UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Characterization of Arabidopsis thaliana Transcription through an Analysis of Nascent Transcripts

Permalink

https://escholarship.org/uc/item/7mw1n7vg

Author

Hetzel, Jonathan

Publication Date

2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA, SAN DIEGO

Characterization of *Arabidopsis thaliana* Transcription through an Analysis of Nascent Transcripts

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Biology

by

Jonathan Andrew Hetzel

Committee in charge:

Professor Joanne Chory, Chair
Professor Mark Estelle
Professor Christopher Glass
Professor Tony Hunter
Professor Julian Schroeder

2016

Copyright

Jonathan Hetzel, 2016

All rights reserved

The Dissertation of Jonathan Andrew Hetzel is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2016

DEDICATION

I dedicate my thesis to my loving parents Kristine and Andrew, who have always supported and encouraged me in all aspects of my life and have given me opportunities beyond anything I could have ever hoped for.

TABLE OF CONTENTS

Signature Page..... iii

Dedication..... iv

Table of Contents..... v

List of Figures and Tables..... vii

Acknowledgements..... ix

Vita..... xi

Abstract of the Dissertation..... xiii

Introduction..... 1

Chapter 1 Nascent RNA sequencing reveals distinct features of plant transcription ……..... 10

1.1 Abstract……..... 10

1.2 Significance Statement..... 11

1.3 Introduction……..... 11

1.4 Results……..... 13

1.5 Discussion……..... 27

1.6 Materials and Methods..... 30

1.7 Acknowledgements…..... 33

1.8 Supplemental Information..... 35

Chapter 2 Identification and Characterization of Pol V Transcripts in *A. thaliana* …... 42

2.1 Introduction..... 42

2.2 Results……..... 44

2.3 Discussion..... 52

2.4 Materials and Methods..... 54

2.5 Acknowledgements……... 57

Chapter 3 Functional Analysis of Critical Residues in the Brassinosteroid Receptor BRI1……..... 58

3.1 Abstract…..... 58

3.2 Introduction……..... 58

3.3 Results…..... 60

3.4 Discussion……..... 69

3.5 Materials and Methods……..... 71

3.6 Acknowledgements…..... 73

3.7 Supplemental Information……..... 74

Perspectives..... 82

References..... 88

LIST OF FIGURES AND TABLES

Figure 1.1 GRO-seq reveals distinct features in *Arabidopsis thaliana* transcription through nascent RNA sequencing…….…….. 15

Figure 1.2 Genome-wide characterization of nascent transcripts in *A. thaliana*..... 19

Figure 1.3 *Arabidopsis* promoter features and motifs…..... 23

Figure 1.4 *Arabidopsis* shows extensive 3' polymerase accumulation..... 26

Figure 1.S1 List of data sets used……..... 35

Figure 1.S2 Divergent transcription is not seen in either corn or *Arabidopsis*, but it is observed in human IMR-90 cells…….……..... 36

Figure 1.S3 Calling transcriptional start sites using 5' GRO-seq and 3-PEAT show strong overlap in signal ……..... 37

Figure 1.S4 RNA stability of IMR-90 cells……..…….... 38

Figure 1.S5 Multiple classes of transcripts are derived from similar promoter structures based on motif analysis……..... 39

Figure 1.S6 Corn 3' pausing correlates with gene length as seen in *Arabidopsis thaliana* but not IMR-90 cells……..... 40

Figure 1.S7 Corn 3' pausing correlates with methylation depletion as seen in *Arabidopsis thaliana* but not IMR-90 cells……..... 41

Figure 2.1Overview of the GRO-seq method..... 45

Figure 2.2 Validation of Pol V transcripts..... 47

Figure 2.3 Length distribution of Pol V transcripts..... 49

Figure 2.4 Pol V transcripts are monophosphorylated..... 50

Figure 2.5 Identification of Pol IV specific transcripts……..... 51

Figure 2.6 Pol V transcript dependence on Pol IV activity……..... 52

Figure 3.1 The glycosylation mutant BRI1^{N154Q}-mCITRINE shows a dwarf phenotype in the *bri1-null* background caused by a lack of protein accumulation and localization.... 61

Figure 3.2 Two BR binding mutant proteins BRI1^{Y597V}-mCITRINE and BRI1^{Y599V}-mCITRINE lead to a semi-dwarf phenotype caused by a reduced binding affinity for BL……..... 65

Figure 3.3 BRI1 residue 727 is critical for coreceptor interaction but not ligand binding..... 68

Table 3.1S List of BRI1 Mutants……..……..... 74

Figure 3.1S. Verification of BRI1 expression by western blot for glycosylation mutants…….... 75

Figure 3.2S. Verification of BRI1 localization for glycosylation mutants by confocal microscopy…….……..... 76

Figure 3.3S. Verification of BRI1 expression by western blot for BR binding and coreceptor binding mutants……..……..... 77

Figure 3.4S. Verification of BRI1 localization for BR binding and coreceptor binding mutants by confocal microscopy……..…….……..... 78

Table 3.2S. BL Binding Assay Kd Values……..……..... 79

Table 3.3S. BRI1 mutagenesis primers……..…..... 80

ACKNOWLEDGEMENTS

Chapter 1, in full, has been submitted for publication as it may appear in *Proceedings of the National Academy of Sciences*. Hetzel, Jonathan A.; Duttke, Sascha H.C.; Liu, Wanlu; Jacobsen, Steve; Chory, Joanne. "Nascent RNA sequencing reveals distinct features in plant transcription", *Proceedings of the National Academy of Sciences* (2013). The author of this dissertation was the primary investigator and author of this paper.

Chapter 2, in part, is being prepared for submission for publication. Hetzel, Jonathan A.; Duttke, Sascha H.C.; Liu, Wanlu; Jacobsen, Steve; Chory, Joanne. “Identification and Characterization of Pol V Transcripts in *A. thaliana*.” The author of this dissertation was the co-primary investigator with Sascha Duttke and Wanlu Liu and will submit as the co-first author of this paper.

Chapter 3, in full, has been prepared by Jonathan Hetzel, Tsegaye Dabi, and Joanne Chory. It does not fit the theme of the rest of the thesis, but is relevant information for members of the brassinosteroid field and therefore has been included here as a standalone chapter. The author of this dissertation was the primary investigator and author of this paper.

Joanne Chory - for taking me into her lab and instilling me with my fundamental training as a scientist, giving me the freedom to grow as an independent researcher, and providing me the resources that no student should be trusted with.

Steve Jacobsen- for first helping introduce me to the world of science and continuing to provide support long after I finished in his lab.

My Committee - Mark Estelle, Tony Hunter, Christopher Glass, and Julian Schroeder for their mentorship and suggestions throughout my PhD.

PBIO - for everyone who has taken an interest in my work and offered ideas and critiques during my time at Salk.

Sascha Duttke and Christopher Benner without whom I would not have been able to accomplish what we did.

Anthony Phan and Sascha Duttke for all of our time together; our house was the best and I can't wait to see what you guys accomplish in the future. Leilani Cruz for her support and pushing me to succeed.

B. Bea and all the members of the SRF- I am very proud of how much we were able to do during my time with the committee and look forward to seeing it continue to grow.

To everyone who I met and worked with during my time in San Diego.

VITA

2007	Bachelor of Science, Harvey Mudd College
2007-2010	Staff Research Associate, University of California, Los Angeles
2010-2016	Graduate Student Researcher, University of California, San Diego
2016	Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Pedmale, U. V., S. S. Huang, M. Zander, B. J. Cole, **J. Hetzel**, K. Ljung, P. A. Reis, P. Sridevi, K. Nito, J. R. Nery, J. R. Ecker and J. Chory. Cryptochromes Interact Directly with PIFs to Control Plant Growth in Limiting Blue Light. Cell 2016 January 14; 164(1-2): 233-245.

Belkhadir Y, Yang L, **Hetzel J**, Dangl JL, Chory J. The growth-defense pivot: Crisis management in plants mediated by LRR-RK surface receptors. Trends Biochem Sci. 2014;39(10):447–456.

Xuehua Zhong, Christopher J. Hale, Minh Nguyen, Israel Ausin, Martin Groth, **Jonathan Hetzel**, Ajay A. Vashisht, Ian R. Hendersona, James A. Wohlschlegel, and Steven E. Jacobsen. DOMAINS REARRANGED METHYLTRANSFERASE3 controls DNA methylation and regulates RNA polymerase V transcript abundance in Arabidopsis. PNAS. 2014 December 11; 112 (3): 911-916.

Du J, Zhong X, Bernatavichute YV, Stroud H, Feng S, Caro E, Vashisht AA, Terragni J, Chin HG, Tu A, **Hetzel J**, Wohlschlegel JA, Pradhan S, Patel DJ, Jacobsen SE. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. Cell. 2012 Sep 28; 151(1):167-80.

Chodavarapu RK, Feng S, Bernatavichute YV, Chen PY, Stroud H, Yu Y, **Hetzel JA**, Kuo F, Kim J, Cokus SJ, Casero D, Bernal M, Huijser P, Clark AT, Krämer U, Merchant SS, Zhang X, Jacobsen SE, Pellegrini M. Relationship between nucleosome positioning and DNA methylation. Nature. 2010 Jul 15; 466(7304):388-92.

Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, **Hetzel J**, Jain J, Strauss SH, Halpern ME, Ukomadu C, Sadler KC, Pradhan S, Pellegrini M, Jacobsen SE. Conservation and divergence of methylation patterning in plants and animals. PNAS. 2010 May 11; 107(19):8689-94.

Wilkes DE, Bennardo N, Chan CW, Chang YL, Corpuz EO, DuMond J, Eboreime JA, Erickson J, **Hetzel J**, Heyer EE, Hubenschmidt MJ, Kniazeva E, Kuhn H, Lum M, Sand A, Schep A, Sergeeva O, Supab N, Townsend CR, Ryswyk LV, Watson HE, Wiedeman AE, Rajagopalan V, Asai DJ. Identification and characterization of dynein genes in Tetrahymena. Methods Cell Biology. 2009; 92:11-30.

ABSTRACT OF THE DISSERTATION

Characterization of *Arabidopsis thaliana* Transcription through an Analysis of Nascent Transcripts

by

Jonathan Andrew Hetzel

Doctor of Philosophy in Biology

University of California, San Diego, 2016

Professor Joanne Chory, Chair

Plant transcription is simultaneously one of the most unique forms of transcription due to the presence of two additional RNA polymerases not found in other eukaryotic species, and one of the least well studied model organisms due to the technical difficulty of isolating and working with plant nuclei. This dissertation focuses on the use of the global run-on sequencing (GRO-seq) technique to study nascent transcripts in *Arabidsopsis thaliana* in order to characterize the distinct aspects of plant transcription and the plant-specific RNA Polymerase V.

By studying nascent transcripts we were able to characterize the core features of plant transcription and identify several interesting characteristics that make plants unique compared to other eukaryotes. Focusing on the promoter, these studies have utilized the whole organism to identify the core promoter of *Arabidopsis*, demonstrate the

unidirectionality of plant transcription, identify the specific DNA motifs found in RNA Polymerase II binding, and that promoter proximal pausing is not found within seedlings. Looking at the termination site of *Arabidopsis* genes, these studies have demonstrated a much more tightly controlled transcriptional termination compared to mammals, and found a distinct feature of plant transcription where polymerase accumulation is observed in the 3' end of genes after the polyA site. Studying non-coding RNAs, we found that plants generally do not utilize non-coding RNAs as derived from RNA Polymerase II, and show very little putative enhancer RNA expression.

Capturing and characterizing the nascent transcripts of the plant specific RNA polymerases has answered fundamental questions of how these polymerases are acting in their roles directing RNA directed DNA methylation and identified the transcripts associated with each. The interdependence of RNA Polymerase IV and V is less connected than previously thought with the identification of a large number of RNA Polymerase V transcripts that are independent of RNA Polymerase IV activity. This result may require a reanalysis of the currently accepted model of RNA Polymerase V recruitment and activity. These results introduce a new method for future studies in the plant biology field and improve the genomic tools available to other researchers.

INTRODUCTION

An individual seed has the ability to take in nutrients, develop into a mature flowering plant, and produce seeds of its own. These aspects of the plant lifecycle are all possible based solely on the coordinated expression of genes at specific points in time. Unsurprisingly, this coordinated expression requires complex cellular machinery that is constantly changing expression based on developmental cues, environmental stimuli, and any number of other factors. The ability to respond to the environment is particularly important for plants as they are rooted in place wherever they grow, and therefore must be able to respond to a host of stimuli. Finding ways to study these changes in gene expression can be difficult in plants since there are well-established techniques for studying steady-state transcript levels, but not nascent transcription. The primary aim of this work was to establish a robust method for analyzing nascent transcription in plants and use that technique to study the unique characteristics of plant transcription.

By 2060 the world population is expected to grow to a staggering 10 billion, with most of this growth occurring in developing countries [1, 2]. The forecasted growth in agricultural yield is currently insufficient to match the needs of the increased population as roughly a 60% total increase in agricultural yield would be required and current methods are unable to meet this growth [1]. This issue is a result of the lack of additional arable land, a smaller labor force engaged in agriculture, and the increased use of crops for animal feed resulting in higher energy use for lower total calories for the individual. To meet the needs of the growing populace without switching the global food source to be purely plant based, more efficient means of production need to be found. Within plant biology that means finding improved agricultural traits to increase overall yield. In order

to find potential targets for improved traits, we need the ability to understand mechanisms of a plant's response to the environment.

To study plants we use the model organism *Arabidopsis thaliana*. *Arabidopsis* is a small tractable member of the mustard family that has a rapid life cycle from an individual seed to producing its own progeny within 2-3 months and requires very few resources for growth. As many researchers have worked with *Arabidopsis* for a significant length of time, there are a wide variety of computation tools, genetic tools, well-characterized methods for performing experiments, and large stocks of mutants from individual screens as well as large-scale T-DNA insertion projects [3, 4]. Additionally, *Arabidopsis* shares highly conserved pathways with more agriculturally relevant crop species, making it easy to establish signaling models in *Arabidopsis* and then verify them in crops for further applied research [5]. Working directly with crops can be difficult with the longer life cycles, higher resource needs, and larger genomes. With all of these resources and the large community surrounding *Arabidopsis* biology, this organism is an excellent starting point for understanding transcription in plants and eventually applying this knowledge to more complex questions such as how *Arabidopsis* responds to stress.

Gene expression is one of the fundamental cellular processes of life and is responsible for an organism's ability to develop and adapt to its changing environment. Gene expression is controlled by two fundamental processes including the rate at which a transcript is being produced and the stabilization of that transcript [6-8]. Transcription rate is a result of initiation, elongation, and termination while stabilization is a result of processing such as polyadenylation, splicing, export, and addition of the 5' cap. If any of these processes are affected, there can be dramatic differences in the stead state level of a transcript and ultimately a change in protein accumulation and cellular function [6-8].

We therefore need to study each of these processes in order to understand gene expression mechanisms.

Generally in eukaryotic systems transcription requires a sequence-specific transcription factor that will bind to DNA and aid in the recruitment of the core RNA polymerase II machinery along with co-regulators of interactions, chromatin-remodeling factors, and a wide array of enzymes responsible for covalent modifications such as methylation and acetylation [9]. This area 50bp upstream and downstream of the transcriptional start site is known as the core promoter and understanding its characteristics and function is arguably the first step in understanding gene expression in any system [10-12]. In general, transcription begins with the formation of the preinitiation complex binding to the core promoter and recruiting the polymerase and additional factors [13-15]. Upon recruitment of the polymerase machinery, transcription can occur and nascent transcripts will be produced. As pol II is responsible for mRNA transcription, Pol II promoter architecture has been extensively studied in other species such as humans, yeast, and bacteria [16-18]. One of the most important features of the promoter is the presence of DNA motifs that are targeted by the transcription machinery [9, 19]. With all of these studies, strong models of how promoter architecture leads to nascent transcripts have been built. How the initiation and termination of transcription occurs within plants is inferred from these other systems, but without directly studying plants, definitive statements cannot be made. In order to characterize transcription in plants, nascent transcripts need to be analyzed through a technique such as PRO-seq [20], NET-seq [21], or global run-on sequencing (GRO-seq) [22].

The main focus of this research has been on the development of new tools for the analysis of whole genome nascent RNA expression through the GRO-Seq method. This technique captures nascent RNA that has just been made across the entire

genome, which differs from techniques such as RNA-seq that study steady-state RNA that was made some time previously and is still stable. In general we can use GRO-seq to identify the position, quantity, and orientation of engaged RNA polymerases by performing a run-on of transcriptionally active nuclei with labeled nucleotides [22-24]. These labeled transcripts can then be sequenced using Illumina sequencing platforms, providing large high-resolution data sets of nascent transcripts. This technique has been performed in other systems, primarily with cell lines or very simple samples such as embryos, but high quality data in plants and other whole organisms in general was lacking when we started this project. By identifying nascent transcripts we are able to build a more accurate model of which genes are active at a specific point in time as well as characterize nascent transcripts to identify aspects of transcription that would otherwise be hidden.

While the GRO-seq method is very powerful, it is not able to accurately identify the 5' ends of nascent transcripts at single nucleotide resolution as we would have hoped. Therefore, we also worked on developing a 5' GRO-seq [25] method in *Arabidopsis* that would allow for precise transcriptional start site (TSS) mapping by enriching for and sequencing transcripts fragment that have a 5' 7meG cap. This enrichment was previously performed using a tobacco acid pyrophosphatase (TAP) enzyme that would remove the 5' cap and leave behind a monophosphate, which can directly be used for the preparation of sequencing libraries. Unfortunately, this enzyme was no longer commercially available so we also set to work finding a replacement for this step in the procedure.

While a definitive plant core promoter model has not been built due to the lack of effective methods and poorly annotated transcriptional start sites, several groups have worked to define several DNA sequence motifs. Some of the strongest motifs include the

TATA box and an initiator, which is also common in other species as some of the strongest elements of the core promoter [26]. In addition, some less common DNA motifs were identified such as DOF, MYB, and MADS box along with a general enrichment of GA sequence content [27]. Later studies that utilized motif discovery tools based on promoters found Y-patch, CA, and GA elements and proposed these may be similar to the CpG islands that are found in many mammalian promoters allowing for a strong prediction of mammalian TSSs [28-33]. While these studies were very valuable for laying the groundwork of studying plant promoter architecture, they all suffered from poor TSS annotations due to the use of relatively low-throughput techniques such as EST/cDNA alignments. Recent work by Morton et al. helped to define TSSs with the development of a 3PEAT method, but unfortunately it still has difficulty calling unstable or non-coding transcripts [26]. As we have been able to build our own TSS maps through direct analysis of TSSs, it has become clear that a more robust method is able to capture far more information about the plant promoter and is able to do it with a much higher resolution and level of confidence. With directly analyzing nascent transcripts, we are also able to capture any unstable or non-coding RNAs and identify new features such as regulatory effects that were not shown previously.

