizu, and Hiroshi Handa. Knock-down of 25 kda subunit of cleavage factor im in hela cells alters alternative polyadenylation within 3'-UTRs. *Nucleic Acids Res.*, 34(21):6264–6271, November 2006.

[164] Uwe Kühn, Miriam Gündel, Anne Knoth, Yvonne Kerwitz, Sabine Rüdel, and Elmar Wahle. Poly(A) tail length is controlled by the nuclear poly(a)-binding protein regulating the interaction between poly(a) polymerase and the cleavage and polyadenylation specificity factor. *J. Biol. Chem.*, 284(34):22803–22814, August 2009.

[165] Kimberly R Kukurba and Stephen B Montgomery. RNA sequencing and analysis. *Cold Spring Harb. Protoc.*, 2015(11):951–969, April 2015.

[166] Ananthanarayanan Kumar, Marcello Clerici, Lena M Muckenfuss, Lori A Passmore, and Martin Jinek. Mechanistic insights into mRNA 3'-end processing. *Curr. Opin. Struct. Biol.*, 59:143–150, December 2019.

[167] G Renuka Kumar and Britt A Glaunsinger. Nuclear import of cytoplasmic poly(a) binding protein restricts gene expression via hyperadenylation and nuclear retention of mRNA. *Mol. Cell. Biol.*, 30(21):4996–5008, November 2010.

[168] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, March 2012.

[169] Cécile Lecampion, Maïna Floris, Jean Raphaël Fantino, Christophe Robaglia, and Christophe Laloi. An easy method for plant polysome profiling. *J. Vis. Exp.*, (114), August 2016.

[170] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[171] Ivano Legnini, Jonathan Alles, Nikos Karaiskos, Salah Ayoub, and Nikolaus Rajewsky. FLAM-seq: full-length mRNA sequencing reveals principles of poly(a) tail length control. *Nat. Methods*, 16(9):879–886, September 2019.

[172] Lei Lei, Junpeng Shi, Jian Chen, Mei Zhang, Silong Sun, Shaojun Xie, Xiaojie Li, Biao Zeng, Lizeng Peng, Andrew Hauck, Haiming Zhao, Weibin Song, Zaifeng Fan, and Jinsheng Lai. Ribosome profiling reveals dynamic translational landscape in maize seedlings under drought stress. *Plant J.*, 84(6):1206–1218, December 2015.

[173] Michael K K Leung, Andrew Delong, and Brendan J Frey. Inference of the human polyadenylation code. *Bioinformatics*, 34(17):2889–2898, September 2018.

[174] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[175] Q Li and A G Hunt. The polyadenylation of RNA in plants. *Plant Physiol.*, 115(2):321–325, October 1997.

[176] Wencheng Li, Bei You, Mainul Hoque, Dinghai Zheng, Wenting Luo, Zhe Ji, Ji Yeon Park, Samuel I Gunderson, Auinash Kalsotra, James L Manley, and Bin Tian. Systematic profiling of Poly(A)+ transcripts modulated by core 3' end processing and splicing factors reveals regulatory rules of alternative cleavage and polyadenylation. *PLoS Genet.*, 11(4):e1005166, April 2015.

[177] Zhongxiao Li, Elva Gao, Juexiao Zhou, Wenkai Han, Xiaopeng Xu, and Xin Gao. Applications of deep learning in understanding gene regulation. *Cell Rep Methods*, 3(1):100384, January 2023.

[178] Steve Lianoglou, Vidur Garg, Julie L Yang, Christina S Leslie, and Christine Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes Dev.*, 27(21):2380–2396, November 2013.

[179] Jaechul Lim, Mihye Lee, Ahyeon Son, Hyeshik Chang, and V Narry Kim. mTAIL-seq reveals dynamic poly(a) tail regulation in oocyte-to-embryo development. *Genes Dev.*, 30(14):1671–1682, July 2016.

[180] L Lim and E S Canellakis. Adenine-rich polymer associated with rabbit reticulocyte messenger RNA. *Nature*, 227(5259):710–712, August 1970.

[181] Juncheng Lin, Ruqiang Xu, Xiaohui Wu, Yingjia Shen, and Qingshun Q Li. Role of cleavage and polyadenylation specificity factor 100: anchoring poly(a) sites and modulating transcription termination. *Plant J.*, 91(5):829–839, September 2017.

[182] Adam J Litterman, Robin Kageyama, Olivier Le Tonqueze, Wenxue Zhao, John D Gagnon, Hani Goodarzi, David J Erle, and K Mark Ansel. A massively parallel 3' UTR reporter assay reveals relationships between nucleotide content, sequence conservation, and mRNA destabilization. *Genome Res.*, 29(6):896–906, June 2019.

[183] Gaofeng Liu, Jin Wang, and Xilin Hou. Transcriptome-Wide N6-Methyladenosine (m6a) methylome profiling of heat stress in pak-choi (brassica rapa ssp. chinensis). *Plants*, 9(9), August 2020.

[184] Min Liu, Jiafu Zhu, and Zhicheng Dong. Immediate transcriptional responses of arabidopsis leaves to heat shock. *J. Integr. Plant Biol.*, 63(3):468–483, March 2021.

[185] Y Liu, H Nie, H Liu, and F Lu. inclusive RNA isoform sequencing (PAIso- seq) reveals wide-spread non-adenosine residues within RNA poly (a) tails. nat commun. 2019; 10 (1): 1–13.

[186] Johnny C Loke, Eric A Stahlberg, David G Strenski, Brian J Haas, Paul Chris Wood, and Qingshun Quinn Li. Compilation of mRNA polyadenylation signals in arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol.*, 138(3):1457–1468, July 2005.

[187] Yanping Long, Jinbu Jia, Weipeng Mo, Xianhao Jin, and Jixian Zhai. FLEP-seq: simultaneous detection of RNA polymerase II position, splicing status, polyadenylation site and poly(a) tail length at genome-wide scale by single-molecule nascent RNA sequencing. *Nat. Protoc.*, 16(9):4355–4381, September 2021.

[188] R Lorenz, S H Bernhart, H Z Siederdissen, H Tafer, C Flamm, P F Stadler, and Others. ViennaRNA package 2.0. algorithm mol biol 6: 26, 2011.

[189] Julius B Lucks, Stefanie A Mortimer, Cole Trapnell, Shujun Luo, Sharon Aviran, Gary P Schroth, Lior Pachter, Jennifer A Doudna, and Adam P Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proc. Natl. Acad. Sci. U. S. A.*, 108(27):11063–11068, July 2011.

[190] K R Luehrsen and V Walbot. Intron creation and polyadenylation in maize are directed by AU-rich RNA. *Genes Dev.*, 8(9):1117–1130, May 1994.

[191] Radoslaw Lukoszek, Peter Feist, and Zoya Ignatova. Insights into the adaptive response of arabidopsis thaliana to prolonged thermal stress by ribosomal profiling and RNA-Seq. *BMC Plant Biol.*, 16(1):221, October 2016.

[192] Guan-Zheng Luo, Alice MacQueen, Guanqun Zheng, Hongchao Duan, Louis C Dore, Zhike Lu, Jun Liu, Kai Chen, Guifang Jia, Joy Bergelson, and Chuan He. Unique features of the m6a methylome in arabidopsis thaliana. *Nat. Commun.*, 5:5630, November 2014.

[193] Weifei Luo, Arlen W Johnson, and David L Bentley. The role of rat1 in coupling mRNA 3'-end processing to transcription termination: implications for a unified allosteric-torpedo model. *Genes Dev.*, 20(8):954–965, April 2006.

[194] Zhenghua Luo and Zhixiang Chen. Improperly terminated, unpolyadenylated mRNA of sense transgenes is targeted by RDR6-mediated RNA silencing in arabidopsis. *Plant Cell*, 19(3):943–958, March 2007.

[195] P M Macdonald and G Struhl. cis-acting sequences responsible for anterior localization of bicoid mRNA in drosophila embryos. *Nature*, 336(6199):595–598, December 1988.

[196] C R Mandel, Y Bai, and L Tong. Protein factors in pre-mRNA 3'-end processing. *Cell. Mol. Life Sci.*, 65(7-8):1099–1122, April 2008.

[197] Nicholas R Markham and Michael Zuker. UNAFold: software for nucleic acid folding and hybridization. *Methods Mol. Biol.*, 453:3–31, 2008.

[198] C A Marotta, B G Forget, S M Weissman, I M Verma, R P McCaffrey, and D Baltimore. Nucleotide sequences of human globin messenger RNA. *Proc. Natl. Acad. Sci. U. S. A.*, 71(6):2300–2304, June 1974.

[199] Yamile Marquez, John W S Brown, Craig Simpson, Andrea Barta, and Maria Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in arabidopsis. *Genome Res.*, 22(6):1184–1195, June 2012.

[200] Lucia Marucci, Matteo Barberis, Jonathan Karr, Oliver Ray, Paul R Race, Miguel de Souza Andrade, Claire Grierson, Stefan Andreas Hoffmann, Sophie Landon, Elibio Rech, Joshua Rees-Garbutt, Richard Seabrook, William Shaw, and Christopher Woods. Computer-Aided Whole-Cell design: Taking a holistic approach by integrating synthetic with systems biology. *Front Bioeng Biotechnol*, 8:942, August 2020.

[201] Andre P Masella, Andrea K Bartram, Jakub M Truszkowski, Daniel G Brown, and Josh D Neufeld. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics*, 13:31, February 2012.

[202] Christine Mayr. Evolution and biological roles of alternative 3'UTRs. *Trends Cell Biol.*, 26(3):227–237, March 2016.

[203] Christine Mayr. What are 3 UTRs doing? *Cold Spring Harb. Perspect. Biol.*, 11(10):a034728, October 2019.

[204] Serina M Mazzoni-Putman and Anna N Stepanova. A plant biologist's toolbox to study translation. *Front. Plant Sci.*, 9:873, July 2018.

[205] J McLauchlan, D Gaffney, J L Whitton, and J B Clements. The consensus sequence YGTGTTYY located downstream from the AATAAA signal is required for efficient formation of mRNA 3' termini. *Nucleic Acids Res.*, 13(4):1347–1368, February 1985.

[206] F Meijlink, T Curran, A D Miller, and I M Verma. Removal of a 67-base-pair sequence in the noncoding region of protooncogene fos converts it to a transforming gene. *Proc. Natl. Acad. Sci. U. S. A.*, 82(15):4987–4991, August 1985.

[207] Alexandre Melnikov, Anand Murugan, Xiaolan Zhang, Tiberiu Tesileanu, Li Wang, Peter Rogov, Soheil Feizi, Andreas Gnirke, Curtis G Callan, Jr, Justin B Kinney, Manolis Kellis, Eric S Lander, and Tarjei S Mikkelsen. Systematic dissection and

optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, 30(3):271–277, February 2012.

[208] Kate D Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christopher E Mason, and Samie R Jaffrey. Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. *Cell*, 149(7):1635–1646, June 2012.

[209] A M Michelson and S H Orkin. The 3' untranslated regions of the duplicated human alpha-globin genes are unexpectedly divergent. *Cell*, 22(2 Pt 2):371–377, November 1980.

[210] A D Miller, T Curran, and I M Verma. c-fos protein can induce cellular transformation: a novel mechanism of activation of a cellular oncogene. *Cell*, 36(1):51–60, January 1984.

[211] W Min Jou, G Haegeman, M Ysebaert, and W Fiers. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature*, 237(5350):82–88, May 1972.

[212] T Miyata, T Yasunaga, and T Nishida. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc. Natl. Acad. Sci. U. S. A.*, 77(12):7328–7332, December 1980.

[213] Weipeng Mo, Bo Liu, Hong Zhang, Xianhao Jin, Dongdong Lu, Yiming Yu, Yuelin Liu, Jinbu Jia, Yanping Long, Xian Deng, Xiaofeng Cao, Hongwei Guo, and Jixian Zhai. Landscape of transcription termination in arabidopsis revealed by single-molecule nascent RNA sequencing. *Genome Biol.*, 22(1):322, November 2021.

[214] B D Mogen, M H MacDonald, R Graybosch, and A G Hunt. Upstream sequences other than AAUAAA are required for efficient messenger RNA 3'-end formation in plants. *Plant Cell*, 2(12):1261–1272, December 1990.

[215] B D Mogen, M H MacDonald, G Leggewie, and A G Hunt. Several distinct types of sequence elements are required for efficient mRNA 3' end formation in a pea rbcs gene. *Mol. Cell. Biol.*, 12(12):5406–5414, December 1992.