Beyond promoter architecture, the general characteristics of plant transcription are largely unknown. For example, studies of mammalian transcription have revealed a huge number of non-coding enhancer RNAs (eRNAs) that aid in gene regulation [34]. Whether or not eRNAs even exist in plants is still an open question that needs to be addressed. Another example is the 5’ pausing of RNA polymerases in mammalian cells after initiation but prior to elongation allowing for a rapid transcriptional response at specific genes. Whether or not plants show 5’ pausing is still unknown. In general these basic questions have been answered in other systems, but not in plants. Given that

plants and animals diverged more than 1.6 billion years ago, it stands to reason that we can expect to find unique characteristics of plant transcription with an effective method to study nascent transcripts.

Plant transcription is particularly interesting to study due to the presence of two additional eukaryotic RNA polymerases not found in other model systems [35, 36]. A typical eukaryotic system makes use of three DNA-dependent RNA polymerases with RNA polymerase II being responsible for the production of mRNAs [37] and thus the most widely studied. RNA polymerase I is responsible for the transcription of rRNA [38] while RNA polymerase III is responsible for 5s rRNA, tRNA, and other small RNAs [39]. These two polymerases are critical, but as they do not give rise to mRNA and because they are considered to be more related to housekeeping genes, they have generally received less attention than RNA polymerase II. Interestingly within *Arabidopsis* and other flowering plants, there are also two additional polyermases, RNA polymerase IV and RNA polymerase V [35, 36]. These polymerases are the result of duplication events and further evolution of specific Pol II subunits resulting in new unique functions over the course of divergent evolution [40, 41]. The largest subunits of pol IV and pol V are NRPD1 and NRPE1, which differ from the pol II subunit NRPB1 due to substitutions and deletions in the active site as well as the C-terminal domains [35, 40, 42, 43]. These mutations, particularly in the C-terminal domain, are believed to result in the specialized functions of the two plant specific polymerases. While neither polymerase has been studied as extensively as pol II, several groups have identified aspects of their function and identified several transcripts produced by each.

The primary function of pol IV is to serve as the starting point of RNA-directed DNA methlyation (RdDM) in order to repress transposable elements and maintain the integrity of the genome [44, 45]. As a brief overview, Pol IV has been shown to produce

transcripts that are processed into double stranded RNAs by RNA-DEPENDENT RNA POLYMERASE (RDR2) at which point they are then cleaved by DICER-LIKE3 (DCL3) into 24 nucleotide small interfering RNAs (siRNAs) that are responsible for the targeting of the RdDM machinery [36, 46-48]. Although pol IV has been shown to be required for the production of the 24 nucleotide siRNAs and is transcriptionally active *in vitro*, there is still little known about the production of pol IV dependent transcripts [42]. These transcripts are particularly difficult to identify due to their relatively low abundance and the rapid cleavage by DCL3 making it difficult to capture before they are processed and no longer accessible. A recent study was able to use a *dlc2/3/4* mutant to disrupt the processing and sequence the RNAs, but there are several substantial caveats to the method where they had to use a strong mutant background and relied on substantial fragmentation, costing them resolution of the full transcripts [49]. Zhai et al. were instead able to directly clone Pol IV transcripts, thus eliminating many of the caveats of a strong mutant background [50]. Using GRO-seq, it is possible to disrupt the other polymerases to reduce the background signal and increase the relative level of pol IV transcripts for easier identification. We are therefore able to validate the proposed pol IV transcripts with a second high-resolution method and ideally identify new characteristics of these transcripts.

Pol V on the other hand is believed to transcribe non-coding regions of the genome that may then be used as a scaffold for the targeting of the DNA methylation machinery [51]. Unlike pol IV, the transcripts produced by pol V are currently not fully known as they are low abundance and rapidly processed, nor are the modifications of those transcripts characterized. However, ChIP-seq studies have shown that pol V itself maps to transposons and repeats that are linked to the 24 nucleotide siRNAs of pol IV as well as sites of cytosine methylation, indicating a recruitment and binding preference

for the sites of RdDM [52-54]. Additionally, it has been shown that pol V transcription is aided by the DDR complex, consisting of several chromatin remodeling factors that target it to sites of RdDM and can disrupt pol V function [55]. Overall, while many studies have focused on this unique plant specific polymerase pol V, little is known about the actual transcripts it produces and few have been mapped. We believed that a modified GRO-seq method could be the key tool to build a genome wide map of their transcripts and their characteristics such as 5' modifications. In order to carry out all of these studies several substantial plant-specific technical hurdles had to be overcome first.

The GRO-seq method relies on the isolation of high quality, transcriptionally active nuclei. Each run-on typically uses approximately 10^6 nuclei per reaction, and poor quality nuclei are far less efficient. Unfortunately, the isolation of large quantities of transcriptionally active plant nuclei has been difficult for several decades and has been a major factor in setting back the study of plant transcription. While other fields were able to move forward with the study of transcription using nuclear run-on experiments, plants fell behind without the ability to isolate high quality nuclei. This initial challenge needed to be overcome before proceeding with any actual studies using the GRO-seq method. After optimizing the large-scale nuclei prep, the GRO-seq method then had to be optimized for the specific conditions of plant nuclei, which differ dramatically from other species due to the presence of contaminants such as chloroplasts, which are an abundant source of RNA and DNA. The effects of high chloroplast content leads to higher background that needs to be filtered out and less overall useable data, reducing the statistical power of any experiments. By overcoming these technical problems in plants, we can build a much more robust method for our studies as well as a technique that can potentially be used in almost any other species for studying a wide variety of tissue types.

With an efficient GRO-seq method the nascent transcripts can be used to characterize the basic characteristics of plant transcription. GRO-seq is simply a modification of the original nuclear run-on protocol to make it genome wide and able to be performed in conjunction with high-throughput sequencing. The first successful use of GRO-seq was performed by Core et al. to show widespread pausing as well as divergent initiation of transcription in human cell lines [22]. While these features of mammalian transcription were suspected, this study was the first to demonstrate this characteristic with a genome-wide view. With an efficient method in place, we are able to characterize everything from the characteristics of the core promoter to the exact site, location, and modifications of the transcript made from a specific polymerase. Much like the mammalian system, we are able to study characteristics of plants that we suspect but cannot prove and see if characteristics of individual plants can be observed genome-wide.

CHAPTER 1 Nascent RNA sequencing reveals distinct features of plant transcription

1.1 Abstract

Transcriptional regulation of gene expression is a major mechanism utilized by plants to confer phenotypic plasticity, yet compared to archaebacteria, eubacteria, and other eukaryotes, very little is known about the design principles of plant transcription. We generated an extensive catalog of nascent and steady state transcripts in *Arabidopsis thaliana* seedlings using global nuclear run-on sequencing (GRO-seq), 5'GRO-seq and RNA-seq and reanalyzed published corn GRO-seq data to identify general characteristics of plant transcription. *Arabidopsis* nascent transcripts were used for *de novo* annotation of coding and non-coding RNAs, which resulted in the identification of several previously unidentified, unstable transcripts. Both coding and non-coding transcripts had comparable promoter chromatin configurations. Motif analysis revealed new regulatory DNA elements including a novel conserved 'TGT' core promoter motif and unreported transcription factor (TF) binding sites. Mapping of engaged RNA polymerases revealed a lack of enhancer RNAs, promoter proximal pausing, and divergent transcription in *Arabidopsis* and corn, which are commonly present in humans and yeast. In contrast, *Arabidopsis* and corn genes accumulate RNA polymerases adjacent to the polyadenylation site immediately downstream of genes, a trend that correlated with increasing gene length and coincided with hypomethylation of CpG residues. Lack of promoter proximal pausing and a higher correlation of nascent and steady state transcripts suggests Arabidopsis transcript levels are more predominantly regulated by transcriptional initiation than in humans. Together, these findings provide insight into plant transcriptional mechanisms, but also eukaryotic transcription in general.

1.2 Significance Statement

Transcription is a fundamental and dynamic step in the regulation of gene expression but the basic characteristics of transcription in plants are poorly understood. We adapted the GRO-seq and 5'GRO-seq methods for plants and analyzed nascent transcripts in *Arabidopsis thaliana* seedlings. We confirmed known and identified novel transcripts and mapped their transcriptional start sites at single nucleotide resolution. This transcript analysis allowed us to identify distinct characteristics of plant transcription and put forward a general method for mapping engaged polymerases and nascent transcripts genome-wide, a key technique that has been lacking in plants and will undoubtedly prove to be valuable for many members of the plant biology field.

1.3 Introduction

Gene expression is a hallmark of life and subject to adaptation in changing environments Steady state gene expression is a result of RNA synthesis through transcription initiation, elongation, and termination, followed by transcript maturation and decay. Alterations of any of these events can result in dramatic differences in transcript levels and ultimately cellular function. During the past 50 years, much has been learned about transcriptional mechanisms using yeast and animal models [12, 56, 57]. In contrast, owing to technical difficulties created by plant cell extracts, there remains a large gap in our knowledge of the design principles of plant transcription. Plants and animals diverged more than 1.6 billion years ago. Studying plant transcription therefore will not only contribute to a better understanding of the world's largest food source, but also the evolution of eukaryotic gene expression as a whole.

The signals recruiting the transcription apparatus to the genome are ultimately integrated at the promoter. Sequence-specific transcription factors (TFs) commonly bind the proximal promoter around -150 to -50 bp upstream of the transcriptional start site

(TSS) [12]. The core promoter, located approximately ±50 bp relative to the TSS, is the primary site where basal transcription factors cooperate with conserved DNA sequence motifs to orchestrate recruitment of the RNA polymerase [12, 58]. For a number of species, transcription by RNA polymerase II has been studied extensively and detailed models of the promoter have been developed [12, 56-58]. However, this has not been the case for plant model systems. Previous studies in plants have focused on enriched sequences at promoters, but a conclusive view of plant transcription has been hindered by the lack of precise TSSs [27, 33]. While annotation has been improved dramatically through the use of high-throughput sequencing techniques such as 3-PEAT [26], the annotation of many plant TSSs still relies exclusively on original 5'RACE or EST/cDNA alignment data. This problem is particularly apparent for unstable or non-coding RNAs, of which only a fraction have been fully annotated.

In order to comprehensively study global transcription it can be invaluable to map all transcription within the cell, regardless of RNA stability. Nascent RNA sequencing methods such as GRO-seq [22], PRO-seq [20] or NET-seq [21] have highlighted the abundance of unstable transcripts in some eukaryotes such as yeast and mammals [59, 60] yet GRO-seq methods have been difficult to perform in plants due to the required number of intact and transcriptionally active nuclei. GRO-seq was recently employed in corn seedlings and provided insight into monocot transcription, reporting a lack of promoter anti-sense transcription typical in mammalian organisms [61]. Profiling nascent RNA in *Arabidopsis*, will facilitate comparisons between monocot and dicot plants and enable the identification of the key features of plant transcription that apply to the kingdom in general.

Here we report an optimization of the GRO-seq method facilitating the genome-wide analysis of engaged RNA polymerases and nascent transcripts in 6-day old

Arabidopsis seedlings. In the analysis, we focus on 7meG-capped transcripts as generated by RNA polymerase II in order to identify protein-coding genes, miRNAs, and other non-coding RNAs. GRO-seq and 5'GRO-seq were paired with RNA-seq to investigate the *Arabidopsis* transcriptome in depth. In addition to stable transcripts, *de novo* annotation of nascent transcripts revealed many unstable non-coding transcripts, although these transcripts were underrepresented in *Arabidopsis* compared to mammals. Motif analysis identified previously unreported promoter motifs and revealed comparable structures for promoters of coding and non-coding transcripts. Nascent RNA sequencing highlighted the lack of divergent transcripts and promoter proximal pausing, but revealed an elevated 3' pausing for RNA polymerase II that was also apparent in corn. Together, these data affirm distinct features of plant transcription and demonstrate a remarkable diversity in the regulation of eukaryotic transcription.

1.4 Results

Global run-on sequencing profiles of nascent transcription in *Arabidopsis thaliana*

To comprehensively characterize the general features of transcription in plants, we adapted GRO-seq and 5'GRO-seq for use with 6-day old seedlings of *Arabidopsis thaliana* (Fig, 1 *A*, Fig. S1). GRO-seq captures nascent RNA independent of RNA stability thereby providing precise maps of engaged RNA polymerases in an orientation specific manner [22]. 5'GRO-seq specifically enriches for cap-protected 5' ends facilitating TSS mapping of nascent transcripts at single nucleotide resolution [25]. We also profiled steady state transcripts using conventional RNA-seq for comparison with nascent transcript levels. As exemplified for the gene At4g10180, GRO-seq reads cover the full transcript including introns, 5'GRO enriches for 5' fragments of the gene, and RNA-seq maps the mature, intron-less transcript (Fig. 1 *B*). In total we observed active transcription covering approximately 35% of the genome by GRO-seq, and 25% by

RNA-seq in 6-day old *Arabidopsis* seedlings. While this number is in part dependent on sequencing depth, it remarkably differs from humans where approximately 75% of the genome was found to be transcribed across different ENCODE cell lines with no individual line transcribing more than 57% [62] (Fig. 1 *C*).

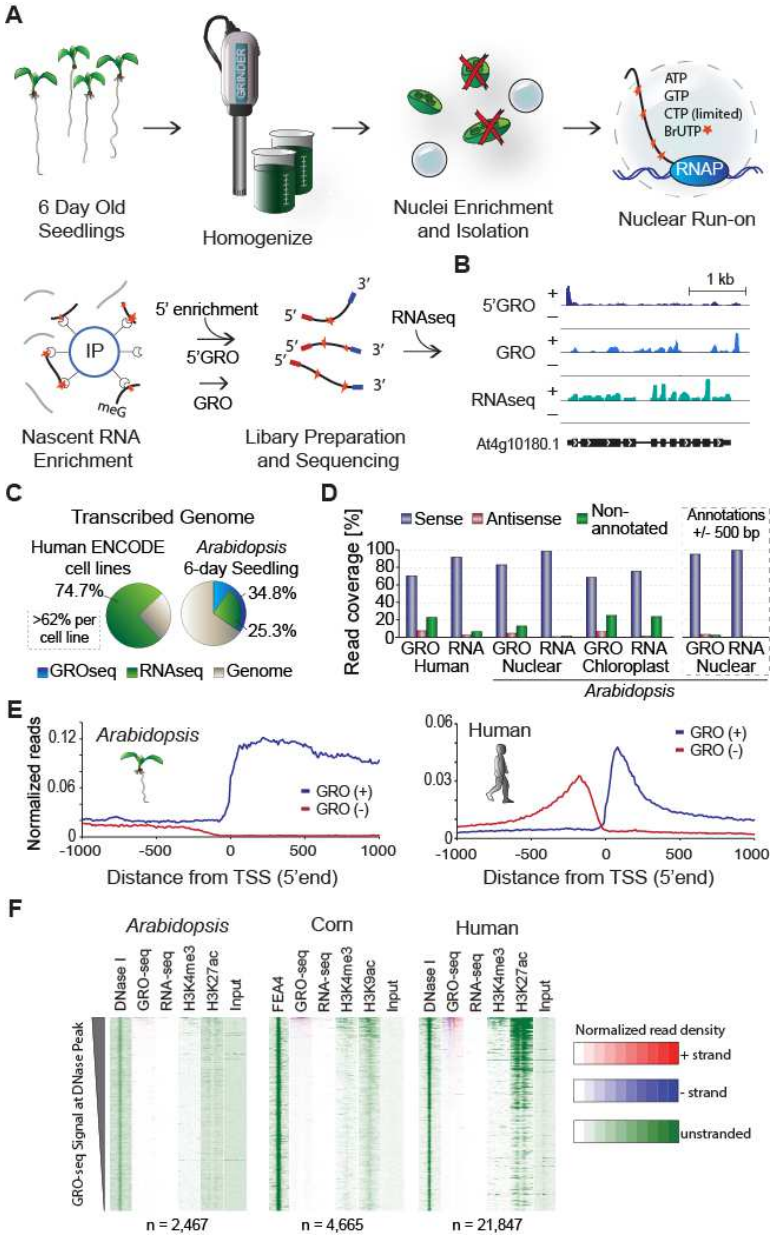


Figure 1.1. GRO-seq reveals distinct features in *Arabidopsis thaliana* transcription through nascent RNA sequencing.
(A) GRO-seq method in *Arabidopsis* (B) Browser shot of sample gene At4g10180.1 with normalized read densities along the y-axis with 5'GRO-seq (dark blue), GRO-seq (light blue), and RNA-seq (green). (C) Percentage of transcribed *Arabidopsis* genome (6 day old seedlings) compared to human ENCODE cell lines [15]. (D) Distribution of RNA-seq and GRO-seq reads relative to annotations or extended annotations (±500bp; right side) with human data from IMR90 cells [8]. (E) Meta plot of GRO-seq signal from annotated genes normalized for reads per bp per gene along y-axis for *Arabidopsis* and human IMR90 cells [8] . (F) Intergenic sites were defined by DNAse-seq peaks (FEA4 ChIP-seq for corn [20]) and heatmaps were generated +/- 1kb from intergenic sites for signal from DNAse-seq, GRO-seq, RNA-seq, H3K4me3, H3K9/27ac, and input in *Arabidopsis*, corn, and IMR-90 cells [8, 13, 19-26]. Sites are sorted based on the total GRO-seq signal observed within 400 bp of the intergenic peak.

In addition to the nuclear and mitochondrial genomes, plants contain a third densely packed chloroplast genome derived from the cyanobacterial lineage. Chloroplast gene expression is controlled by a bacterial type plastid encoded polymerase (PEP) and two nuclear encoded plastid polymerases (NEP) [63]. While we depleted chloroplasts during nuclei isolation and selected against 5'-monophosphorylated RNAs, the organelle is so abundant that a substantial number of nascent transcripts were still captured. We found approximately 76% of the chloroplast genome was actively engaged demonstrating pervasive transcription in this organelle. Comparing GRO-seq to RNA-seq revealed novel transcripts in the chloroplast genome (Fig. 1 D). This result demonstrates the potential for characterizing prokaryotic or viral polymerases using the GRO-seq method even though this was not the goal of our study.