[216] A Moreira, M Wollerton, J Monks, and N J Proudfoot. Upstream sequence elements enhance poly(a) site efficiency of the C2 complement gene and are phylogenetically conserved. *EMBO J.*, 14(15):3809–3819, August 1995.

[217] L Morello and D Breviario. Plant spliceosomal introns: Not only cut and paste. *Curr. Genomics*, 9(4):227–238, 2008.

[218] Rajiv Movva, Peyton Greenside, Georgi K Marinov, Surag Nair, Avanti Shrikumar, and Anshul Kundaje. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. *PLoS One*, 14(6):e0218073, June 2019.

[219] D Muhlrad, C J Decker, and R Parker. Deadenylation of the unstable mRNA encoded by the yeast MFA2 gene leads to decapping followed by 5'–¿3' digestion of the transcript. *Genes Dev.*, 8(7):855–866, April 1994.

[220] Shingo Nagaya, Kazue Kawamura, Atsuhiko Shinmyo, and Ko Kato. The HSP terminator of arabidopsis thaliana increases gene expression in plant cells. *Plant Cell Physiol.*, 51(2):328–332, February 2010.

[221] F Nagy, G Morelli, R T Fraley, S G Rogers, and N H Chua. Photoregulated expression of a pea rbcs gene in leaves of transgenic plants. *EMBO J.*, 4(12):3063–3068, December 1985.

[222] So Nakagawa, Yoshihito Niimura, Takashi Gojobori, Hiroshi Tanaka, and Kin-Ichiro Miura. Diversity of preferred nucleotide sequences around the translation initiation codon in eukaryote genomes. *Nucleic Acids Res.*, 36(3):861–871, February 2008.

[223] Douglas Kyung Nam, Sanggyu Lee, Guolin Zhou, Xiaohong Cao, Clarence Wang, Terry Clark, Jianjun Chen, Janet D Rowley, and San Ming Wang. Oligo(dT) primer generates a high frequency of truncated cDNAs through internal poly(a) priming during reverse transcription. *Proc. Natl. Acad. Sci. U. S. A.*, 99(9):6152–6156, April 2002.

[224] Reena Narsai, Katharine A Howell, A Harvey Millar, Nicholas O'Toole, Ian Small, and James Whelan. Genome-wide analysis of mRNA decay rates and their determinants in arabidopsis thaliana. *Plant Cell*, 19(11):3418–3436, November 2007.

[225] Barbara J Natalizio and Susan R Wente. Postage for the messenger: designating routes for nuclear mRNA export. *Trends Cell Biol.*, 23(8):365–373, August 2013.

[226] S B Needleman and C D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, 48(3):443–453, March 1970.

[227] Jonathan Neve, Radhika Patel, Zhiqiao Wang, Alastair Louey, and André Martin Furger. Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biol.*, 14(7):865–890, July 2017.

[228] J L Nichols. N6-methyladenosine in maize poly(a)-containing RNA. *Plant Sci. Lett.*, 15(4):357–361, August 1979.

[229] Scott J Nicholson and Vibha Srivastava. Transgene constructs lacking transcription termination signal induce efficient silencing of endogenous targets in arabidopsis. *Mol. Genet. Genomics*, 282(3):319–328, September 2009.

[230] Panos Oikonomou, Hani Goodarzi, and Saeed Tavazoie. Systematic identification of regulatory elements in conserved 3' UTRs of human transcripts. *Cell Rep.*, 7(1):281–292, April 2014.

[231] Fatih Ozsolak, Philipp Kapranov, Sylvain Foissac, Sang Woo Kim, Elane Fishilevich, A Paula Monaghan, Bino John, and Patrice M Milos. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143(6):1018–1029, December 2010.

[232] Joseph M Paggi and Gill Bejerano. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. *RNA*, 24(12):1647–1658, December 2018.

[233] Matthew T Parker, Katarzyna Knop, Anna V Sherwood, Nicholas J Schurch, Katarzyna Mackinnon, Peter D Gould, Anthony Hall, Geoffrey J Barton, and Gordon G Simpson. Nanopore direct RNA sequencing maps an arabidopsis N6 methyladenosine epitranscriptome. July 2019.

[234] Matthew T Parker, Katarzyna Knop, Anna V Sherwood, Nicholas J Schurch, Katarzyna Mackinnon, Peter D Gould, Anthony Jw Hall, Geoffrey J Barton, and Gordon G Simpson. Nanopore direct RNA sequencing maps the complexity of arabidopsis mRNA processing and m6a modification. *Elife*, 9, January 2020.

[235] Lori A Passmore and Jeff Coller. Roles of mRNA poly(a) tails in regulation of eukaryotic gene expression. *Nat. Rev. Mol. Cell Biol.*, 23(2):93–106, February 2022.

[236] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and Others. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, 32, 2019.

[237] Rupali P Patwardhan, Choli Lee, Oren Litvin, David L Young, Dana Pe'er, and Jay Shendure. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.*, 27(12):1173–1175, December 2009.

[238] Molly Perchlik, Alexander Sasse, Sara Mostafavi, Stanley Fields, and Josh T Cuperus. Impact of random 50-base sequences inserted into an intron on splicing in*Saccharomyces cerevisiae*. June 2023.

[239] Ariadna Peremarti, Richard M Twyman, Sonia Gómez-Galera, Shaista Naqvi, Gemma Farré, Maite Sabalza, Bruna Miralpeix, Svetlana Dashevskaya, Dawei Yuan, Koreen

Ramessar, Paul Christou, Changfu Zhu, Ludovic Bassie, and Teresa Capell. Promoter diversity in multigene transformation. *Plant Mol. Biol.*, 73(4-5):363–378, July 2010.

[240] Ana Pérez-González and Elena Caro. Effect of transcription terminator usage on the establishment of transgene transcriptional gene silencing. *BMC Res. Notes*, 11(1):511, July 2018.

[241] R P Perry and D E Kelley. Existence of methylated messenger RNA in mouse L cells. *Cell*, 1(1):37–42, January 1974.

[242] G Pesole, F Mignone, C Gissi, G Grillo, F Licciulli, and S Liuni. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, 276(1-2):73–81, October 2001.

[243] N J Proudfoot. Sequence analysis of the 3' non-coding regions of rabbit alpha- and beta-globin messenger RNAs. *J. Mol. Biol.*, 107(4):491–525, November 1976.

[244] N J Proudfoot and G G Brownlee. 3' non-coding region sequences in eukaryotic messenger RNA. *Nature*, 263(5574):211–214, September 1976.

[245] N J Proudfoot and J I Longley. The 3' terminal sequences of human alpha and beta globin messenger RNAs: comparison with rabbit globin messenger RNA. *Cell*, 9(4 PT 2):733–746, December 1976.

[246] Nick Proudfoot. New perspectives on connecting messenger RNA 3' end formation to transcription. *Curr. Opin. Cell Biol.*, 16(3):272–278, June 2004.

[247] Nick J Proudfoot. Ending the message: poly(a) signals then and now. *Genes Dev.*, 25(17):1770–1782, September 2011.

[248] Nick J Proudfoot, Andre Furger, and Michael J Dye. Integrating mRNA processing with transcription. *Cell*, 108(4):501–512, February 2002.

[249] Paola Punzo, Stefania Grillo, and Giorgia Batelli. Alternative splicing in plant abiotic stress responses. *Biochem. Soc. Trans.*, 48(5):2117–2126, October 2020.

[250] Cheng Qin, Nongnong Shi, Mei Gu, Hang Zhang, Bin Li, Jiajia Shen, Atef Mohammed, Eugene Ryabov, Chunyang Li, Huizhong Wang, Yule Liu, Toba Osman, Manu Vatish, and Yiguo Hong. Involvement of RDR6 in short-range intercellular RNA silencing in nicotiana benthamiana. *Sci. Rep.*, 2:467, June 2012.

[251] Qiudeng Que, Mary-Dell M Chilton, Cheryl M de Fontes, Chengkun He, Michael Nuccio, Tong Zhu, Yuexuan Wu, Jeng S Chen, and Liang Shi. Trait stacking in transgenic crops: challenges and opportunities. *GM Crops*, 1(4):220–229, 2010.

[252] Michal Rabani, Lindsey Pieper, Guo-Liang Chew, and Alexander F Schier. A massively parallel reporter assay of 3' UTR sequences identifies in vivo rules for mRNA degradation. *Mol. Cell*, 70(3):565, May 2018.

[253] Anireddy S N Reddy, Yamile Marquez, Maria Kalyna, and Andrea Barta. Complexity of the alternative splicing landscape in plants. *Plant Cell*, 25(10):3657–3683, October 2013.

[254] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11:129, March 2010.

[255] William A Ricci, Zefu Lu, Lexiang Ji, Alexandre P Marand, Christina L Ethridge, Nathalie G Murphy, Jaclyn M Noshay, Mary Galli, María Katherine Mejía-Guerra, Maria Colomé-Tatché, Frank Johannes, M Jordan Rowley, Victor G Corces, Jixian Zhai, Michael J Scanlon, Edward S Buckler, Andrea Gallavotti, Nathan M Springer, Robert J Schmitz, and Xiaoyu Zhang. Widespread long-range cis-regulatory elements in the maize genome. *Nat Plants*, 5(12):1237–1249, December 2019.

[256] Lewis F Richardson. *Weather Prediction by Numerical Process*. University Press, 1922.

[257] L J Richter, Y Thanavala, C J Arntzen, and H S Mason. Production of hepatitis B surface antigen in transgenic plants for oral immunization. *Nat. Biotechnol.*, 18(11):1167–1171, November 2000.

[258] Daniel Rodríguez-Leal, Zachary H Lemmon, Jarrett Man, Madelaine E Bartlett, and Zachary B Lippman. Engineering quantitative trait variation for crop improvement by genome editing. *Cell*, 171(2):470–480.e8, October 2017.

[259] Alexander B Rosenberg, Rupali P Patwardhan, Jay Shendure, and Georg Seelig. Learning the sequence determinants of alternative splicing from millions of random sequences. *Cell*, 163(3):698–711, October 2015.

[260] Remus R E Rosenkranz, Sarah Ullrich, Karin Löchli, Stefan Simm, and Sotirios Fragkostefanakis. Relevance and regulation of alternative splicing in plant heat stress response: Current understanding and future directions. *Front. Plant Sci.*, 13:911277, June 2022.

[261] Emanuel Rosonina, Syuzo Kaneko, and James L Manley. Terminating the transcript: breaking up is hard to do. *Genes Dev.*, 20(9):1050–1056, May 2006.

[262] H M Rothnie. Plant mRNA 3'-end formation. *Plant Mol. Biol.*, 32(1-2):43–61, October 1996.

[263] H M Rothnie, J Reid, and T Hohn. The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3-end formation in plants. *EMBO J.*, 13(9):2200–2210, May 1994.

[264] Paul J Sample, Ban Wang, David W Reid, Vlad Presnyak, Iain J McFadyen, David R Morris, and Georg Seelig. Human 5' UTR design and variant effect prediction from a massively parallel translation assay. *Nat. Biotechnol.*, 37(7):803–809, July 2019.

[265] H Sanfaçon, P Brodmann, and T Hohn. A dissection of the cauliflower mosaic virus polyadenylation signal. *Genes Dev.*, 5(1):141–149, January 1991.

[266] F Sanger, J E Donelson, A R Coulson, H Kössel, and D Fischer. Use of DNA polymerase I primed by a synthetic oligonucleotide to determine a nucleotide sequence in phage fl DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 70(4):1209–1213, April 1973.

[267] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.*, 74(12):5463–5467, December 1977.

[268] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology*, 24:104–108, 1992.

[269] Kengo Sato, Manato Akiyama, and Yasubumi Sakakibara. RNA secondary structure prediction using deep learning with thermodynamic integration. *Nat. Commun.*, 12(1):941, February 2021.

[270] Andrew Savinov, Benjamin M Brandsen, Brooke E Angell, Josh T Cuperus, and Stanley Fields. Effects of sequence motifs in the yeast 3' untranslated region determined from massively parallel assays of random sequences. *Genome Biol.*, 22(1):293, October 2021.

[271] Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342, May 2011.

[272] Z Schwarz-Sommer, L Leclercq, E Göbel, and H Saedler. Cin4, an insert altering the structure of the A1 gene in zea mays, exhibits properties of nonviral retrotransposons. *EMBO J.*, 6(13):3873–3880, December 1987.

[273] Ankeeta Shah, Briana E Mittleman, Yoav Gilad, and Yang I Li. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biol.*, 22(1):291, October 2021.

[274] Ophir Shalem, Eilon Sharon, Shai Lubliner, Ifat Regev, Maya Lotan-Pompan, Zohar Yakhini, and Eran Segal. Systematic dissection of the sequence determinants of gene 3' end mediated expression control. *PLoS Genet.*, 11(4):e1005147, April 2015.