Analysis of GRO-seq data revealed 83% and 68% of engaged RNA polymerases occupy the sense strand of annotated transcripts of the *Arabidopsis* nuclear and chloroplast genomes, respectively. Approximately 4% occupied the antisense strand of genes on the nuclear genome and 13% mapped to unannotated regions. These numbers were significantly higher for the chloroplast genome with 7% and 25% mapping to antisense genic and unannotated regions, respectively (Fig. 1*D*). By comparison, 98% of the RNA-seq mapped to the sense strand of nuclear genes. Together these findings suggest that nuclear RNA polymerases are heavily engaged on the sense strand in *Arabidopsis*, particularly when compared to humans [22]. Indeed, expanding the annotations by 500bp to either side increases the number of nuclear engaged RNA polymerase to 95%, suggesting the majority of non-annotated transcription occur directly adjacent to the annotated TSS and transcription termination site (TTS).

Strikingly, when compared to other organisms such as humans [22] *Arabidopsis* lacks significant divergent transcription as well as 5' pausing at transcription start sites

(Fig. 1 *E*). These results contrast previous reports of divergent transcription as an inherent feature of eukaryotic promoters [22, 64, 65]. Analzying our *Arabidopsis* GRO-seq data with the removal of promoters within 1kb of each other to prevent signal overlap appearing as promoter anti-sense transcription and reanalyzing corn GRO-seq data [61] conclusively demonstrates the lack of divergent transcription and pausing in both systems (Fig. S2). The lack of promoter proximal pausing argues that plant transcription is more predominantly regulated at the level of initiation.

To search for the presence of enhancer RNAs in plants we identified putative enhancer elements using published DNase hypersensitivity (DNase-Seq) data [66] for *Arabidopsis* and FEA4 ChIP-seq [67] peaks for corn (due to a lack of DNAse-seq data). In total 2,467 putative intergenic enhancers were identified in *Arabidopsis* and 4,665 in corn compared to 21,847 in the IMR-90 human cell line. Each site was sorted based on their GRO-seq and RNA-seq signals and heat maps were generated for +/- 1kb from the intergenic site of signal for chromatin accessibility, GRO-seq, RNA-seq, and epigenetic modifications commonly found at either promoters (H3K4me3) or enhancers (H3K9/27ac)(Fig. 1*F*) [22, 61, 66-73]. Based on the heat map, there is very little GRO-seq, RNA-seq, or enhancer associated chromatin marks in *Arabidopsis* and corn compared to humans, and both plants lack the distinctive bidirectional transcription common at mammalian enhancers. Given these data it would appear that if plants have enhancer elements, they rarely, if at all, produce transcripts from these regions.

Nascent Transcript Identification

Unlike RNA-seq, which measures steady-state levels of RNA species, GRO-seq captures nascent transcripts independent of transcript stability [22, 25]. This can be exemplified by the miRNA MIR158A, which responds to nitrogen starvation and has been predicted to target a pair of fucosyltransferase protein transcripts (Fig. 2 *A*) [74].

The annotated miRNA used in previous studies is 100 bp while the actual primary transcript as mapped by GRO-seq is more than 1kb in length and initiates several hundred bp upstream of the current annotation [74]. Additionally, GRO-seq captures transcripts previously undetected by RNA-seq or 3-PEAT (Fig. 2 *A*).

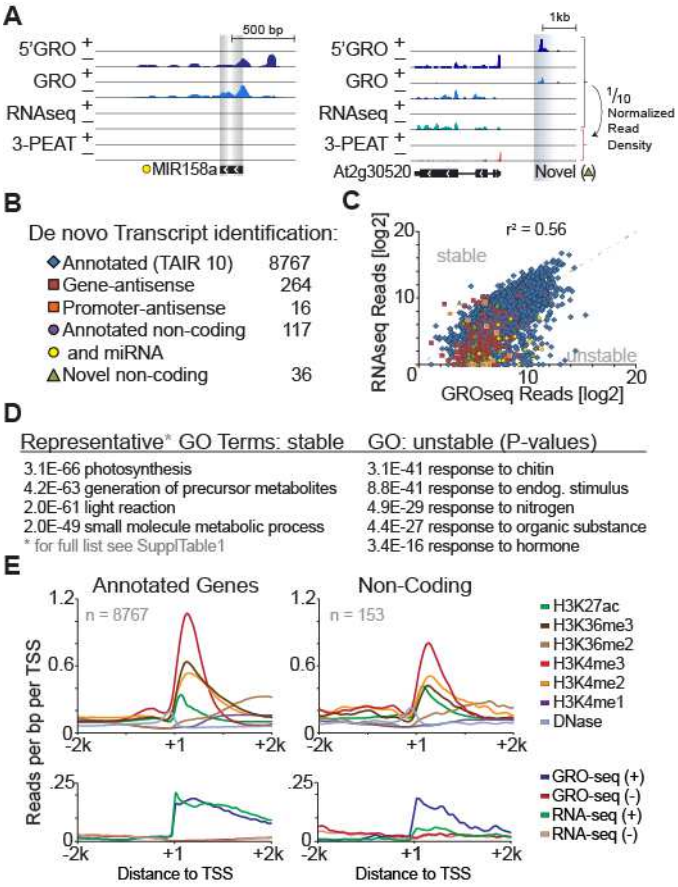


Figure 1.2. Genome-wide characterization of nascent transcripts in *A.thaliana*.
(A) Browser shot for non-coding transcripts (left) of miRNA MIR158a with 5' GRO-seq (dark blue) and GRO-seq (light blue). Relative read densities for 5'GRO/GRO and RNA-seq (green) were scaled by 10% to enable visualization alongside the 3-PEAT TSS mapping data at these loci (red). (B) Classification of *de novo* identified high-confidence transcripts with HOMER. (C) RNA stability plot of [log2] RNA-seq reads vs [log2] GRO-seq reads for identified transcript groups in Fig. 2 *B* (D) Representative list of GO terms with p-values identified for the most stable and unstable transcripts from Fig. 2 *B.* Stable transcripts were defined as having 3-fold more RNA-seq than GRO-seq, while unstable transcripts were defined as having 3-fold more GRO-seq than RNA-seq. (E) Meta plot at the TSS of 8,767 annotated genes and 153 non-coding transcripts from Fig. 2 *B* with normalized reads per bp per TSS for chromatin modifications [19, 24, 26] (top) and with RNA-seq and GRO-seq (bottom).

We next created an unbiased atlas of *Arabidopsis* transcription using *de novo* transcript annotation based on GRO-seq expression loci. We complemented this strategy with 5'GRO-seq to accurately annotate the 5' ends of each transcript. We identified 9200 high confidence transcripts comprised of 8767 annotated protein-coding genes, 264 gene-antisense, 16 promoter-antisense, 117 annotated non-coding, and 36 novel intergenic transcripts (Fig. 2 *B*, Table S1). The number of intergenic transcription

units is significantly less than described in human cell lines [75]. However, gene-antisense RNAs, which were described as modulators of gene expression, are more enriched in *Arabidopsis*, suggesting an increase in antisense gene regulation in plants compared to humans [61, 75, 76]. The TSSs for annotated transcripts called using 5'GRO-seq showed strong overlap to those also found by 3-PEAT [26], increasing the confidence in the data sets (Fig. S3).

Transcript Stability

Comparison of GRO-seq and RNA-seq transcript levels provides a measure for transcript stability. Stable transcripts were defined as transcripts with a low ratio of GRO-seq to RNA-seq signal while transcripts with high ratios of GRO-seq to RNA-seq signal were assigned as unstable transcripts. Plotting the different transcript classes with respect to stability revealed a range of stability for annotated protein-coding genes compared to general instability for non-coding annotated RNAs, miRNAs, promoter-antisense transcripts, and novel non-coding transcripts (Fig. 2 *C*). Performing the same comparison of RNA-seq vs GRO-seq for exon coverage of human IMR-90 cells revealed a higher variance than in *Arabidopsis* (r^2 = 0.56 , Human r^2 = 0.22 , Fig. S4), underlining a much tighter correlation between transcription and steady-state RNA levels in *Arabidopsis*. This correlation suggests *Arabidopsis* transcription is more predominantly regulated at the level of transcription initiation compared to humans.

To investigate if RNA stability was associated with genes with specific biological functions, we performed functional enrichment analysis with annotations from the Gene Ontology (GO) of the most stable and unstable transcripts (Fig. 2 *D*, Table S2). Stable transcripts were associated with cell metabolism and photosynthesis functions while unstable transcripts were enriched for stimulus response genes that are regulated by pathogens, hormones, and other chemicals. These findings are consistent with the

biological theme that transcripts associated with essential processes are stable while regulated genes tend to be less stable.

We next analyzed the epigenetic status of chromatin found at the loci of *de novo* transcripts identified using only GRO-seq data. Analysis of the transcript loci found at annotated protein-coding genes revealed DNase hypersensitivity and transcription-related chromatin marks at the expected position relative to the TSS [77-79] (Fig. 2E). H3K4me3, H3K27ac, and H3K36me3 peaked in proximity to the +1 nucleosome and H3K4me2 slightly downstream [71, 73]. DNase hypersensitivity overlapped with the promoter region, and H3K4me1 and H3K36me2 increase along the body of the gene [66, 71, 73]. The loci for non-coding transcripts showed a nearly identical pattern of chromatin modifications, suggesting they have a similar promoter region to those found at protein-coding gene promoters. The non-coding RNAs showed less RNA-seq signal compared to GRO-seq as expected while annotated protein-coding genes showed similar RNA-seq and GRO-seq signal (Fig. 2 *E*).

***Arabidopsis* promoter structures and identification of promoter motifs**

5'GRO-seq provides high-resolution TSS data allowing for the investigation of promoter elements in a distance-specific manner. Using the 9200 transcripts identified from GRO-seq and 5'GRO-seq data, we found that the *Arabidopsis* core promoter region (-50 to +50 bp with respect to the TSS) shows GC content at 38%. *Arabidopsis* promoters display a strong A/T-enrichment around -30 bp from the TSS suggesting a predominant role of the TATA-box core promoter motif (Fig. 3 *A*). This contrasts human core promoters that are ordinarily GC-rich and only slightly enriched for the TATA box. Over 80% of plant transcripts initiated on an A, followed by 13% G, 5% T and 2% C. *De novo* motif analysis using HOMER underlined the strong prevalence of Inr-like motif (44.8%) and variations thereof (Fig. 3 *B*). The *Arabidopsis* Inr consensus sequences

‘TYA(+1)YYN’ and ‘TYA(+1)GGG’ differ from the traditional Inr ‘TCAKTY’ in *Drosophila* [10]. Also notable was the enrichment of guanosine triplenucleotides in proximity to the TSS but their functional significance remains to be explored (Fig 3B). Some of the elements around the transcripts initiating on ‘C’ matched the reported TCT motif consensus sequence, yet no apparent enrichment in the promoters of genes involved in translation was observed [80].

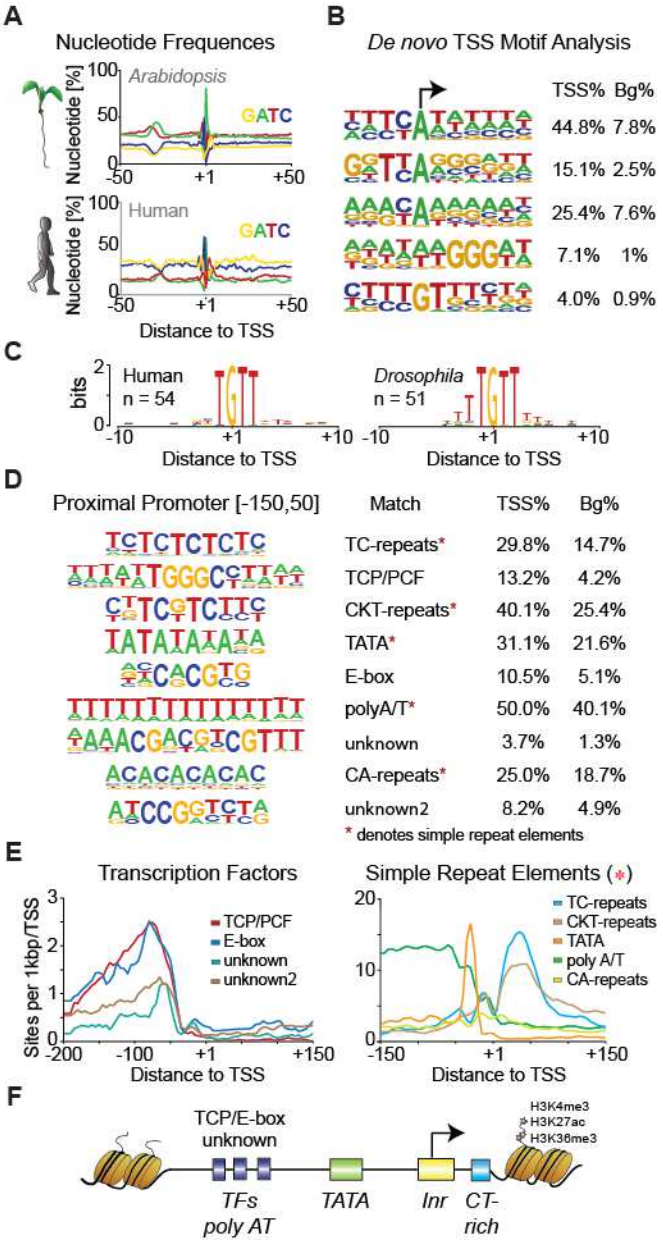


Figure 1.3. ***Arabidopsis*** **promoter features and motifs.**
(A) Meta plot of nucleotide frequency with respect to the +1 TSS as defined by 5'GRO-seq for annotated transcripts at the core promoter region of *Arabidopsis* and human HeLa samples [35]. (B) Position restricted *de novo* motif analysis of *Arabidopsis* initiating nucleotides using HOMER. Percentage of motifs at TSSs as compared to background levels. (C) TGT motif as identified at *Drosophila* S2 and human (HeLa) TSSs [35]. (D) *De novo* motif analysis of the proximal promoter region from -150 to +50 with respect to the TSS using HOMER. Identified motifs with possible matches (left) and percent of TSSs containing the motif (right) along with background levels. (E) Meta plot of the TF binding sites (left) and simple repeat elements (right) with respect to the distance from the TSS along the x-axis. (F) Schematic of regulatory sequence features at *Arabidopsis* promoters.

Our analysis further identified a novel ‘initiator’ which we termed the ‘TGT motif’ at approximately 4% of *Arabidopsis* TSSs (Fig. 3 *B*). Analysis of HeLa [81] and *Drosophila* S2 5’GRO-seq data revealed the TGT motif to be also present in humans and flies (Fig. 3 *C*) where it is rare but conserved.

The region that mediates the intricate interplay among sequence specific transcription factors that modulate gene expression is commonly found in the proximal promoter region, located from -150bp to +50bp relative to the TSS. *De novo* motif analysis using HOMER highlighted a predominance of simple sequence repeats in *Arabidopsis* but also identified transcription factor binding sites, some of which are unknown (Fig. 3 *D,E*) The most prevalent sequence patterns were TC- and CKT-repeats present around the TSS and just downstream. These may resemble 5′-terminal oligopyrimidine tracts (TOPs) which were reported to play a role in mRNA translation [82]. The presence of CT-repeats upstream of the initiator has previously been observed as the ‘Y-Patch’ [33]. CA- repeats were found dispersed across the core promoter region. Simple repeats of polyA/T occur upstream of the TSS with an additional enrichment of ‘TATA’ around -30 bp, with respect to the TSS. The polyA/T-enrichment is known to inhibit nucleosome formation, and aids DNA flexibility and thus transcription factor recruitment [83]. The TATA box was found in 31% of plant promoters, similar to the 29% reported previously [27]. This percentage is higher than in *Drosophila* or Humans and suggests that the -30 and +1 regions are particularly relevant in *Arabidopsis* core-promoters.

The most common transcription factor motifs enriched in Arabidopsis proximal promoters were specifically found upstream of the TATA box with the E-box and TCP/PCF elements being the most prevalent at 10.5% and 13.2%. Additionally, two unidentified motifs were found that have not been reported previously (Fig. 3 *D,E*). An

oversimplified model for the basal *Arabidopsis* promoter structure is proposed (Fig. 3 *F*). Given our classification of coding, non-coding, and anti-sense transcripts, we compared the promoters from each class (Fig. S5). While there were some minor differences, overall the classes contained similar sequence motifs and general sequence composition. Transcriptional stability and transcript properties are thus unlikely to be encoded within the promoter.

3' RNA polymerase Accumulation

Although promoter-proximal pausing in *Arabidopsis* seedlings was not readily apparent, we noticed a sharp transcript termination, possibly due to the compact genome, and an accumulation of polymerase at the 3' polyadenylation site (PAS) of genes when compared to *Drosophila* or humans [22, 84] (Fig. 4 *A*). To confirm the build-up of nascent RNA downstream of the PAS, we plotted published Pol II ChIP-seq [85] to independently characterize the polymerase pausing in this region genome-wide. While the RNA-seq signal goes to zero as expected, Pol II ChIP-seq shows a clear signal increase just downstream of the polyA site (Fig. 4 *B*), in line with previous reports describing a connection between 3'pausing and polyadenylation *in vitro* [84, 86, 87].

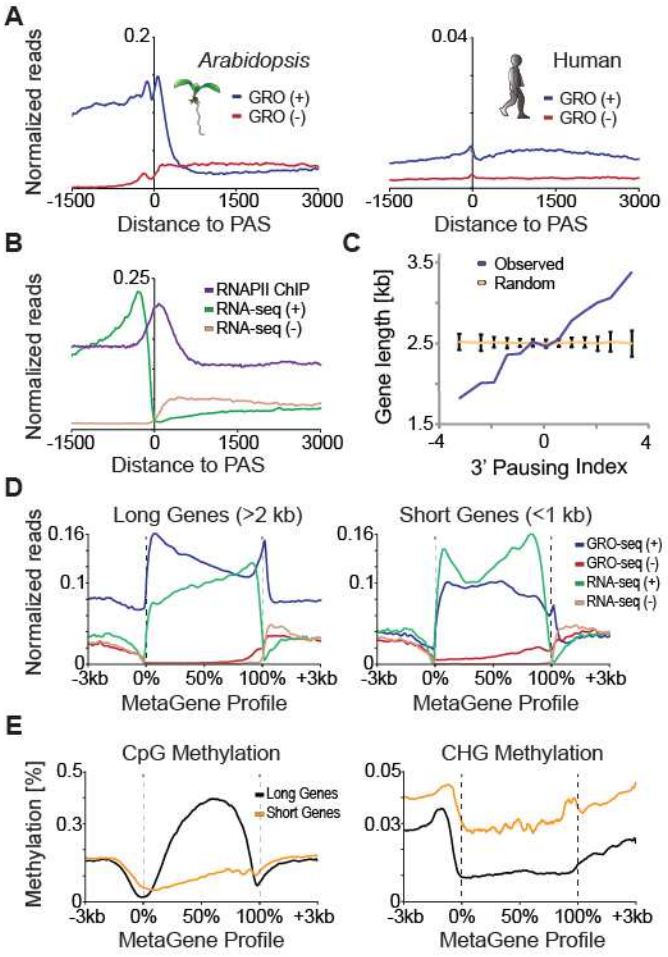


Figure 1.4. *Arabidopsis* shows extensive 3' RNA polymerase accumulation.
(A) Meta plot of GRO-seq signal from Tair10 annotated genes for *Arabidopsis* and human samples [8]. Reads were normalized as reads per bp per gene. (B) RNA polymerase II accumulation as shown by ChIP-Seq (purple) [39] with RNA-seq (+) (green) and (-) (tan) signal around the poly adenlylation site (PAS). (C) 3' pausing index for expressed genes (gene body FPKM>5 and at least ten reads from -250 to 0 relative to the PAS). Index calculated as the ratio of the reads from 0 to +250 relative to the PAS compared to the reads from 0 to -250. Randomized data is shown in tan and standard deviations calculated based on 1000 randomizations. (D) Meta plots anchored by TSS (0%) and PAS (100%) of GRO-seq and RNA-seq for genes >2.0kb (left) and <1kb (right) in total length. Reads were normalized as per bp per gene. (E) CpG methylation (left) and CHG methylation (right) [53] were plotted as percent methylation along the normalized gene body.

We found no defining chromatin marks, polyadenylation signals, or nucleotide frequency differences between paused and unpaused genes, but found an association between 3' pausing and both gene length and CpG methylation (Fig. 4 C, E). A breakdown of genes by length revealed several distinct characteristics between the

GRO-seq signal, RNA-seq signal, and amount of 3' RNA polymerase accumulation (Fig. 4 *D*). Longer genes show a higher accumulation in the GRO-seq reads compared to RNA-seq, suggesting these transcripts are generally less stable than shorter genes. Reanalyzing corn GRO-seq data [61] showed similar strong 3' accumulation of nascent RNA that increased with gene length (Fig. S6) suggesting that the higher level of 3' pausing for longer genes is a distinct characteristic of plant transcription.

In addition to gene length, CpG methylation was found to be associated with 3' pausing. In *Arabidopsis* CpG methylation was excluded from the 3' pausing sites and promoter regions of transcribed genes, while it is only excluded from the promoter region in mammalian systems [70] (Fig. 4 *E*). In aggregate, long genes show an average maximum of CpG body methylation at approximately 40% in the middle of the gene, which drops below 10% at their 5' and 3' ends. However, short genes never show a body methylation peak above 15% and have a comparatively minimal drop at the 3' end demonstrating a strong connection between CpG methylation and 3' pausing. In contrast, CHG methlyation exhibited a distinctly different pattern with a drop at the promoter but no decrease at the pause site, suggesting a specific exclusion of CpG methylation from the PAS in *Arabidopsis*. While a comparable methylC-seq data set does not exist in corn, we were able to reanalyze methyl-DNA immunoprecipitation (MeDIP) [72] to show that methylation is generally excluded from both the 3' and 5' ends of genes as seen in *Arabidopsis*, with a very strong drop at the 3' end of long genes (Fig. S7). The data from corn suggests the connection between 3' pausing of engaged RNA polymerase and CpG methylation is another distinct characteristic of plant transcription.

1.5 Discussion

This study has put forward an application of the GRO-seq method for mapping engaged RNA polymerases by isolating nascent transcripts at a genome-wide level in

Arabidopsis. The identification of nascent transcripts and definition of TSSs revealed unique characteristics of *Arabidopsis* transcription and their connection to other eukaryotic systems. The lack of divergent transcription and promoter proximal pausing in both corn and *Arabidopsis* challenges the notion that eukaryotic promoters are inherently divergent [22, 64, 65]. The absence of promoter proximal pausing, the rate-limiting step after transcriptional initiation in mammals [88], together with higher correlation between nascent and steady state transcript levels argues that transcript levels are more strongly regulated at the level of initiation in plants, revealing distinct differences among Eukaryotes.

Utilizing 5'GRO-seq, we mapped 9200 high confidence TSSs in 6-day old *Arabidopsis* seedlings yet identified only 36 novel and 117 annotated non-coding transcripts. This result suggests plants rely less on non-coding regulatory RNAs as generated by Pol II, particularly when compared to mammals. Notably, plants lack eRNAs as well as the negative elongation factor (NELF) involved in promoter proximal pausing. eRNAs were reported to mediate release of NELF-dependent pausing [89]. Therefore, given the absence of NELF, potential eRNAs may not have provided the same selective advantages in plants, which would impede eRNA or mammalian-like enhancer evolution. In contrast, however, Zhu et al. [73] predicted over 10,000 plant enhancers based on chromatin signatures in leaves and flowers. Without tissue-matched GRO-seq data for these predicted enhancers or targeted disruption it is difficult to validate their *in vivo* role or potential for enhancer transcription, but future work may be able to clarify these results.

Arabidopsis promoters were strongly enriched for Inr-like motifs and the TATA-box with an incidence of over 80% and 30%, respectively. The prevalence of strong core promoter motifs may suggest *Arabidopsis* lacks core promoter diversity compared to

other systems. Plants, despite commonly containing more than one copy of the TATA-box binding protein (TBP), lack TBP-related factors (TRFs) [90]. In bilateral symmetric animals, TRFs were shown to support different transcription systems enabling regulatory diversity through core promoter motif diversity [91, 92]. *Arabidopsis* on the other hand encodes two additional eukaryotic RNA polymerases Pol IV and Pol V, which are integral to the repression of a subset of genes and transposons through RNA-directed DNA methylation [44]. These additional RNA polymerases may reflect a different evolutionary approach to increasing the regulatory diversity of the genome [61].

The improvement in capturing nascent transcripts enabled characterization of the basic features of *Arabidopsis* transcription. RNA polymerase pausing, a major regulator of transcription elongation in mammals [88], was observed predominantly downstream of the PAS in *Arabidopsis* and corn. The underlying mechanisms of RNA polymerase accumulation observed downstream of the PAS is unknown, but it is likely a common feature in plant transcription based on the identical patterns seen in corn GRO-seq. Previous *in vitro* yeast work has proposed that increased pausing downstream of the polyA signal results in increased surveillance time for the mRNA and therefore a higher chance of degradation [84, 86]. This idea may hold true in plants based on the higher GRO-seq signal compared to RNA-seq for longer *Arabidopsis* genes, which also show higher amounts of 3’ pausing compared to shorter genes. How CpG methylation fits in is still unclear given our general lack of understanding of gene body methylation in plants, but it is clearly excluded from both the 5’ and 3’ ends in plants rather than just the 5’ end.

In summary, we have described a method for the analysis of nascent transcription in primary tissue from *Arabidopsis thaliana* and used it to provide a high-resolution map of transcription that has been lacking for plant biology in general. With the identification of distinct characteristics of plant transcription we have opened up new

avenues of study for plant biology and contributed to a better understanding of our own eukaryotic transcription and evolution.

1.6 Materials and Methods

Plant Material and Growth Conditions

The *A. thaliana* accession Columbia (Col-0) was used as the wild type genetic background for experiments. All seeds were sterilized with chlorine gas by mixing 100ml of bleach and 5 ml of concentrated HCl then grown on plates containing half Linsmaier and Skoog Medium (Caisson Laboratories). Plates were placed at 4°C for 3 days for vernalization then placed in growth chambers with 24 hours light and 22°C for 6 days before tissue collection.

Nuclei Isolation

Approximately 20 grams of 6-day-old seedlings were homogenized with an OMNI International General Laboratory Homogenizer in ice cold grinding buffer (300 mM sucrose, 20 mM Tris, pH 8.0, 5 mM $MgCl_2$, 5 mM KCl, 0.2% Triton X-100, 5 mM β-mercaptoethanol, 35% glycerol) at 4°C. Samples were filtered twice through a 250 micron nylon mesh and miracloth before being split into 50ml conical tubes and spun for 10 minutes at 5000g. Supernatant was discarded, and the pellets were resuspended in grinding buffer using a Kimble loose dounce homogenizer (Fisher Scientific). Wash step was repeated twice. Nuclei were resuspended in freezing buffer (50 mM Tris, pH 8.0, 5 mM $MgCl_2$, 20% glycerol, 5 mM β-mercaptoethanol) and frozen in liquid N_2.

GRO-seq and 5'GRO-seq Library Preparation

GRO-seq and 5' GRO-seq were performed similar to described in [22, 25, 81] with modifications as follow. Approximately $5x10^6$ nuclei were used for run-ons at room temperature for 5' and CTP concentration was limited to 20nM. After run-on, samples were treated with RQ1 DNase (Promega) for 20 minutes at 37°C followed by Trizol LS

(Fisher Scientific) extraction. Samples were then treated with Terminator 5'-Phosphate-Dependent Exonuclease (Epicentre) for 1 hour at 30°C. RNA fragmentation was carried out with Fragmentation Reagents kit (Ambion) for 11' at 70°C. Treatments were used to deplete ribosomal and chloroplast RNA. Nascent RNA was enriched by double pull down using BrdU antibody beads (Santa Cruz Biotechnology) for 80', washed three times with GRO-Binding buffer (0.25x SSPE, 0.05% Tween, 37.5 mM NaCl , 1 mM EDTA) then extracted with Trizol LS (Fisher Scientific). Remaining steps were carried out as described previously [81] except de-capping using TAP was replaced with RppH (New England Biolabs) and NEB Next Small RNA Library Prep Set (New England Biolabs) was used for library preparation. De-capping was performed in 3' Ligation buffer at 37°C in the absence of 3' adapter. 3'Adapter ligation was performed at 20°C instead of 25°C so RppH is inactive.

RNA-seq Library Preparation

Total RNA was extracted from seedlings frozen in liquid nitrogen using RNeasy Plant Mini Kit (Qiagen). 10µg of total RNA was used for extraction of mRNA with the Poly(A)Purist MAG Kit (Ambion) according to manufacturer's instructions. Isolated mRNA was fragmented with Fragmentation Reagents (Ambion) for 14'. Fragmented RNA was incubated with T4 polynucleotide kinase (New England Biolabs) in low pH buffer (0.1 M MES, 10 mM MgCl2, 10 mM mercaptoethanol, 300 mM NaCl - pH 5.6) for 3' repair, then T4PNK buffer for 5' phosphorylation. The Small RNA Library Prep Set was (New England Biolabs) was used for library preparation.

Analysis for NGS data

All data generated by this study and previously published data were reanalyzed in a consistent manner as described below. Adapter sequences were trimmed from the 3' ends of all RNA-seq, GRO-seq, and 5'GRO-seq reads and the reads were then

aligned using STAR (v2.4.0k) [93] to the appropriate genome: TAIR10 (*Arabidopsis*), AGPv3 (corn), or hg19 (human). All ChIP-seq and DNase-Seq data was aligned to the appropriate genome using Bowtie 2 [94]. Only reads that aligned to the genome at a single unique location were considered for downstream analysis. All analysis or visualizations of NGS data were normalized to a read depth of 10^7 total uniquely aligned reads per experiment. Genome browser tracks depicting normalized read densities for each experiment were generated using HOMER [95] and visualized in IGV [96]. Meta gene plots, histograms, heatmaps, GO enrichment analysis, and gene expression values containing normalized read densities were generated using HOMER, using gene and feature annotations from TAIR10 (Arabidopsis), Ensembl (release 30/AGPv3, corn), and RefSeq (human).

De novo Transcript Identification

Transcript discovery and annotation were performed using routines in HOMER as described below. Transcription units were identified directly from GRO-seq data using the HOMER program findPeaks (-style groseq with default parameters) as described in Wang et al. [97], which looks for regions of continuous, strand-specific GRO-seq read coverage to identify transcripts. Transcription start sites (TSS) were found using findPeaks (-style tss) as described in Lam et al. [25], which identifies strand specific peaks from 5'GRO-seq data using traditional GRO-seq as a control. Since 5'GRO-seq is only an enrichment for capped protected RNA fragments, only peaks containing 3-fold more 5'GRO-seq reads compared to traditional GRO-seq were considered as bonafide TSS. TSS discovered by 5'GRO-seq were then assigned to GRO-seq transcripts if they were found within 1 kb of the 5' end of the GRO-seq transcript. In cases where multiple TSS could be assigned to the same transcript, the TSS with the highest read density was used. Only de novo identified transcripts that were assigned a valid 5'GRO-seq TSS

were considered in downstream analysis. Transcripts were annotated into different classes as follows: First, transcripts that strand-specifically overlap any known transcript (TAIR10) were assigned the accession number/annotation of the known transcript. Next transcripts that overlapped any known transcript on the opposite strand are assigned as 'antisense transcripts'. Next, transcripts found upstream of an annotated TSS in the opposite strand of the annotated genes were assigned as 'promoter-antisense transcripts'. Transcripts failing to meet any of these critierion were assigned as 'novel' transcripts.

Motif Discovery

De novo motif discovery and calculation of motif positions were performed using HOMER. Two strategies were employed to identify proximal-promoter enriched motifs and Initiator elements, respectively. Proximal promoter motifs were identified by applying de novo motif discovery to the sequence from -150 to +50 bp relative to TSS discovered using 5'GRO-seq, using random regions of the *Arabidopsis* genome as background. In order to identify Initiator motifs, position-restricted motif discovery was performed by specifically analyzing the sequences from -4 to +6 bp relative to the TSS for 11 bp motifs, using randomly selected 11 bp sequences from the surrounding 100 bp of promoter sequences as background.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo

1.7 Acknowledgements

We thank J. Kadonaga and C. Glass for feedback and suggestions. JH was supported by the NIH (T32GM007240), the Rose Hills Foundation, and the H.A. and Mary K. Chapman Charitable Trust. S.H.C.D. is the recipient of the University of California at San Diego Molecular Biology/Cancer Center Fellowship.

Chapter 1, in full, has been submitted for publication as it may appear in *Proceedings of the National Academy of Sciences*. Hetzel, Jonathan A.; Duttke, Sascha H.C.; Liu, Wanlu; Jacobsen, Steve; Chory, Joanne. "Nascent RNA sequencing reveals distinct features in plant transcription", *Proceedings of the National Academy of Sciences* (2013). The author of this dissertation was the primary investigator and author of this paper.

1.8 Supplemental Information

Experiment	Uniquely Alignable Reads
5'GRO-seq	7840925
GRO-seq rep1	20003079
GRO-seq rep2	13659841
RNA-seq rep1	19159534
RNA-seq rep2	17491818
Arabidopsis public data used in this study:	NCBI Acc (GEO/SRA/ENA):
Dnase-Seq	GSE53322
H3K4me3 ChIP-seq	SRA010097
H3K4me2 ChIP-seq	SRA010097
H3K36me3 ChIP-seq	SRA010097
H3K36me2 ChIP-seq	SRA010097
H3K27ac ChIP-seq	SRA174578
H3K4me1 ChIP-seq	SRA174578
methylC-Seq	GSE53033
PEAT (TSS mapping)	SRA171647
Human public data used in this study:	NCBI Acc (GEO/SRA):
GRO-seq (IMR90)	GSE13518
RNA-seq (IMR90)	GSE30567
H3K4me3 ChIP-seq (IMR90)	GSE16256
H3K27ac ChIP-seq (IMR90)	GSE16256
Dnase-Seq (IMR90)	GSE18927
Zea Mays Public Data Used in this Study:	NCBI Acc (GEO/SRA):
GRO-seq	GSE54166
RNA-seq	GSE76939
H3K4me3 ChIP-Seq	GSE15286
H3K36me3 ChIP-seq	GSE15286
H3K9ac ChIP-seq	GSE15286
McrBCSeq (DNA methylation ChIP)	GSE15286

Figure 1.S1 List of Data Sets Used.
Summary of sequencing experiments performed and published data reanalyzed for this study.

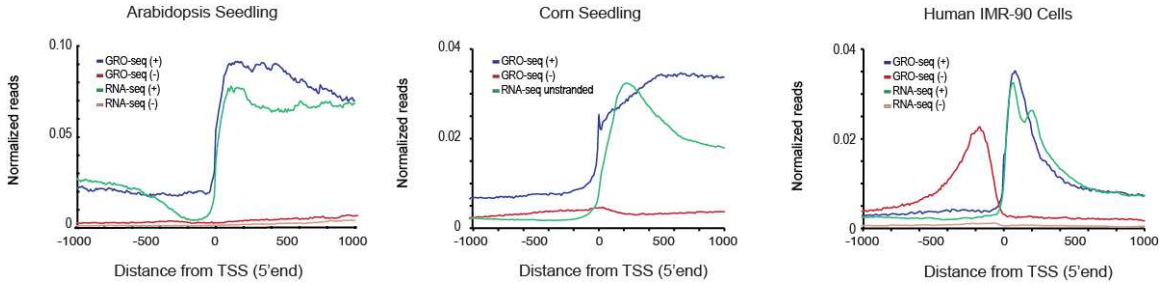


Figure 1.S2 Divergent transcription is not seen in either corn or *Arabidopsis*, but it is observed in human IMR-90 cells.

Meta plot of GRO-seq and RNA-seq signal from unidirectional promoters of annotated genes. Only isolated TSS were the closest TSS for another gene is at least 1 kb away were used. GRO/RNA-seq data is presented normalized for reads per bp per gene along y-axis for *Arabidopsis*, corn, and human IMR-90 cells [8, 13, 22]. Distance from TSS is plotted along x-axis.

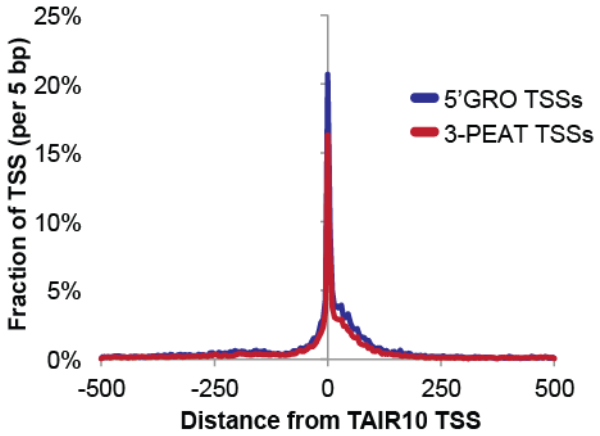


Figure 1.S3 Calling transcriptional start sites using 5' GRO-seq and 3-PEAT show strong overlap in signal.
Distribution of TSSs identified from 5'GRO-seq (blue) and 3-PEAT (red) data relative to the TAIR10 TSS annotation.