[275] Eilon Sharon, Yael Kalma, Ayala Sharp, Tali Raveh-Sadka, Michal Levo, Danny Zeevi, Leeat Keren, Zohar Yakhini, Adina Weinberger, and Eran Segal. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat. Biotechnol.*, 30(6):521–530, May 2012.

[276] A J Shatkin. Methylated messenger RNA synthesis in vitro by purified reovirus. *Proc. Natl. Acad. Sci. U. S. A.*, 71(8):3204–3207, August 1974.

[277] G Shaw and R Kamen. A conserved AU sequence from the 3' untranslated region of GM-CSF mRNA mediates selective mRNA degradation. *Cell*, 46(5):659–667, August 1986.

[278] M D Sheets, S C Ogg, and M P Wickens. Point mutations in AAUAAA and the poly (a) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro. *Nucleic Acids Res.*, 18(19):5799–5805, October 1990.

[279] Lisha Shen, Zhe Liang, Chui Eng Wong, and Hao Yu. Messenger RNA modifications in plants. *Trends Plant Sci.*, 24(4):328–341, April 2019.

[280] Yingjia Shen, Guoli Ji, Brian J Haas, Xiaohui Wu, Jianti Zheng, Greg J Reese, and Qingshun Quinn Li. Genome level analysis of rice mRNA 3-end processing signals and alternative polyadenylation. *Nucleic Acids Res.*, 36(9):3150–3161, April 2008.

[281] Alexander Sherstnev, Céline Duc, Christian Cole, Vasiliki Zacharaki, Csaba Hornyik, Fatih Ozsolak, Patrice M Milos, Geoffrey J Barton, and Gordon G Simpson. Direct sequencing of arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. *Nat. Struct. Mol. Biol.*, 19(8):845–852, August 2012.

[282] Harshraj Shinde, Ambika Dudhate, Ulhas S Kadam, and Jong Chan Hong. RNA methylation in plants: An overview. *Front. Plant Sci.*, 14:1132959, March 2023.

[283] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, July 2017.

[284] Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. October 2018.

[285] Xiaomin Si, Huawei Zhang, Yanpeng Wang, Kunling Chen, and Caixia Gao. Manipulating gene translation in plants by CRISPR–Cas9-mediated genome editing of upstream open reading frames. *Nat. Protoc.*, 15(2):338–363, January 2020.

[286] David A Siegel, Olivier Le Tonqueze, Anne Biton, Noah Zaitlen, and David J Erle. Massively parallel analysis of human 3' UTRs reveals that AU-rich element length and registration predict mRNA destabilization. *G3*, 12(1), January 2022.

[287] Gordon G Simpson, Paul P Dijkwel, Victor Quesada, Ian Henderson, and Caroline Dean. FY is an RNA 3' end-processing factor that interacts with FCA to control the arabidopsis floral transition. *Cell*, 113(6):777–787, June 2003.

[288] Stacy D Singer, Kerik D Cox, and Zongrang Liu. Enhancer-promoter interference and its prevention in transgenic plants. *Plant Cell Rep.*, 30(5):723–731, May 2011.

[289] Jaswinder Singh, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.*, 10(1):5407, November 2019.

[290] Ratnesh Singh, Ray Ming, and Qingyi Yu. Comparative analysis of GC content variations in plant genomes. *Trop. Plant Biol.*, 9(3):136–149, September 2016.

[291] Robin P Smith, Leila Taher, Rupali P Patwardhan, Mee J Kim, Fumitaka Inoue, Jay Shendure, Ivan Ovcharenko, and Nadav Ahituv. Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat. Genet.*, 45(9):1021–1028, September 2013.

[292] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: modeling sequencing errors in unique molecular identifiers to improve quantification accuracy. *Genome Res.*, 27(3):491–499, March 2017.

[293] Ashish Kumar Srivastava, Yuming Lu, Gaurav Zinta, Zhaobo Lang, and Jian-Kang Zhu. UTR-Dependent control of gene expression in plants. *Trends Plant Sci.*, 23(3):248–259, March 2018.

[294] Alexander O Subtelny, Stephen W Eichhorn, Grace R Chen, Hazel Sive, and David P Bartel. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, 508(7494):66–71, April 2014.

[295] Alessandra M Sullivan, Andrej A Arsovski, Janne Lempe, Kerry L Bubb, Matthew T Weirauch, Peter J Sabo, Richard Sandstrom, Robert E Thurman, Shane Neph, Alex P Reynolds, Andrew B Stergachis, Benjamin Vernot, Audra K Johnson, Eric Haugen, Shawn T Sullivan, Agnieszka Thompson, Fidencio V Neri, 3rd, Molly Weaver, Morgan Diegel, Sanie Mnaimneh, Ally Yang, Timothy R Hughes, Jennifer L Nemhauser, Christine Queitsch, and John A Stamatoyannopoulos. Mapping and dynamics of regulatory DNA and transcription factor networks in a. thaliana. *Cell Rep.*, 8(6):2015–2030, September 2014.

[296] Jialei Sun, Na He, Longjian Niu, Yingzhang Huang, Wei Shen, Yuedong Zhang, Li Li, and Chunhui Hou. Global quantitative mapping of enhancers in rice by STARR-seq. *Genomics Proteomics Bioinformatics*, 17(2):140–153, April 2019.

[297] Yadong Sun, Yixiao Zhang, Keith Hamilton, James L Manley, Yongsheng Shi, Thomas Walz, and Liang Tong. Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc. Natl. Acad. Sci. U. S. A.*, 115(7):E1419–E1428, February 2018.

[298] Yanqing Sun, Lianguang Shang, Qian-Hao Zhu, Longjiang Fan, and Longbiao Guo. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.*, 27(4):391–401, April 2022.

[299] Naeem H Syed, Maria Kalyna, Yamile Marquez, Andrea Barta, and John W S Brown. Alternative splicing in plants–coming of age. *Trends Plant Sci.*, 17(10):616–623, October 2012.

[300] Emese Xochitl Szabo, Philipp Reichert, Marie-Kristin Lehniger, Marilena Ohmer, Marcella de Francisco Amorim, Udo Gowik, Christian Schmitz-Linneweber, and Sascha Laubinger. Metabolic labeling of RNAs uncovers hidden features and dynamics of the arabidopsis transcriptome. *Plant Cell*, 32(4):871–887, April 2020.