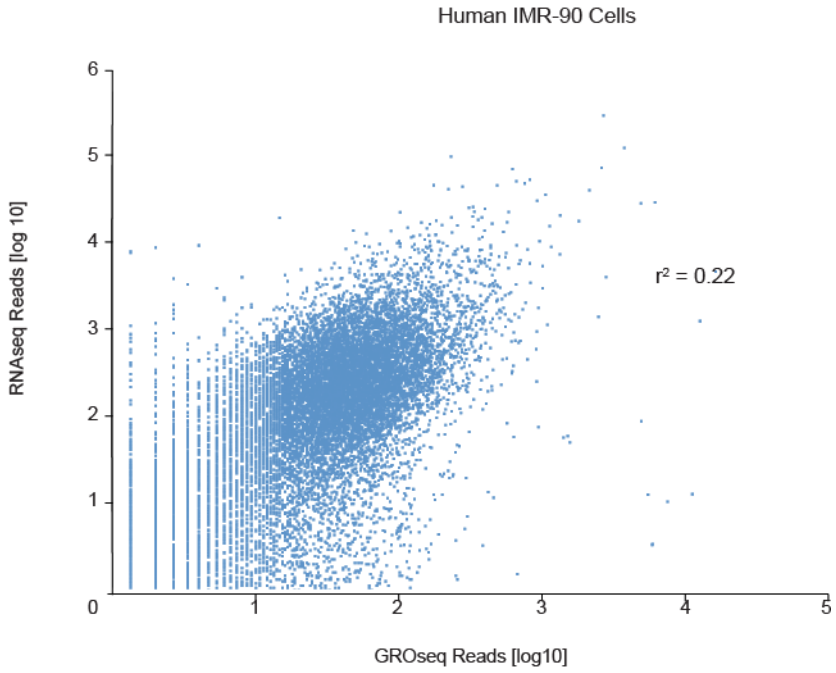


Figure 1.S4 RNA stability of IMR-90 cells.
RNA stability plot of [log10] normalized RNA-seq reads vs [log10] normalized GRO-seq reads for RefSeq genes in IMR-90 cells [8, 22].

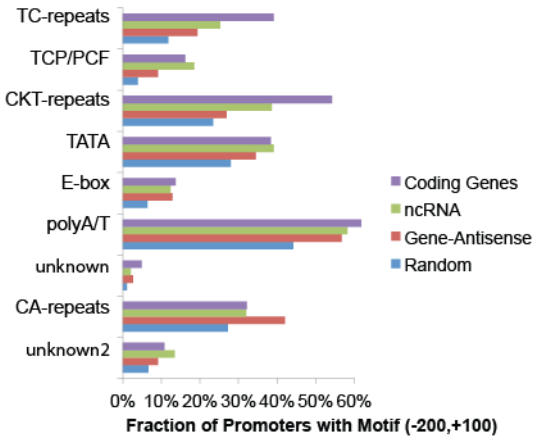


Figure 1.S5 Multiple classes of transcripts are derived from similar promoter structures based on motif analysis.
The percentage of promoters containing each motif in the region 200bp upstream and 100bp downstream of the TSS. Each motif was analyzed for the transcript classes of coding genes (purple), non-coding transcripts (green), and gene-antisense transcripts (red) as well as a random genomic regions as a control (blue).

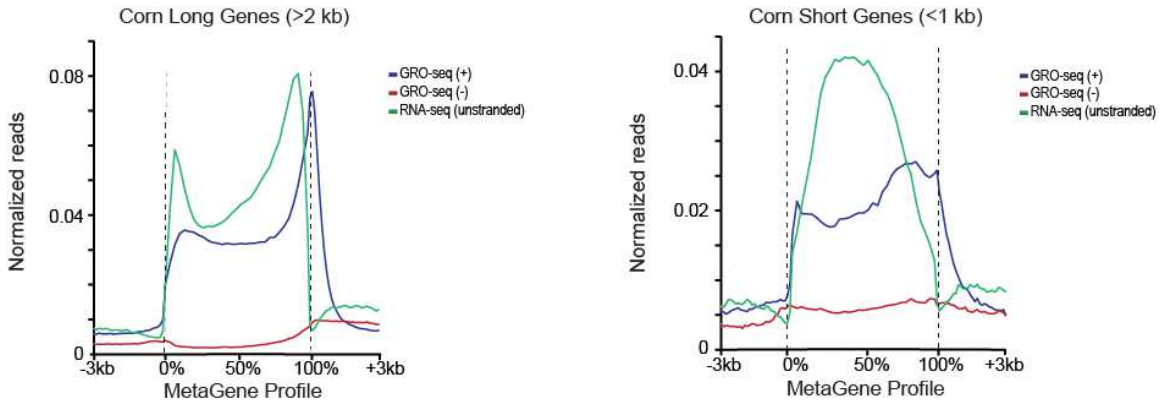


Figure 1.S6 Corn 3' Pausing correlates with gene length as seen in *Arabidopsis thaliana* but not IMR-90 cells.
Meta Gene plots anchored by TSS (0%) and PAS (100%) of corn GRO-seq [13] and RNA-seq (unstranded) for genes >2.0kb (left) and <1kb (right). Reads were normalized per bp per gene.

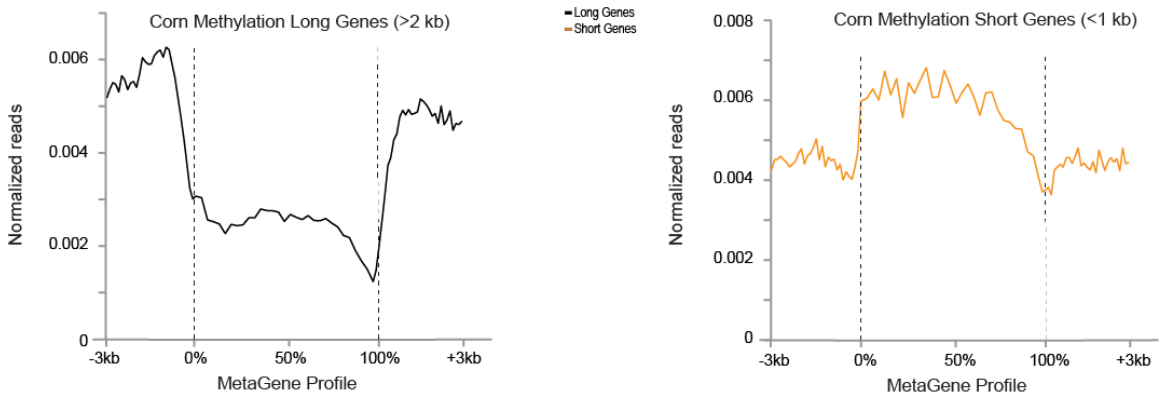


Figure 1.S7 Corn 3' pausing correlates with methylation depletion as seen in *Arabidopsis thaliana* but not IMR-90 cells.
Methylation of corn using DNA methylation ChIP for long genes >2kb (left) and short genes <1kb (right) [25]. Reads were normalized per bp per gene.

CHAPTER 2 Identification and Characterization of Pol V Transcripts in *A. thaliana*

2.1 Introduction

Plant transcription is unique due to the evolution of two additional non-essential eukaryotic RNA polymerases Pol IV and Pol V in addition to the standard three RNA polymerases found in other eukaryotic systems Pol I, Pol II, and Pol III. An analysis of Pol IV and Pol V using mass spectrometry [40, 98, 99] and phylogenetic analysis [36, 41, 100] showed these two polymerases are derived from Pol II with distinct large subunits NRPD1 and NRPE1 as well as three and six additional distinct subunits respectively and are still evolving via a multistep process [101]. It is therefore possible that these two polymerases share some characteristics of Pol II, such as sequence-specific transcription factor binding or preference for specific epigenetic marks along DNA, but without a strong definition of these transcripts, there is still a substantial gap in our understanding of these polymerases.

Functionally, in *Arabidospis thaliana* Pol IV is critical for the synthesis of noncoding RNAs responsible for transcriptional silencing of transposons, repetitive elements, and some genes through RNA directed DNA methylation (RDDM) [35, 36, 44, 46-48, 50]. The current model proposes that Pol IV transcribes precursor RNAs, RNA-DEPENDENT RNA POLYMERASE 2 (RDR2) makes the precursors double stranded, DCL3 (DICER-LIKE 3) dices the double stranded RNAs, and the resulting siRNAs are loaded into ARGONAUTE 4 (AGO4) [36, 42, 46-48]. Recent work by Zhai et al. has mapped the Pol IV dependent RNAs and proposed that they are small enough that there is generally only one precursor RNA per siRNA produced [50]. The identification of Pol IV transcripts is dependent on the disruption of the members of the DCL family, DCL2, DCL3, and DCL4.

While Pol IV generates the preceursor RNAs for RDDM, Pol V is believed to produce noncoding RNAs that serve as a scaffold for the Pol IV siRNAs to target sites for methylation by DOMAIN REARRANGED METHYLTRANSFERASE 2 (DRM2). Pol V transcription is dependent on a complex known as DDR consisting of DEFECTIVE IN RNA-DIRECTED DNA METHYLATION (DRD1), a chromatin remodeling factor, RNA-DIRECTED DNA METHYLATION 1 (RDM1), a protein that binds single stranded DNA, and DMS3 (DEFECTIVE IN MERISTEM SILENCING 3), a cohesin-like protein. After transcription RDM1, Pol V, DRM2, and AGO4, all interact to catalyze DNA methylation [102]. Pol V transcription is believed to be dependent on the activity of Pol IV through the generation of siRNAs and the establishment of DNA methlyation at the sites of interest.

Though we now have a much stronger understanding of Pol IV transcription through the cloning and mapping of Pol IV transcripts by Zhai et al., we are lacking similar knowledge for Pol V transcription. Only a few Pol V dependent transcripts have been identified and we only know that they have been shown to be potentially monophosphorylated [51] and are resistant to alpha-Amanitin treatment, which disrupts Pol II transcription [42]. They are believed to be long and noncoding, but their length has not yet been validated. Additionally, given that the Pol V transcripts are noncoding, traditional methods such as RNA-seq are not effective for their detection.

In this study, we set out to define the fundamental characteristics of Pol V transcripts using a modified version of *Arabidopsis* GRO-seq. We were able to identify Pol V transcripts, demonstrate that these transcripts are on average much shorter than predicted at only 50 bp, show that many Pol V transcripts are not dependent on Pol IV indicating Pol V is less dependent on RDDM than expected, define the 5’ end of Pol V transcripts as being monophosphorylated, demonstrate a strong overlap with Pol V

ChIP-seq data, enrich for Pol V transcripts using alpha-Amanitin treatments, and validate the method by confirming previously published Pol IV data.

2.2 Results

Capture of Pol V RNA transcripts through global run-on sequencing

In order to characterize transcripts produced by Pol V we used a modified GRO-seq method with immature floral tissue from *Arabidopsis thaliana*. GRO-seq captures nascent RNA transcripts, which allows for the identification of noncoding RNAs such as those produced by Pol V (Figure 2.1A). Immature floral tissue was used for GRO-seq library preparation due to the enrichment of Pol IV and Pol V transcripts in those tissues [103]. Treatment with alpha-Amanitin was used to deplete Pol II signal, thus enriching for Pol V transcript signal and allowing for the calling of these transcripts. Pol V ChIP-seq was used for comparison to GRO-seq as an additional way to validate potential Pol V transcripts. GRO-seq libraries were prepared from Col, Col with alpha-Amanitin, *nprd1-4* mutants, and *nrpe1-11* mutants. NRPD1 and NRPE1 encode the largest subunit of Pol IV and Pol V respectively, and their disruption eliminates associated transcripts. A sample browser shot shows Pol V dependent transcripts are captured with GRO-seq, enriched by alpha-Amanitin, overlap with sites previously identified using NRPE1 ChIP-seq data, and eliminated in the *nrpe1* mutant background (Figure 2.1B).

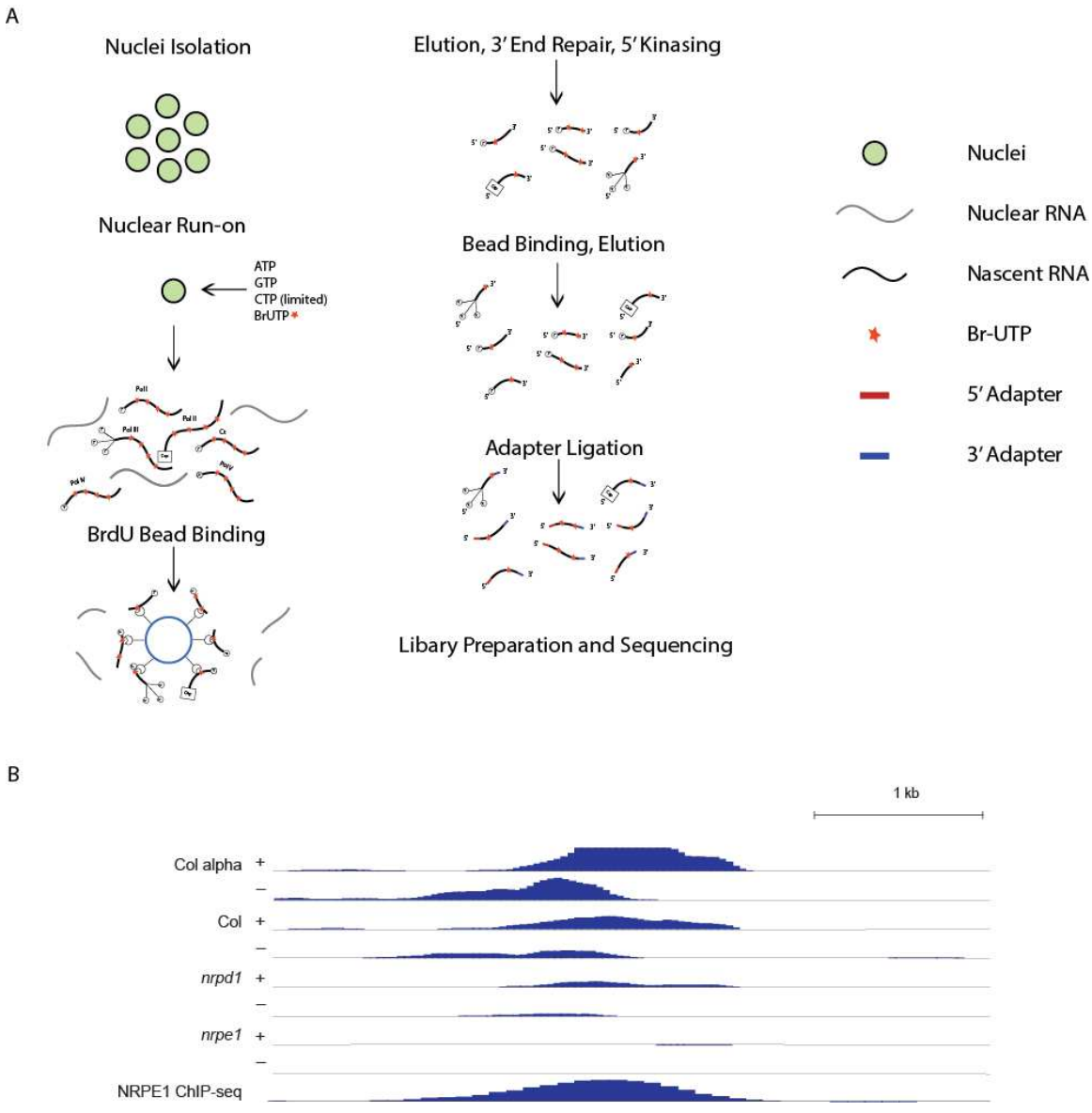


Figure 2.1 Overview of the GRO-seq method.
A) Overview of the modified GRO-seq method that captures monophosphorylated transcripts associated with Pol V. See materials and methods for details. B) Browser shot of sample Pol V specific transcripts with normalized read densities along the y-axis. Samples from top to bottom are Col treated with alpha-Amanitin (+) and (-) strands, Col untreated, *nrpd1* mutant, *nrpe1* mutant, and NRPE1 ChIP-seq.

Globally, we were able to identify 2,607 Pol V peaks based on the enrichment of transcript signal in Col treated with alpha-Amanitin compared to Col with no treatment and their depletion in the *nrpe1* background (Figure 2.2A). Approximately 1600

transcripts were shared by both strands, 570 were plus strand specific, and 425 were minus strand specific. Of the 7.6MB regions covered by Pol V ChIP-seq, 2.3MB of Pol V GRO-seq sites overlapped and 0.5MB did not (Figure 2.2B). Due to the low resolution of ChIP-seq compared to GRO-seq, it is not surprising that the total GRO-seq sites were lower, but the additional 0.5MB not found in ChIP-seq was surprising. Overall, the ChIP-seq served as a validation of the Pol V GRO-seq data.