[301] Y Takagaki and J L Manley. RNA recognition by the human polyadenylation factor CstF. *Mol. Cell. Biol.*, 17(7):3907–3914, July 1997.

[302] Yongjun Tan, Xiaohao Yan, Jialei Sun, Jing Wan, Xinxin Li, Yingzhang Huang, Li Li, Longjian Niu, and Chunhui Hou. Genome-wide enhancer identification by massively parallel reporter assay in arabidopsis. *Plant J.*, 116(1):234–250, October 2023.

[303] A Tanaka, S Mita, S Ohta, J Kyozuka, K Shimamoto, and K Nakamura. Enhancement of foreign gene expression by a dicot intron in rice but not in tobacco is correlated with an increased level of mRNA and an efficient splicing of the intron. *Nucleic Acids Res.*, 18(23):6767–6770, December 1990.

[304] Patrick E Thomas, Xiaohui Wu, Man Liu, Bobby Gaffney, Guoli Ji, Qingshun Q Li, and Arthur G Hunt. Genome-wide control of polyadenylation site choice by CPSF30 in arabidopsis. *Plant Cell*, 24(11):4376–4388, November 2012.

[305] Bin Tian and Joel H Graber. Signals for pre-mRNA cleavage and polyadenylation. *Wiley Interdiscip. Rev. RNA*, 3(3):385–396, 2012.

[306] Bin Tian, Jun Hu, Haibo Zhang, and Carol S Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.*, 33(1):201–212, January 2005.

[307] Chenfei Tian, Yixin Zhang, Jianhua Li, and Yong Wang. Benchmarking intrinsic promoters and terminators for plant synthetic biology research. *BioDesign Research*, 2022, 2022.

[308] Wei Tian, Xi Huang, and Xinhao Ouyang. Genome-wide prediction of activating regulatory elements in rice by combining STARR-seq with FACS. *Plant Biotechnol. J.*, 20(12):2284–2297, December 2022.

[309] Jackson Tonnies, Mueth Nicholas Arthur, Sayeh Gorjifard, Jonah Chu, and Christine Queitsch. Scalable transfection of maize mesophyll protoplasts. *Journal of Visualized Experiments*, page (in press), 2023.

[310] Jackson Tonnies, Nicholas A Mueth, Sayeh Gorjifard, Jonah Chu, and Christine Queitsch. Scalable transfection of maize mesophyll protoplasts. *J. Vis. Exp.*, (196), June 2023.

[311] M Tosi, R A Young, O Hagenbüchle, and U Schibler. Multiple polyadenylation sites in a mouse alpha-amylase gene. *Nucleic Acids Res.*, 9(10):2313–2323, May 1981.

[312] Fumiaki Uchiumi. *Gene Expression and Regulation in Mammalian Cells: Transcription Toward the Establishment of Novel Therapeutics*. BoD – Books on Demand, February 2018.

[313] Igor Ulitsky, Alena Shkumatava, Calvin H Jan, Alexander O Subtelny, David Koppstein, George W Bell, Hazel Sive, and David P Bartel. Extensive alternative polyadenylation during zebrafish development. *Genome Res.*, 22(10):2054–2066, October 2012.

[314] Ilya Vainberg Slutskin, Adina Weinberger, and Eran Segal. Sequence determinants of polyadenylation-mediated regulation. *Genome Res.*, 29(10):1635–1647, October 2019.

[315] Ilya Vainberg Slutskin, Shira Weingarten-Gabbay, Ronit Nir, Adina Weinberger, and Eran Segal. Unraveling the determinants of microRNA mediated regulation using a massively parallel reporter assay. *Nat. Commun.*, 9(1):529, February 2018.

[316] A Valsamakis, S Zeichner, S Carswell, and J C Alwine. The human immunodeficiency virus type 1 polyadenylylation signal: a 3' long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylylation. *Proceedings of the National Academy of Sciences*, 88(6):2108–2112, 1991.

[317] Krishnan Venkataraman, Kirk M Brown, and Gregory M Gilmartin. Analysis of a noncanonical poly(a) site reveals a tripartite mechanism for vertebrate poly(a) site recognition. *Genes Dev.*, 19(11):1315–1327, June 2005.

[318] Quentin Vicens, Jeffrey S Kieft, and Olivia S Rissland. Revisiting the closed-loop model and the nature of mRNA 5'-3' communication. *Mol. Cell*, 72(5):805–812, December 2018.

[319] Albrecht G von Arnim, Qidong Jia, and Justin N Vaughn. Regulation of plant translation by upstream open reading frames. *Plant Sci.*, 214:1–12, January 2014.

[320] E Wahle. A novel poly(a)-binding protein acts as a specificity factor in the second phase of messenger RNA polyadenylation. *Cell*, 66(4):759–768, August 1991.

[321] E Wahle. 3'-end cleavage and polyadenylation of mRNA precursors. *Biochim. Biophys. Acta*, 1261(2):183–194, April 1995.

[322] Bo Wang, Michael Regulski, Elizabeth Tseng, Andrew Olson, Sara Goodwin, W Richard McCombie, and Doreen Ware. A comparative transcriptional landscape of maize and sorghum obtained by single-molecule sequencing. *Genome Res.*, 28(6):921–932, June 2018.

[323] Maojun Wang, Pengcheng Wang, Fan Liang, Zhengxiu Ye, Jianying Li, Chao Shen, Liuling Pei, Feng Wang, Jiang Hu, Lili Tu, Keith Lindsey, Daohua He, and Xianlong Zhang. A global survey of alternative splicing in allopolyploid cotton: landscape, complexity and regulation. *New Phytol.*, 217(1):163–178, January 2018.

[324] Po-Hao Wang, Sandeep Kumar, Jia Zeng, Robert McEwan, Terry R Wright, and Manju Gupta. Transcription Terminator-Mediated enhancement in transgene expression in maize: Preponderance of the AUGAAU motif overlapping with Poly(A) signals. *Front. Plant Sci.*, 11, 2020.

[325] Wei Wang, Dong-Hui Fang, Jia Gan, Yi Shi, Hui Tang, Huai Wang, Mao-Zhong Fu, and Jun Yi. Evolutionary and functional implications of 3' untranslated region length of mRNAs by comprehensive investigation among four taxonomically diverse metazoan species. *Genes Genomics*, 41(7):747–755, July 2019.

[326] Wei Wang, Zhi Wei, and Hongzhe Li. A change-point model for identifying 3UTR switching by next-generation RNA sequencing. *Bioinformatics*, 30(15):2162–2170, April 2014.

[327] Xingang Wang, Lyndsey Aguirre, Daniel Rodríguez-Leal, Anat Hendelman, Matthias Benoit, and Zachary B Lippman. Dissecting cis-regulatory control of quantitative trait variation in a plant stem cell circuit. *Nat Plants*, 7(4):419–427, April 2021.

[328] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

[329] Jacob D Washburn, Maria Katherine Mejia-Guerra, Guillaume Ramstein, Karl A Kremling, Ravi Valluru, Edward S Buckler, and Hai Wang. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proc. Natl. Acad. Sci. U. S. A.*, 116(12):5542–5549, March 2019.

[330] J D Watson and F H Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, May 1953.

[331] J D Watson and F H Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[332] Kyle E Watters, Timothy R Abbott, and Julius B Lucks. Simultaneous characterization of cellular RNA structure and function with in-cell SHAPE-Seq. *Nucleic Acids Res.*, 44(2):e12, January 2016.

[333] C M Wei and B Moss. Methylation of newly synthesized viral messenger RNA by an enzyme in vaccinia virus. *Proc. Natl. Acad. Sci. U. S. A.*, 71(8):3014–3018, August 1974.

[334] Michael A White, Connie A Myers, Joseph C Corbo, and Barak A Cohen. Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. *Proc. Natl. Acad. Sci. U. S. A.*, 110(29):11952–11957, July 2013.

[335] E Whitelaw and N Proudfoot. Alpha-thalassaemia caused by a poly(a) site mutation reveals that transcriptional termination is linked to 3' end processing in the human alpha 2 globin gene. *EMBO J.*, 5(11):2915–2922, November 1986.

[336] Marvin Wickens, Philip Anderson, and Richard J Jackson. Life and death in the cytoplasm: messages from the 3 end. *Curr. Opin. Genet. Dev.*, 7(2):220–232, April 1997.

[337] Mary Ann Winters and Mary Edmonds. A Poly(A) polymerase from calf thymus: CHARACTERIZATION OF THE REACTION PRODUCT AND THE PRIMER REQUIREMENT. *J. Biol. Chem.*, 248(13):4763–4768, July 1973.

[338] Yu Mi Woo, Yeonui Kwak, Sim Namkoong, Katla Kristjánsdóttir, Seung Ha Lee, Jun Hee Lee, and Hojoong Kwak. TED-Seq identifies the dynamics of Poly(A) length during ER stress. *Cell Rep.*, 24(13):3630–3641.e7, September 2018.

[339] Hsin-Yen Larry Wu and Polly Yingshan Hsu. A custom library construction method for super-resolution ribosome profiling in arabidopsis. *Plant Methods*, 18(1):115, October 2022.

[340] Jing Wu, Ligeng Ma, and Ying Cao. Alternative polyadenylation is a novel strategy for the regulation of gene expression in response to stresses in plants. *Int. J. Mol. Sci.*, 24(5), March 2023.

[341] L Wu, T Ueda, and J Messing. The formation of mRNA 3'-ends in plants. *Plant J.*, 8(3):323–329, September 1995.

[342] Xiaohui Wu, Guoli Ji, and Qingshun Quinn Li. Prediction of plant mRNA polyadenylation sites. *Methods Mol. Biol.*, 1255:13–23, 2015.

[343] Xiaohui Wu, Man Liu, Bruce Downie, Chun Liang, Guoli Ji, Qingshun Q Li, and Arthur G Hunt. Genome-wide landscape of polyadenylation in arabidopsis provides evidence for extensive alternative polyadenylation. *Proc. Natl. Acad. Sci. U. S. A.*, 108(30):12533–12538, July 2011.

[344] Xiaohui Wu, Tao Liu, Congting Ye, Wenbin Ye, and Guoli Ji. scAPAtrap: identification and quantification of alternative polyadenylation sites from single-cell RNA-seq data. *Brief. Bioinform.*, 22(4), July 2021.

[345] Zheng Xia, Lawrence A Donehower, Thomas A Cooper, Joel R Neilson, David A Wheeler, Eric J Wagner, and Wei Li. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3'-UTR landscape across seven tumour types. *Nat. Commun.*, 5:5274, November 2014.

[346] Zhihao Xia, Yu Li, Bin Zhang, Zhongxiao Li, Yuhui Hu, Wei Chen, and Xin Gao. DeeReCT-PolyA: a robust and generic deep learning method for PAS identification. *Bioinformatics*, 35(14):2371–2379, July 2019.

[347] Shang-Qian Xie, Yue Han, Xiao-Zhou Chen, Tai-Yu Cao, Kai-Kai Ji, Jie Zhu, Peng Ling, and Chuan-Le Xiao. ISOdb: A comprehensive database of Full-Length isoforms generated by Iso-Seq. *Int. J. Genomics Proteomics*, 2018:9207637, November 2018.

[348] Denghui Xing and Qingshun Quinn Li. Alternative polyadenylation and gene expression regulation in plants. *Wiley Interdiscip. Rev. RNA*, 2(3):445–458, 2011.

[349] Sinian Xing, Kunling Chen, Haocheng Zhu, Rui Zhang, Huawei Zhang, Bingbing Li, and Caixia Gao. Fine-tuning sugar content in strawberry. *Genome Biol.*, 21(1):230, September 2020.

[350] Tun Xu, Xiaofei Yang, Yanyan Jia, Zihang Li, Guangbo Tang, Xiujuan Li, Bo Wang, Tingjie Wang, Jiadong Lin, Li Guo, and Kai Ye. A global survey of the transcriptome of the opium poppy (papaver somniferum) based on single-molecule long-read isoform sequencing. *Plant J.*, 110(2):607–620, April 2022.

[351] Chenxiao Xue, Fengti Qiu, Yuxiang Wang, Boshu Li, Kevin Tianmeng Zhao, Kunling Chen, and Caixia Gao. Tuning plant phenotypes by precise, graded downregulation of gene expression. *Nat. Biotechnol.*, March 2023.

[352] D Yaffe, U Nudel, Y Mayer, and S Neuman. Highly conserved sequences in the 3' untranslated region of mRNAs coding for homologous proteins in distantly related species. *Nucleic Acids Res.*, 13(10):3723–3737, May 1985.