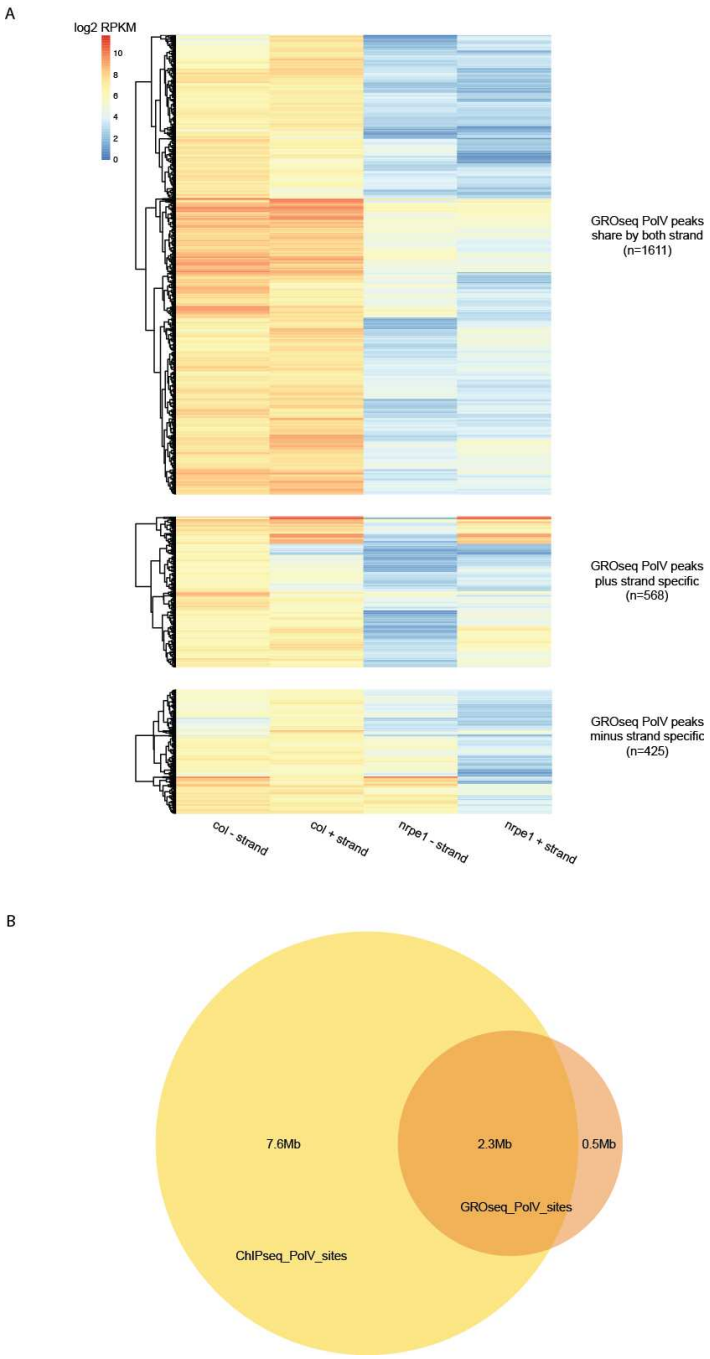


Figure 2.2 Validation of Pol V transcripts.
A) Called Pol V transcripts were called based on their increase of expression in the alpha-Amanitin background (two left columns) and their decrease in expression in the *nrpe1* background (two right columns). Differential expression is based on the RPKM of the transcripts. B) Overlap of the Pol V ChIP-seq data and the transcripts called by GRO-seq shows 2.3MB of overlap with 7.6MB specific to ChIP-seq and 0.5MB specific to GRO-seq.

Pol V transcript length

Without a comprehensive analysis of Pol V transcripts, previous groups assumed they were long noncoding RNAs based on the identification of a handful of transcripts. With our global analysis of Pol V transcripts we mapped the transcript length distribution of our called transcripts to show that the highest frequency appeared at approximately 50bp for all transcripts that were less than 100bp (Figure 2.3A). When looking at all Pol V transcripts that were greater than 100bp there is a very strong peak around 1kb, which represents an artifact in the library preparation that is consistently seen across all data sets (Figure 2.3B). Based on the lack of peaks for transcripts greater than 100bp it appears clear that the strongest peak of Pol V transcripts sits at 50bp in contrast to earlier expectations.

While examining the distribution of transcript lengths there is an unexpected bump in the 20 to 35 bp region that persists in the *nrpd1* mutant background (Figure 2.3A). This suggests that there is a class of Pol V transcripts that are independent of RDDM mechanisms and are specifically smaller than most of the other identified Pol V transcripts. This result may indicate that some sites with CG and CHG methylation specifically rely on the use of small transcripts from Pol V. Based on the 25bp and 50bp peaks it appears that there are two specific size classifications of Pol V transcripts with potentially specific purposes for each class. The 50bp class probably have the standard scaffolding RNA role, though we are unable to know the specific details of their function without directed disruption.

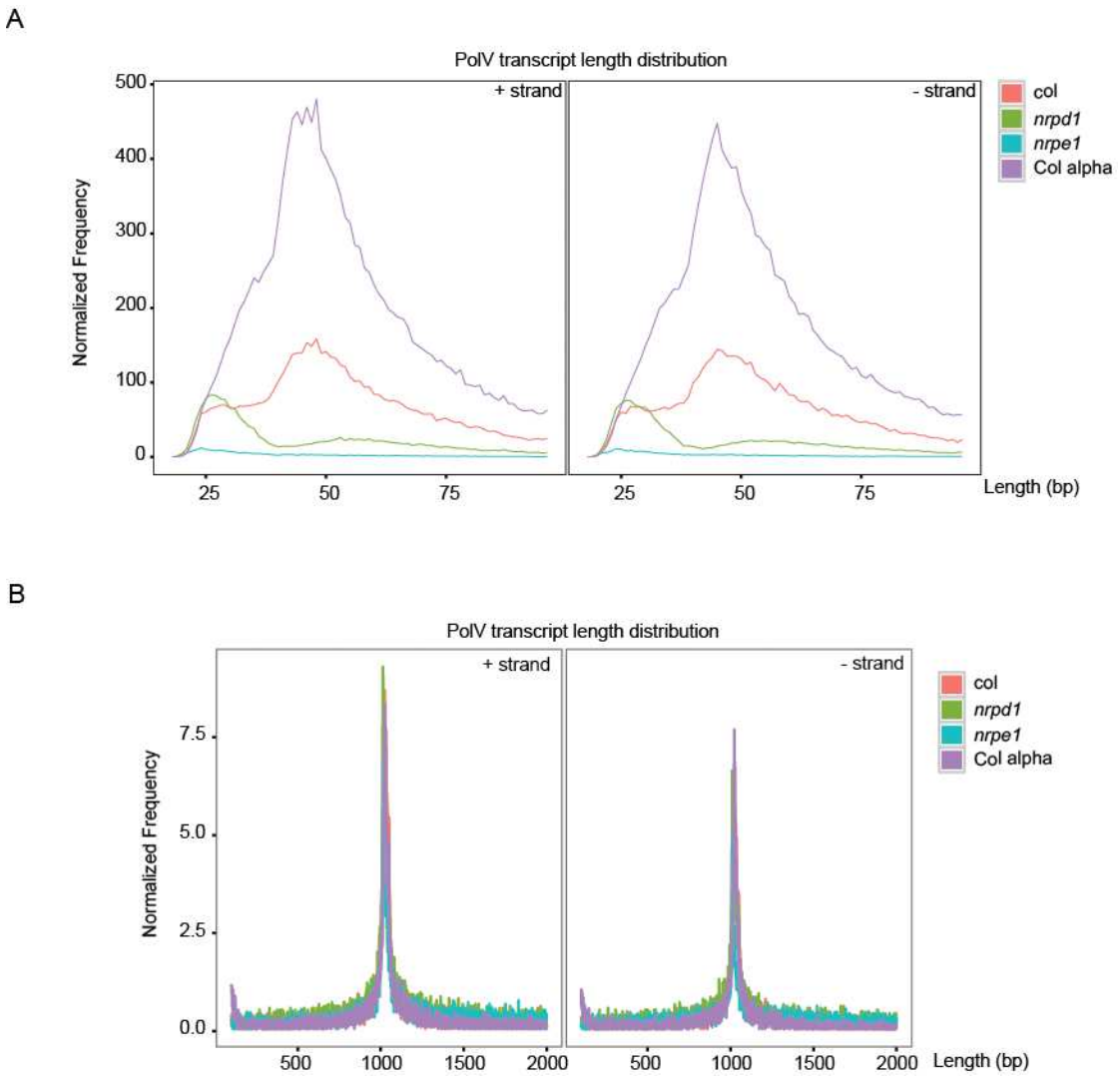


Figure 2.3 Length distribution of Pol V transcripts.
A) Normalized frequency of Pol V transcript length was plotted for Col, Col treated with alpha-Amanitin, *nrpd1*, and *nrpe1*. Only transcripts less than 100bp in length were plotted. B) Same plots from (A) were generated for transcripts greater than 100bp in length.

Pol V transcripts show a monophosphate 5' end structure

Previous analysis of the 5' end structure of one Pol V dependent transcript known as IGN5 showed Terminator Exonuclease (Exo), an enzyme that degrades RNAs with a 5' monophosphate, reduced IGN5 signal as measured by RT-PCR [51]. Tobacco Acid Pyrophosphatase (TAP), which converts a 5' cap to a monophosphate showed increased reduction of IGN5 in combination with Exo, suggesting that the majority of

IGN5 RNA had a 5’ monophosphate, while some small fraction had a 5’ cap. To analyze this in-depth we modified our GRO-seq method to include a shrimp alkaline phosphatase (rSAP) step. The rSAP removes 5’ phosphates that are essential for library generation, and therefore any 5’ monophosphorylated transcripts will not be detected after an rSAP treatment. Analysis of the resulting libraries shows that most Pol V transcripts were eliminated or at least substantially reduced (Figure 2.4). Some signal does persist at specific loci and whether or not that is the result of Pol II compensation protecting the transcripts, or simply inefficiencies of the rSAP enzyme remains to be seen, though enzyme inefficiency seems more likely.

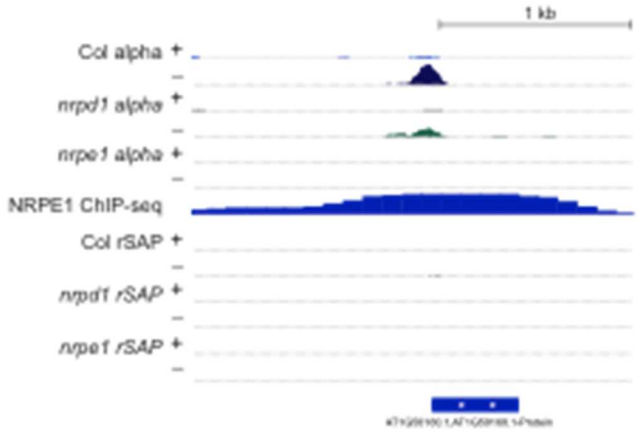


Figure 2.4 Pol V transcripts are monophosphorylated.
Browser shot of sample Pol V specific transcripts with normalized read densities along the y-axis. Samples from top to bottom are Col treated with alpha-Amanitin (+) and (-) strands, *nrpd1* treated with alpha-Amanitin, *nrpe1* treated with alpha-Amanitin, NRPE1 ChIP-seq, Col treated with rSAP, *nrpd1* treated with rSAP, and *nrpe1* treated with rSAP.

Confirmation of Pol IV transcripts

With the classification of Pol V transcripts, we used previously identified Pol IV transcripts [50] to verify the GRO-seq technique was working as expected. Given that Pol IV transcripts are believed to have similar monophosphorylated 5’ ends they should therefore be identified along with Pol V transcripts. Given the difficulties of identifying the

Pol IV transcripts in a wild type background such as Col-0, we examined several of the highest expressed transcripts and found that they were indeed cleanly identified as Pol IV dependent and Pol V independent (Figure 2.5). With the identification of Pol IV transcripts, we increased our confidence in the technique and confirmed previously identified transcripts.

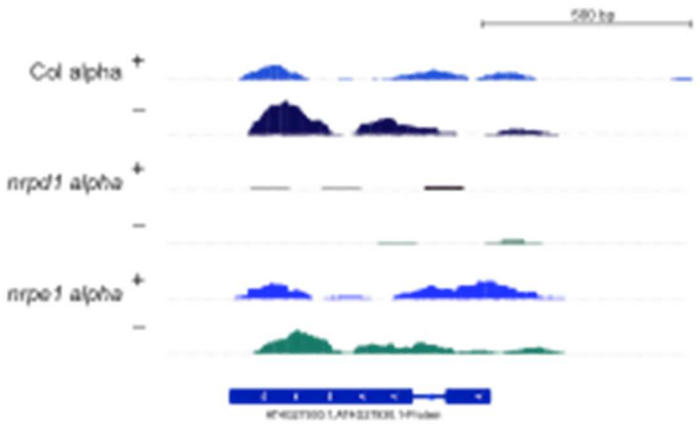


Figure 2.5 Identification of Pol IV specific transcripts.
Browser shot of sample Pol IV specific transcripts with normalized read densities along the y-axis. Samples from top to bottom are Col treated with alpha-Amanitin (+) and (-) strands, *nrpd1* treated with alpha-Amanitin, *nrpe1* treated with alpha-Amanitin.

Identification of Pol IV dependent and independent Pol V transcripts

Pol V transcription is dependent on the presence of DNA methylation due to the feedback between DNA methylation, histone methylation, and recruitment of the DDR complex [104]. Therefore while Pol IV transcript expression appears to be primarily controlled by Pol IV [50], this is not the case for Pol V transcripts. We set out to identify Pol V transcripts that were RDDM dependent and independent using the GRO-seq from the *nrpd1* mutant background as *nrpd1* is deficient in CHH methylation but still maintains CG and CHG methylation. These methylation differences are due to DRM2 controlling CHH methylation while CG and CHG methylation are largely controlled by CMT3 and

MET1. We were able to identify Pol V transcripts that are Pol IV dependent and independent (Figure 2.6A,B). Unsurprisingly, a large percentage of Pol V transcripts were dependent on Pol IV, suggesting that those Pol V transcription are largely dependent on CHH methylation. This result fits with the currently accepted model, but the large amount of Pol IV independent transcripts was unexpected (Figure 2.6A). These Pol IV independent transcripts suggest Pol V can utilize CHG and CG methylation for recruitment at a much higher rate than expected.

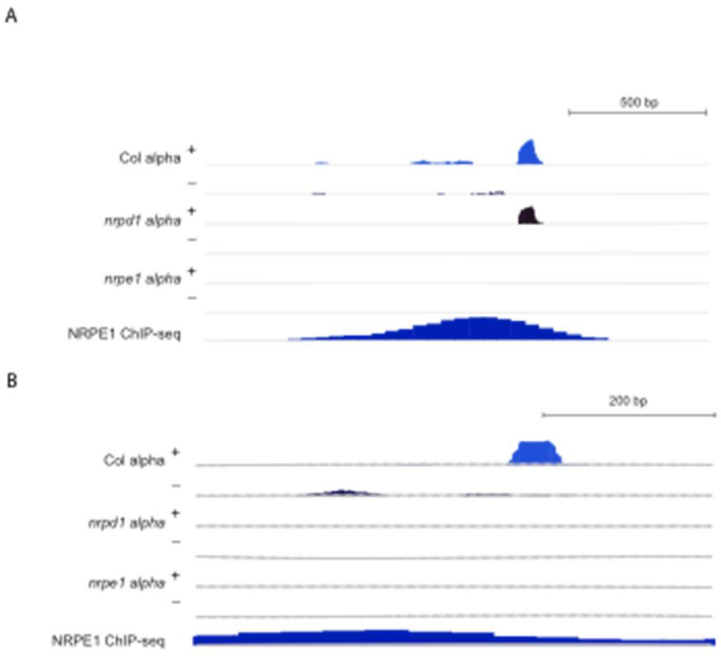


Figure 2.6 Pol V transcript dependence on Pol IV activity.
A) Browser shot of sample Pol V specific transcripts with normalized read densities along the y-axis. Samples from top to bottom are Col treated with alpha-Amanitin (+) and (-) strands, *nrpd1* treated with alpha-Amanitin, *nrpe1* treated with alpha-Amanitin. (B) Same browser shot with the same conditions as (A) but demonstrating a Pol IV dependent transcript instead of an independent one

2.3 Discussion

In this study we have put forward a method for analyzing transcripts produced by Pol IV and Pol V in *Arabidopsis thaliana* and used it to identify previously

uncharacterized Pol V transcripts and confirm known Pol IV transcripts. This method is an important demonstration of the modification of the GRO-seq technique for a variety of applications and the data gained from this study will be valuable for advancing our understanding of noncoding RNAs in plants and the possible alternative evolutionary pathway plants have taken to achieve regulatory diversity of gene expression without the use of enhancers.

This study is the first global identification of Pol V transcripts and there were several unexpected results. The overlap with ChIP-seq is an important validation of the called transcripts and also a demonstration of the much higher resolution of GRO-seq compared to ChIP-seq given the general overlap between the two data sets with ChIP-seq being much less specific to the actual transcript. Unexpectedly, there is a small amount of GRO-seq signal outside of the ChIP-seq signal. Whether this is simply an issue with sequencing depth and peak calling or a poor antibody we cannot say, but it will require more in depth validation to determine what is occurring at these sites.

The length of Pol V transcripts was expected to be several hundred base pairs given previous studies focusing on long noncoding RNAs, but our data clearly shows a strong peak around 50bp that is much shorter than expected. Additionally, the subset of even smaller Pol V transcripts that persist in the *nrpd1* background suggest there may be two different classes of Pol V transcripts. In the *nrpd1* background CHH methylation is disrupted, but CG and CHG persist, so it is possible that the smaller Pol V transcripts are specifically recruited to these different methylation types. Further work will be needed to verify if this is indeed happening, but it would open up new avenues to explore regarding the function of Pol V and RDDM. Additionally, this result suggests more diversity in transcripts produced and possibly function for Pol V compared to Pol IV, which only produces small noncoding transcripts for siRNA production.

Interestingly, Pol V transcripts demonstrated a strand preference in less than half of the total transcripts called. This result indicates that once the DNA is accessible for transcript production, it is less reliant on a specific sequence orientation. In the current model of Pol V transcripts being used primarily as a scaffold this result may not be particularly surprising, but it is still an important demonstration that Pol V transcripts may have faced less evolutionary pressure due to their function. It is also an important point regarding how Pol V functions as a polymerase in that it may be less stringent in its transcript production. Additional study will be helpful in determining if there is a particular motif, or sequence composition at these different transcripts and how they relate to function.

The identification of Pol IV dependent and independent Pol V transcripts presents several new questions regarding Pol V function. We have identified roughly half of Pol V transcripts to be independent of Pol IV function. This result is unexpected given the current model where Pol IV siRNA biogenesis results in the recruitment of Pol V and subsequent transcript production. Instead, it would appear that a substantial amount of Pol V activity is independent of Pol IV and may be recruited by CG or CHG methylation, which are maintained by MET1, CMT3, and CMT2. These results open up new avenues to explore in DNA methylation and the role of Pol V transcripts.

2.4 Materials and Methods

Plant Material and Growth Conditions

The *A. thaliana* accession Columbia (Col-0) was used as the wild type genetic background for this study. The mutant alleles of *nrpd1-4* [46] and *nrpe1-12* [48] used in this study have been characterized previously and were in the Col-0 background. All plants were grown on soil under long day conditions (16 hours light, 8 hours dark).

Inflorescence tissues was collected and used for the GRO-seq protocol.

Nuclei Isolation

Approximately 10 grams of inflorescence and meristem tissue was collected from plants and immediate placed in ice cold grinding buffer (300 mM sucrose, 20 mM Tris, pH 8.0, 5 mM MgCl2, 5 mM KCl, 0.2% Triton X-100, 5 mM β-mercaptoethanol, 35% glycerol). Nuclei were isolated as described previously (Hetzel et al. 2016). Briefly, samples were ground with an OMNI International General Laboratory Homogenizer at 4°C until well homogenized, filtered through a 250μm nylon mesh, 100μm nylon mesh, miracloth, and finally a 40μm cell strainer before being split into 50ml conical tubes. Samples were spun for 10 minutes at 5250g, the supernatant was discarded, and the pellets were pooled and resuspended in 25ml of grinding buffer using a dounce homogenizer. The wash step was repeated at least once more and nuclei were resuspended in 1ml of freezing buffer (50 mM Tris, pH 8.0, 5 mM MgCl2, 20% glycerol, 5 mM β-mercaptoethanol).

GRO-seq Library Preparation

GRO-seq was performed as described previously (Lam et al. 2013, Duttke et al. 2015) with modifications as follows. Approximately 10^7 nuclei were run-on at room temperature for 4 minutes with 20nM limiting CTP concentration. When indicated, 5-10 μg/ml αAmanitin was added to inhibit Pol II transcript elongation, thereby increasing the signal from Pol V. Following the run-on, samples were treated with RQ1-DNAse (Promega) for 20’ and subsequently stopped with 80 μl of STOP mix (20mM EDTA; 200 mM NaCl; 1% SDS; 0.3 mg/mL glycogen; 0.3 mg/mL ProteinaseK) and incubated at room temp for 30’. RNA was extracted using Trizol LS Reagent (Thermo Fisher Scientific) as described by the manufacturer. Purified RNA was resuspended in H2O + 0.05% Tween20 and, if indicated, fragmented using RNA fragmentation reagents

(Ambion). Nascent RNA was enriched through two IPs against BrdU (Santa Cruz Biotechnology). When fragmented, 3' repair and 5' kinasing were performed as described in Duttke et al. (2015). If indicated, samples were de-capped using RppH (New England Biolabs), phophatase treated using Polyphosphatase (Epicentre) or dephosphorylated with rSAP (New England Biolabs). Preparation of the sRNA libraries for Illumina platforms using the NEBNext Small RNA Library Prep Set (New England Biolabs).

For the small RNA_GRO, fragmentation and RppH treatment were skipped in order to enrich for small, monophosphorylated transcripts associated with Pol IV and Pol V.

Analysis for NGS data

All data generated by this study and previously published data were reanalyzed in a consistent manner as described below. Adaptors and primers were trimmed with cutadapt from raw reads. Trimmed reads were then mapped to Tair10 genome with Bowtie [94] by setting seeds length to 25 (-l 25) and allowing 3 mismatches for seeds (-n 3). Only paired end reads aligned within 1kb of each other were kept.

Only reads that aligned to the genome at a single unique location were considered for downstream analysis. All analysis or visualizations of NGS data were normalized to a read depth of 10^7 total uniquely aligned reads per experiment. Genome browser tracks depicting normalized read densities for each experiment were generated using HOMER [95] and visualized in IGV [96]. Meta gene plots, histograms, heatmaps, GO enrichment analysis, and gene expression values containing normalized read densities were generated using HOMER, using gene and feature annotations from TAIR10.

De novo Transcript Identification

Transcript discovery and annotation were performed using routines in HOMER as described below. Transcription units were identified directly from GRO-seq data using the HOMER program findPeaks (-style groseq with default parameters) as described in Wang et al. [97], which looks for regions of continuous, strand-specific GRO-seq read coverage to identify transcripts.

2.5 Acknowledgements

Chapter 2, in part, is being prepared for submission for publication. Hetzel, Jonathan A.; Duttke, Sascha H.C.; Liu, Wanlu; Jacobsen, Steve; Chory, Joanne. "Identification and Characterization of Pol V Transcripts in *A. thaliana*." The author of this dissertation was the co-primary investigator with Sascha Duttke and Wanlu Liu and will submit as the co-first author of this paper.

CHAPTER 3 Functional Analysis of Critical Residues in the Brassinosteroid Receptor BRI1

3.1 Abstract

The leucine rich repeat receptor like kinase (LRR-RLK) BRI1 is the primary receptor for the plant steroid hormones brassinosteroids (BR) in *Arabidopsis thaliana*. Significant advances in the understanding of early BR signaling events came with the resolution of the crystal structure of BRI1, its coreceptor BAK1, and the most biologically active brassinosteroid brassinolide (BL) in a ternary complex. Using this structure as a model for early BR signaling we worked to test the predicted critical amino acid residues in the extracellular domain of BRI1 involved in N-linked glycoslyation, steroid perception, and coreceptor interaction. Each residue was tested with single amino acid changes in the BRI1 protein and subsequent complementation of a *bri1-null* mutant. Complementation of the *bri1-null* was followed by protein localization and accumulation using an mCITRINE tag, etiolated growth, steroid binding, and coreceptor interaction. We were able to identify a single glycosylation site at $BRI1^{N154}$ that has a role in protein stability, two critical residues of the steroid-binding pocket at $BRI1^{Y597}$ and $BRI1^{Y599}$ that reduce binding affinity, and one site at $BRI1^{M727}$ that is critical for coreceptor interaction. Our study supports a model of early BR signaling where BRI1 binds to BL and is then able to bind to BAK1, provides insight into the post translational regulation of BRI1, and demonstrates the use of structural data to tune BR signaling.

3.2 Introduction

Brassinosteroids are a group of steroid hormones that are required for proper growth and development in plants. These hormones have been shown to be involved in a number of processes including cell elongation and differentiation, pathogen defense, reproductive development, and general growth. When the endogenous BR signaling

pathway is disrupted the resulting plants have severe growth defects including short hypocotyls and petioles, dark green leaves, dwarfed stature, and reduced male fertility[105, 106]. In *Arabidopsis thaliana* the BR signaling pathway has been extensively studied and the pathway is connected from ligand perception by the receptor at the plasma membrane to the transcription factor response in the nucleus [107].

At the molecular level BR signaling begins with steroid perception by the plasma membrane localized receptor BRI1 and subsequent interaction with the coreceptor BAK1 [108-110]. BRI1 was first identified as a plasma membrane localized leucine rich repeat receptor-like kinase from forward genetic screens for BR insensitivity more than ten years ago [105, 111]. The extracellular domain contains 25 leucine rich repeats with a 70 amino acid island domain between LRRs 21 and 22, which was identified as the steroid binding pocket [112]. Several gain and loss-of-function mutations mapped to the region encoding the island domain and provided further evidence for it as the site of hormone interaction. The intracellular domain contains a kinase domain, an activation loop, and the C-terminal tail, which has auto-inhibitory effects on the kinase domain in low BR conditions[113].

Recently, x-ray crystallography work has shown the interactions between the extracellular domains of BRI1, the coreceptor BAK1, and the steroid brassinolide (BL) expressed in insect cells [114-117]. The crystal structure of the BRI1 extracellular domain showed the LRRs forming a super helix with the island domain forming a BL binding surface with the LRR. A later crystal structure of the ternary complex showed BL interacting with both the BRI1 and BAK1 extracellular domains to hold them together and allow the interaction of their kinase domains.

In this current study we used the BRI1 extracellular domain crystal structure to select several residues for study related to N-linked glycosylation, steroid binding, and

potential sites of interaction with the coreceptor BAK1. Using site directed mutagenesis, we disrupted each of these residues to determine their function through phenotypic analysis and characterization. We found one glycosylation site that is critical for protein stability at BRI1^{N154}, two sites in the binding pocket that are important for BL perception at BRI1^{Y597} and BRI1^{Y599}, and one site involved in the interaction with the coreceptor BAK1 at BRI1^{M727}. These findings support the proposed interaction sites seen in the crystal structure of BRI1, BAK1, and BL in complex, provide evidence for BRI1 and BL binding occurring before BRI1 and BAK1 interact, and gives insight into post-translation regulation of BRI1.

3.3 Results

A Single Glycosylation Site is Critical for BRI1 Stability

Based on the crystal structure of the extracellular domain of BRI1, nine N-glycosylation sites were predicted to mask large surfaces of the N-terminal half of the extracellular domain but not the C-terminal portion where the steroid binds (Figure 1a). Two of these sites were particularly interesting as they mapped to the interior of the super helix, which was predicted to play a role in stability, and one site was found that has contact with the island domain, suggesting a possible role in steroid binding (Table S1).

Figure 1

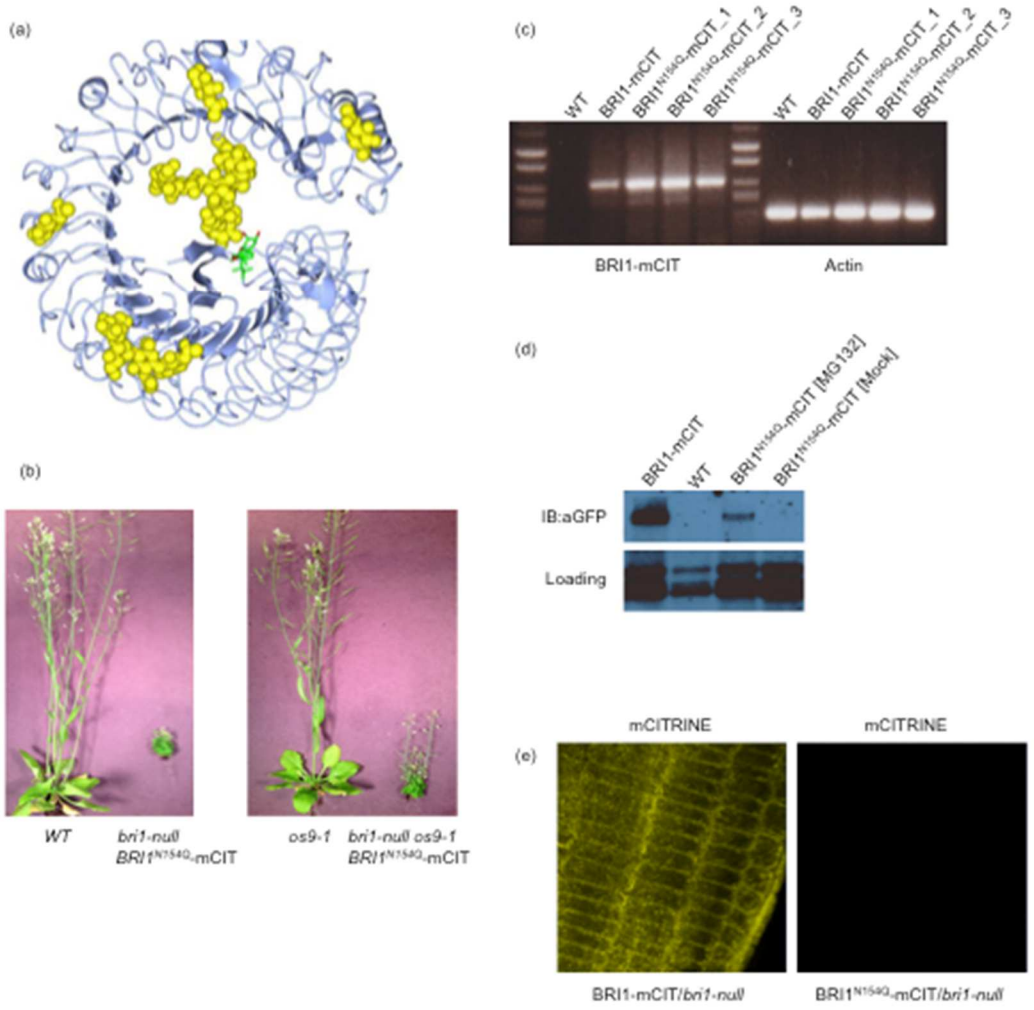


Figure 3.1. The glycosylation mutant BRI1^{N154Q}-mCITRINE shows a dwarf phenotype in the *bri1-null* background caused by a lack of protein accumulation and localization.
(a) Nine predicted N-linked glycosylation sites mask large surfaces of the BRI1 extracellular domain. BRI1 residues are colored lavender and the glycans are colored yellow. (b) BRI1^{N154Q}-mCITRINE is not able to rescue the *bri1-null* mutant phenotype and the *os-9* background cannot rescue the phenotype of the double mutant. (c) Transcript levels were analyzed with semi-quantitative PCR using primers spanning *BRI1* and the mCITRINE tag. Actin primers were analyzed as a standard. (d) Protein accumulation was analyzed by western blot using anti-GFP antibodies on microsomal fractions derived from seven day-old seedlings. A four-hour 50μM MG132 treatment was used to inhibit protein degradation. Cross-reacting bands from the anti-GFP antibodies were used as loading controls. (e) Protein localization and accumulation of BRI1^{N154Q}-mCITRINE in root tips from seven day-old seedlings was visualized by confocal microscopy.

To disrupt these N-glycosylation sites, mutations resulting in Asp to Glu were made to prevent glycosylation while minimizing the impact on neighboring residues. Each of the glycosylation mutant proteins was tagged with mCITRINE, a monomeric yellow variant of GFP, and expression was under the control of the endogenous 1.7kb BRI1 promoter. Constructs were transformed into a heterozygous *bri1-null* mutant of *Arabidopsis* yielding more than 20 T1 plants per construct. The *bri1-null* mutant was a T-DNA insertion in the BRI1 gene GABI_134E10 and showed the characteristic severe dwarfism and BR deficient phenotype [109]. A minimum of five homozygous lines was selected for independent transformation of each construct and characterized to select lines expressing BRI1 similar to wild type levels. Phenotypes in a homozygous *bri-null* background were scored and protein accumulation and localization was determined by western blot and confocal microscopy respectively. BRI1-mCITRINE without any mutations rescues the *bri1-null* mutant phenotype [109], accumulates and localizes to the plasma membrane and endosomes as shown for wild type [118], and was used as a control during these analyses.

Each construct complemented the *bri1-null* phenotype with the exception of the $BRI1^{N154Q}$-mCITRINE lines, which were severely dwarfed and appear to have a similar phenotype to *bri1-null* (figure 1b). Transcript levels of $BRI1^{N154Q}$-mCITRINE were similar to the BRI1-mCITRINE control, which was previously shown to rescue the *bri1-null* phenotype (figure 1c). $BRI1^{N154Q}$-mCITRINE protein levels however, were almost undetectable by western blot or confocal microscopy (Figure 1d, e). This low level of protein may account for the inability to complement *bri1-null*. All other glycosylation mutant proteins accumulated and localized to the plasma membrane (Supplemental Figure 1,2). The $BRI1^{N154Q}$-mCITRINE construct was crossed to an endoplasmic-reticulum-associated protein degradation (ERAD) mutant known as *os-9* (Figure 1b).

The *os-9* mutant was previously shown to be able to rescue the phenotype of *bri1-5,* which has a mutation in the N-terminal cap causing ER retention and a dwarf phenotype [119, 120]. The *os-9* mutant background, however, did not rescue BRI1^{N154Q}-mCITRINE function. The main difference between *bri1-5* and BRI1^{N154Q}-mCITRINE is that the BRI1 protein is still detected in *bri1-5* even though it is not secreted. The BRI1^{N154Q}-mCITRINE and *os-9* double mutant phenotype suggests that degradation rather than retention is the likely cause of the dwarf phenotype.

To investigate degradation of the BRI1^{N154Q}-mCITRINE protein, seven day-old seedlings were treated with 50μM MG132 for four hours and protein accumulation was analyzed by western blot (Figure 1c). MG132 reduces protein degradation and was predicted to rescue the BRI1^{N154Q}-mCITRINE protein. Whole seedlings were used for western blot analysis and the mutant protein accumulates, though it does not reach the same levels seen in the wild type BRI1-mCitrine lines. Together these results suggest that protein degradation is the reason BRI1^{N154Q}-mCITRINE cannot complement the *bri1-null* mutant.

Mutational Analysis of the BL Binding Site Within BRI1

The structure of the extracellular domain of BRI1 bound to BL shows a binding pocket formed by the BRI1 island domain and the last five LRRs (Figure 2a). To characterize these potential steroid interaction sites we made mutations to remove interactions between BL and BRI1 (Table S1). The constructs were tagged and transformed as described for the glycosylation mutants. Three of the BRI1 mutants *BRI1^{W564H}*-mCITRINE*, BRI1^{M657A}*-mCITRINE*,* and *BRI1^{F681A}*-mCITRINE complemented the *bri1-null* mutant with wild type protein accumulation and localization (Supplemental Figure 3, 4). Two lines *BRI1^{Y597V}*-mCITRINE and *BRI1^{Y599V}*-mCITRINE showed a semi-dwarf phenotype, which is not as severe as a *bri-null* and is still fertile. These mutant

proteins accumulated and localized to plasma membrane similar to wild type BRI1 across multiple lines (Figure 2c, d).

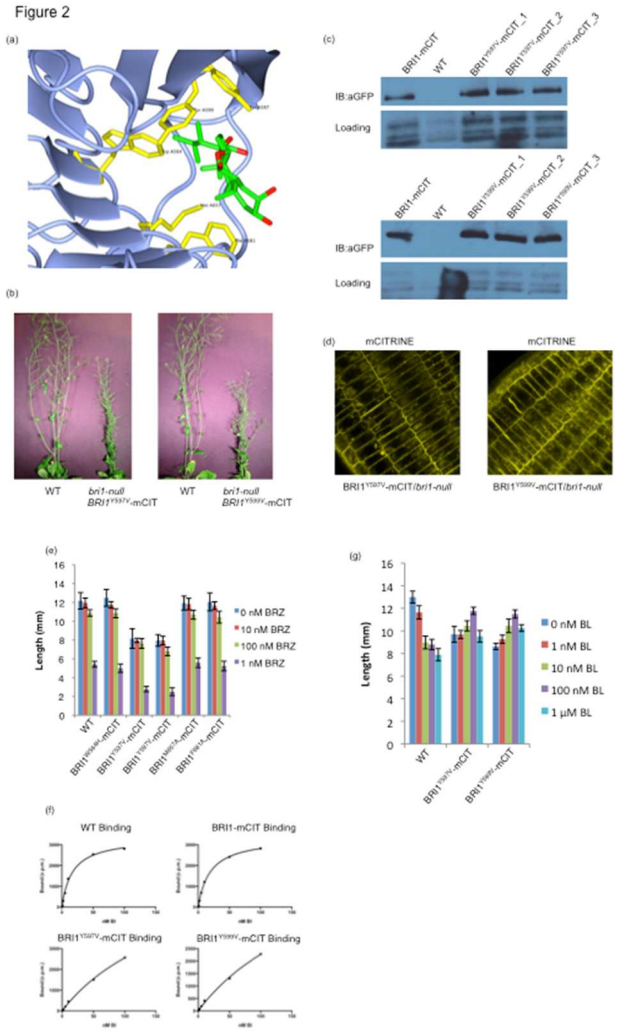


Figure 3.2. Two BR binding mutant proteins BRI1^{Y597V}-mCITRINE and BRI1^{Y599V}-mCITRINE lead to a semi-dwarf phenotype caused by a reduced binding affinity for BL.
(a) Five potential steroid binding residues were disrupted to test their functions. The BRI1 residues are colored lavender, the BL steroid is colored green, and the five target residues are colored yellow. (b) BRI1^{Y597V}-mCITRINE and BRI1^{Y599V}-mCITRINE show a semi-dwarf phenotype in a *bri1-null* background. (c) Protein accumulation was analyzed by western blot using anti-GFP antibodies on microsomal fractions purified from seven day-old seedlings. Multiple independent lines were compared for each construct. Cross-reacting bands from the anti-GFP antibodies were used as loading controls. (d) Confocal microscopy images of root tips of seven day-old seedlings show accumulation and localization of the protein to the plasma membrane. (e) Hypocotyl length of six day-old etiolated seedlings was measured under varying concentration of BRZ. Error bars represent the standard error from N=25 samples measured for one independent transformation and BL concentration. (f) Using [^{3}H]-BL, specific binding was determined from 100 µg of total protein of microsomal fractions of seven day-old seedlings. Non-specific binding was determined using 100-fold cold inhibitor and subtracted from total binding. (g) Hypocotyl growth of six-day etiolated seedlings was measured under increasing concentrations of exogenous BL. Error bars represent the standard error from N=25 samples from one independent transformation and BL concentration.