[353] Shotaro Yamasaki, Atsunobu Suzuki, Yasuaki Yamano, Harunori Kawabe, Daishin Ueno, Taku Demura, and Ko Kato. Identification of 5'-untranslated regions that function as effective translational enhancers in monocotyledonous plant cells using a novel method of genome-wide analysis. *Plant Biotechnol.*, 35(4):365–373, December 2018.

[354] Youli Yao, Luhua Song, Yael Katz, and Gad Galili. Cloning and characterization of arabidopsis homologues of the animal CstF complex that regulates 3' mRNA cleavage and polyadenylation. *J. Exp. Bot.*, 53(378):2277–2278, November 2002.

[355] Valeria Yartseva, Carter M Takacs, Charles E Vejnar, Miler T Lee, and Antonio J Giraldez. RESA identifies mRNA-regulatory sequences at high resolution. *Nat. Methods*, 14(2):201–207, February 2017.

[356] Congting Ye, Yuqi Long, Guoli Ji, Qingshun Quinn Li, and Xiaohui Wu. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics*, 34(11):1841–1849, June 2018.

[357] Congting Ye, Danhui Zhao, Wenbin Ye, Xiaohui Wu, Guoli Ji, Qingshun Q Li, and Juncheng Lin. QuantifyPoly(A): reshaping alternative polyadenylation landscapes of eukaryotes with weighted density peak clustering. *Brief. Bioinform.*, 22(6), November 2021.

[358] Congting Ye, Qian Zhou, Xiaohui Wu, Guoli Ji, and Qingshun Quinn Li. Genome-wide alternative polyadenylation dynamics in response to biotic and abiotic stresses in rice. *Ecotoxicol. Environ. Saf.*, 183:109485, November 2019.

[359] Wenbin Ye, Qiwei Lian, Congting Ye, and Xiaohui Wu. A survey on methods for predicting polyadenylation sites from DNA sequences, bulk RNA-seq, and single-cell RNA-seq. *Genomics Proteomics Bioinformatics*, 21(1):67–83, February 2023.

[360] Xuhong Yu, Pascal G P Martin, and Scott D Michaels. BORDER proteins protect expression of neighboring genes by promoting 3' pol II pausing in plants. *Nat. Commun.*, 10(1):4359, September 2019.

[361] Zhibo Yu, Liwei Hong, and Qingshun Q Li. Signatures of mRNA alternative polyadenylation in arabidopsis leaf development. *Front. Genet.*, 13:863253, April 2022.

[362] Zhibo Yu, Juncheng Lin, and Qingshun Quinn Li. Transcriptome analyses of FY mutants reveal its role in mRNA alternative polyadenylation. *Plant Cell*, 31(10):2332–2352, October 2019.

[363] Wei Zeng, Xinhua Dai, Jing Sun, Yifeng Hou, Xuan Ma, Xiaofeng Cao, Yunde Zhao, and Youfa Cheng. Modulation of auxin signaling and development by polyadenylation machinery. *Plant Physiol.*, 179(2):686–699, February 2019.

[364] Huawei Zhang, Xiaomin Si, Xiang Ji, Rong Fan, Jinxing Liu, Kunling Chen, Daowen Wang, and Caixia Gao. Genome editing of upstream open reading frames enables translational control in plants. *Nat. Biotechnol.*, 36(9):894–898, October 2018.

[365] Jian Zhang, Yi-Zhe Zhang, Jing Jiang, and Cheng-Guo Duan. The crosstalk between epigenetic mechanisms and alternative RNA processing regulation. *Front. Genet.*, 11:998, August 2020.

[366] Jingxian Zhang, Balasubramanyam Addepalli, Kil-Young Yun, Arthur G Hunt, Ruqiang Xu, Suryadevara Rao, Qingshun Q Li, and Deane L Falcone. A polyadenylation factor subunit implicated in regulating oxidative signaling in arabidopsis thaliana. *PLoS One*, 3(6):e2410, June 2008.

[367] Yong Zhang, Lianfeng Gu, Yifeng Hou, Lulu Wang, Xian Deng, Runlai Hang, Dong Chen, Xiansheng Zhang, Yi Zhang, Chunyan Liu, and Xiaofeng Cao. Integrative genome-wide analysis reveals HLP1, a novel RNA-binding protein, regulates plant flowering by targeting alternative polyadenylation. *Cell Res.*, 25(7):864–876, July 2015.

[368] Hongwei Zhao, Denghui Xing, and Qingshun Quinn Li. Unique features of plant cleavage and polyadenylation specificity factor revealed by proteomic studies. *Plant Physiol.*, 151(3):1546–1556, November 2009.

[369] Jing Zhao, Linda Hyman, and Claire Moore. Formation of mRNA 3 ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiol. Mol. Biol. Rev.*, 63(2):405–445, June 1999.

[370] Liangzhen Zhao, Hangxiao Zhang, Markus V Kohnen, Kasavajhala V S K Prasad, Lianfeng Gu, and Anireddy S N Reddy. Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and Nanopore-Based direct RNA sequencing. *Front. Genet.*, 10:253, March 2019.

[371] Wenxue Zhao, Joshua L Pollack, Denitza P Blagev, Noah Zaitlen, Michael T McManus, and David J Erle. Massively parallel functional annotation of 3' untranslated regions. *Nat. Biotechnol.*, 32(4):387–391, April 2014.

[372] Xuelian Zheng, Wei Deng, Keming Luo, Hui Duan, Yongqin Chen, Richard McAvoy, Shuiqing Song, Yan Pei, and Yi Li. The cauliflower mosaic virus (CaMV) 35S promoter sequence alters the level and patterns of activity of adjacent tissue- and organ-specific gene promoters. *Plant Cell Rep.*, 26(8):1195–1203, August 2007.

[373] Andy Zhou, Liam D Kirkpatrick, Izaiah J Ornelas, Lorenzo J Washington, Niklas F C Hummel, Christopher W Gee, Sophia N Tang, Collin R Barnum, Henrik V Scheller, and Patrick M Shih. A suite of constitutive promoters for tuning gene expression in plants. *ACS Synth. Biol.*, 12(5):1533–1545, May 2023.

[374] Jan Zrimec, Christoph S Börlin, Filip Buric, Azam Sheikh Muhammad, Rhongzen Chen, Verena Siewers, Vilhelm Verendel, Jens Nielsen, Mats Töpel, and Aleksej Zelezniak. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. *Nat. Commun.*, 11(1):6141, December 2020.

[375] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, 31(13):3406–3415, July 2003.