BR deficient mutants show reduced etiolated growth, which can be quantified with a dose response of the BR biosynthesis inhibitor brassinazole (BRZ) [121]. We measured, etiolated growth for these mutant proteins in the presence of BRZ at a concentration of 0 nM, 10 nM, 100 nM, and 1 μM (Figure 2e). In both wild type and mutant plants the etiolated growth was significantly impaired with higher concentrations of BRZ. However, while the semi-dwarf plants expressing *BRI1*^{Y597V}-mCITRINE and *BRI1*^{Y599V}-mCITRINE have reduced etiolated growth as expected, they are not hypersensitive to BRZ. This result makes their inability to complement *bri-null* likely from a reduced ability to perceive endogenous BR rather than an unknown mechanism. There is no gain of function or resistance to BRZ seen in any other BL binding site mutants (Figure 2e).

To directly test the steroid binding abilities of these mutants, tritium-labeled BL was used for saturation binding experiments and the calculation of K_d values (Figure 2f). The calculated K_d of wild type BRI1 using microsomal fractions from 7 day-old seedlings was 16.9 ± 2.2 nM, which is in agreement with previously published results [122]. As a control for the mCITRINE tagged lines, BRI1-mCITRINE had a K_d of 19.0 ± 1.4 nM, which is also in agreement with previously published results. For *BRI1*^{Y597V}-mCITRINE the calculated K_d was 188.2 ± 27.2 nM and BRI1^{Y599V}-mCITRINE had a K_d of 203.4 ± 45.1 nM. This significant reduction in binding affinity resulting in a reduced ability to respond to endogenous BRs provides an explanation for the semi-dwarf phenotype seen in these two mutants lines. No significant binding affinity differences were seen in the other mutants compared to BRI1-mCITRINE (Table S2).

To confirm the reduced binding affinity, increased concentrations of exogenous BL were used to rescue the etiolated growth phenotype of BRI1^{Y597V}-mCITRINE and

BRI1^{Y599V}-mCITRINE expressing lines (Figure 2g). As expected, higher concentration of BL had the characteristic growth inhibition effects in wild type [123]. However, the etiolated growth phenotype of BRI1^{Y597V}-mCITRINE and BRI1^{Y599V}-mCITRINE expressing lines was partially rescued with exogenous BL, and higher concentrations of BL are required to cause growth inhibition. Together, these results suggest the semi-dwarf phenotype of the mutant protein expressing lines results from a reduced ability of the mutant proteins to bind BL.

Identification of the BAK1 Interaction Surface

Prior to the resolution of the BRI1, BAK1, and BL ternary structure, the BRI1 extracellular domain was used to build a homology model predicting sites of BRI1 and BAK1 interaction. Mutations disrupting candidate amino acid residues were made as described for the glycosylation mutants (Table S1). Lines expressing BRI1^{M727D}-mCITRINE, showed a semi-dwarf phenotype in the *bri1-null* background (Figure 3a). Protein accumulation and localization appeared similar to wild type by western blot and confocal microscopy across multiple lines of this mutant (Figure 3b, c).

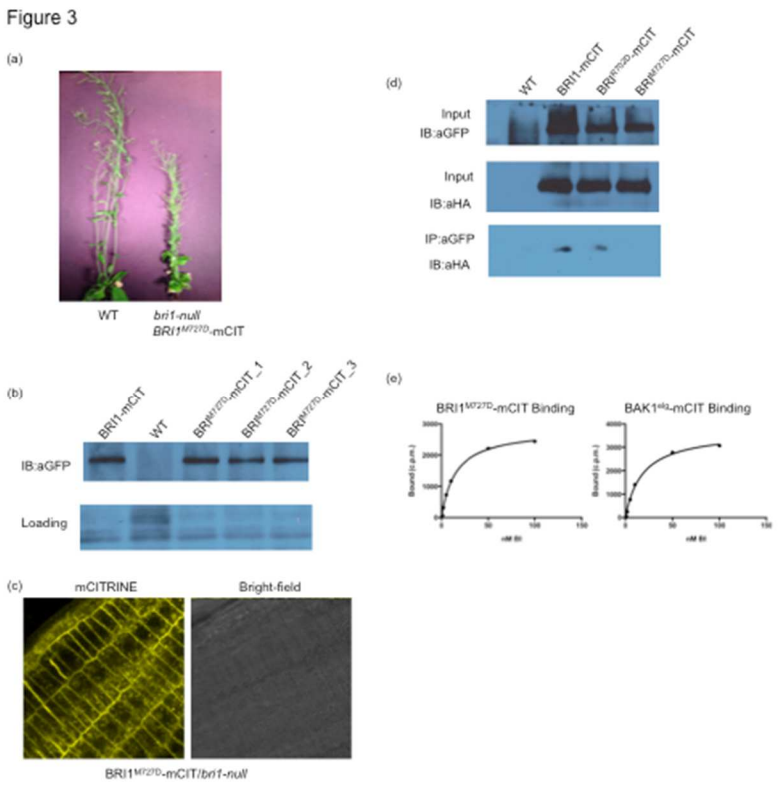


Figure 3.3. BRI1 residue 727 is critical for coreceptor interaction but not ligand binding. (a) The mutant BRI1^{M727D}-mCITRINE shows a semi-dwarf phenotype in a *bri1-null* background. (b) Protein accumulation of BRI1^{M727D}-mCITRINE in seven day-old seedlings was analyzed by western blot of microsomal fractions using anti-GFP antibodies. Cross-reacting bands from the anti-GFP antibodies were used as loading controls. (c) Confocal microscopy of root tips from seven day-old seedlings shows BRI1^{M727D}-mCITRINE accumulation and localization to the plasma membrane. (d) Total protein from microsomal fractions of seven day-old seedlings was immunoprecipitated with anti-GFP antibody conjugated magnetic beads and subjected to immunoblot as indicated. (e) [^{3}H]-BL was used to determine the BR binding affinity for BRI1^{M727D}-mCITRINE and BAK1prom:bak1^{elg}-mCITRINE from seven day-old seedlings.

We next crossed the BRI1^{M727D}-mCITRINE line with another line expressing BAK1prom::BAK1:6xHA for co-immunoprecipitation experiments (Figure 3d). For controls, BRI1^{R702D}-mCITRINE, which complemented the *bri1-null*, and BRI-mCITRINE were also crossed with BAK1prom::BAK1:6xHA. BAK1::6xHA coimmunoprecipitated with both BRI1^{R702D}-mCITRINE and BRI-mCITRINE, but not with BRI1^{M727D}-mCITRINE. These results suggest BRI1 residue M727 is necessary for BAK1 interaction.

To ensure the phenotype of *BRI1^{M727D}*-mCITRINE expressing lines was not due to a deficiency in steroid binding, saturation binding experiments were performed using

the tritium-labeled BL (Figure 3e). The calculated K_d value of 14.9 ± 1.4 nM was not significantly different compared to wild type suggesting that *BRI1*^{M727D}-mCITRINE disrupts the BRI1-BAK1 interaction and not BRI1-BL interaction. This data also supports a model where BRI1 activation by steroid binding occurs before coreceptor interaction because steroid binding is not affected in the *BRI1*^{M727D}-mCITRINE mutant. To further test this model, saturation binding experiments were also performed in plants expressing BAK1prom:bak1^{elg}::CITRINE in a *bak1-3* background (Figure 3e). *bak1*^{elg} is a gain-of-function mutant, which is hyper sensitive to BR treatment and displays long twisted petioles and elongated leaves [124]. The calculated K_d of 18.1 ± 2.5 nM also is not significantly different compared to wild type, which provides further evidence that BR binding by BRI1 is independent of BAK1.

3.4 Discussion

In conclusion, this study has identified key BRI1 residues related to protein stability, steroid interaction, and BAK1 coreceptor binding. This work tests the models of interaction proposed by the crystal structures of BRI1, BAK1, and BL in a ternary complex with *in vivo* evidence of functional importance of key residues. This study also supports a model in which BRI1 steroid perception precedes coreceptor interaction and downstream signaling.

It was unexpected that only one N-glycosylation site in BRI1 was determined to be critical when compared to other LRR-RLKs such as the mammalian TLR2 which has four such residues, all of which were determined to be functionally important for protein stability and secretion [125]. Based on the TLR2 work we expected to see similar functional importance of N-glycosylation in BRI1, but this was not the case with the identification of only one critical residue. In general the importance of N-glycosylation marks on plant proteins is not fully understood and is still being studied. It is possible

that while the plant and mammalian N-glycosylation pathways are highly conserved, plant proteins may not rely as strongly on glycosylation sites for quality control [126]. It is also possible that the BRI1 protein is unique in that it shows significant glycosylation but we do not yet understand its function.

While we do not see dwarf phenotypes from the removal of most glycosylation sites in BRI1, the degradation of the BRI1^{N154Q}-mCITRINE protein provides evidence for a quality control mechanism that has not yet been identified. Genetic screens that suppress the phenotype of the lines expressing BRI1^{N154Q}-mCITRINE may yield insight into the targeting and degradation of aberrantly glycosylated proteins. The identification of members of this pathway could be critical for understanding the importance of glycosylation in plants. It may also be the case that only specific sequences or structures are accessible by the degradation machinery, hence only one critical site was found in BRI1. Further study with analysis of additional glycoproteins and mutant backgrounds deficient in glycosylation in *Arabidopsis* would be necessary to further address this topic.

The BRI1 steroid-binding pocket was previously shown to be critical for steroid perception, but we lacked details of how BL was being bound by BRI1. This study tested the structure of the pocket by directly targeting individual amino acids and reducing steroid binding affinity without abolishing protein activity. It also further emphasizes the importance of the BRI1 island domain for steroid perception. Of the mutants tested, only the island domain connection was shown to affect steroid binding even though several candidate residues in the LRR core were also tested. It is possible that multiple simultaneous mutations made in the LRR core would show BR deficient phenotypes, but individual amino acid substitutions are not sufficient on their own. Most importantly, these results validate the BRI1 structure steroid binding pocket *in vivo* and demonstrate the ability to manipulate the BR growth pathway through minor mutations. These results

open up the possibility of targeting the steroid-binding pocket for inhibition with a small molecule or through using the genetic data to design minimal mutations to control growth in agriculturally relevant plants.

Finally, this study provides evidence for the early activation of BR signaling being dependent on BRI1 perception of BL followed by interaction between BRI1 and BAK1. The BL binding of *BRI*^{M727D} and *BAK1*^{elg} is similar to wild type BRI1 despite their BR hyposensitive and BR hypersensitive phenotypes respectively. If BL binding was dependent on the BRI1-BAK1 interaction we would expect to see a decrease in binding affinity for plants expressing *BRI1*^{M727D} and in increase in binding affinity for plants expressing *BAK1*^{elg}. The *BRI*^{M727D} inability to stably bind BAK1 also shows *in vivo* evidence of the interaction surface between the two proteins. Further mutations in the BAK1 structure to disrupt BL-BAK1 and BRI1-BAK1 interactions would also serve to confirm this interaction surface. It is important to note that it may be difficult to see a phenotype based on redundancy of other members of the SERK family. It may also be possible to modify the BL structure to only bind to BRI1 but not BAK1 to further characterize this interaction.

3.5 Materials and Methods

Plant Material and Growth Conditions

The *A. thaliana* accession Columbia (Col-0) was used as the wild type genetic background for experiments. All seeds were sterilized and grown on plates containing half Linsmaier and Skoog Medium (Caisson Laboratories) supplemented with basta selection for BRI1 mutant expressing lines. For etiolated growth assays plants were exposed to white light for 2 hours at 22° to induce germination then transferred to the dark at 22° for 6 days. For phenotypic observation plants were grown on soil under long day conditions (16 hours light, 8 hours dark).

Construction of BRI1 Vectors

The vectors for transgenic BRI1 were constructed with MultiSite Gateway Three-Fragment Vector Construction Kit (Invitrogen). All BRI1 mutants were generated with PCR based mutagenesis of BRI1 in pDONR221 using the primers listed in Table S1. PCR mutagenesis was achieved by outcompeting the template DNA with primers containing the mutation of interest. The 1.7kb native BRI1 promoter was subcloned into the pDONRP4-P1R vector and the mCitrine tag was subcloned into the pDONRP2R-P3 vector. The three components were subcloned into the destination vector pB7CWG2 and transformed into the GV3101 agrobacterium strain. Completed vectors were used to transform *Arabidopsis* through floral dip of agrobacterium using 0.05% Silwet-L77.

Microsomal Fraction Isolation

100 mg of tissue from seven day-old seedlings was ground in ice-cold sucrose buffer [20 mM Tris, pH 8; 0.33 M Sucrose; 1 mM EDTA, pH 8; protease inhibitor (Roche)] using a polytron (Brinkman). Samples were centrifuged at 5,000 x g for 10 minutes at 4°C and the supernatant was saved. Samples were centrifuged again at 20,000 x g for 45 minutes at 4°C to pellet microsomes. Pellets were resuspended in 1 ml of immunoprecipitation buffer (50 mM Tris pH 8, 150 mM NaCl, 1% Triton X-100) using a 2-mL potter-Elvehjem homogenizer (Wheaton). Finally, samples were pelleted again at 20,000 x g for 10 minutes at 4°C and the supernatants that contained microsomal associated proteins were saved. Total protein in each sample was quantified by Bradford assay and used for further experiments.

BL Binding Assays

Tritium-labeled BL was made by Perkin Elmer with a specific activity of 3.4 Ci/mmol. Microsomal fractions were prepared from 7 day-old seedlings grown on LS medium plates and resuspended at 2mg/ml of protein in BL binding buffer (0.25 M

mannitol, 10 mM Tris-2-[N-morpholino] ethanesulphonic acid (MES), pH 5.7, 5 mM $MgCl_2$, 0.1 mM $CaCl_2$ and cOmplete mini EDTA-Free protease inhibitor cocktail from Roche). Each reaction contained 100 µg of protein in a total reaction volume of 100 µl and was incubated for 30 minutes at 22°C with the indicated amount of [^{3}H]-BL and unlabelled steroid. After incubation the reactions were filtered through a glass-fibre filter (Whatman GF/F) and washed with 10ml of cold BL binding buffer. The bound fraction was quantified by scintillation counting. Specific binding was determined by subtracting binding with 100-fold excess of unlabeled steroid from total binding. Binding data was analyzed and plotted with GraphPad Prism 6.0. All experiments were repeated a minimum of three times with minimal variation.

Confocal Microscopy of BRI1 Lines

To observe the fluorescent signal from the mCitrine YFP, root tips from 7 day-old seedlings were imaged using a Zeis 710 microscope with 40x objective lens.

3.6 Acknowledgements

We thank J. Woodson and B. Willige for revisions on the manuscript, U. Pedmale for advice for experimental design, Y. Jaillais for vectors containing the BRI1 promoter, mCITRINE tag, and BRI1 cds as well as seeds for BAK1::6xHA and bak1^{elg}-mCITRINE.

Chapter 3, in full, has been prepared by Jonathan Hetzel, Tsegaye Dabi, and Joanne Chory. It does not fit the theme of the rest of the thesis, but is relevant information for members of the brassinosteroid field and therefore has been included here as a standalone chapter. The author of this dissertation was the primary investigator and author of this paper.

3.7 Supplemental Information

Table 3.1S. List of BRI1 Mutants.
List of mutants analyzed for glycosylation, steroid interaction, and coreceptor binding.

Glycosylation Mutants

N112Q	N401Q
N154Q interior of superhelix	N438Q
N233Q	N545Q island domain contact
N275Q interior of superhelix	N573Q
N351Q	

Steroid Binding Mutants

W564H	M657A
Y597V	F681A
Y599V	

Coreceptor Binding Mutants

R702D
M727D

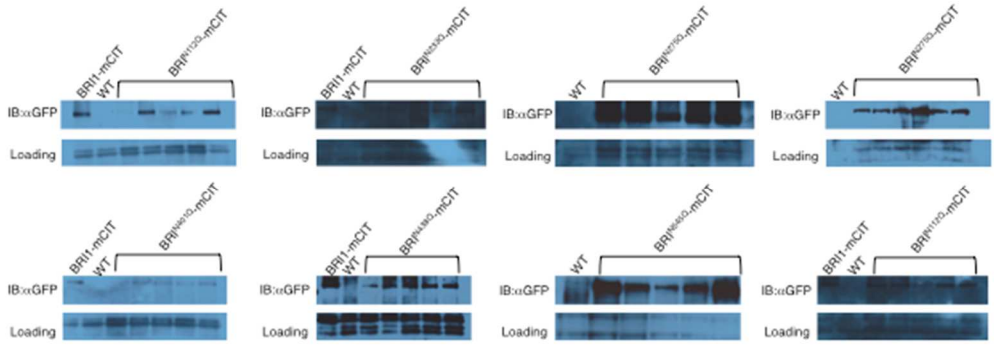


Figure 3.1S. Verification of BRI1 expression by western blot for glycosylation mutants. Probing with an anti-GFP antibody, protein accumulation was analyzed by western blot for multiple lines of each glycosylation mutant. Lines expressing similar to wild type were used for further analysis.

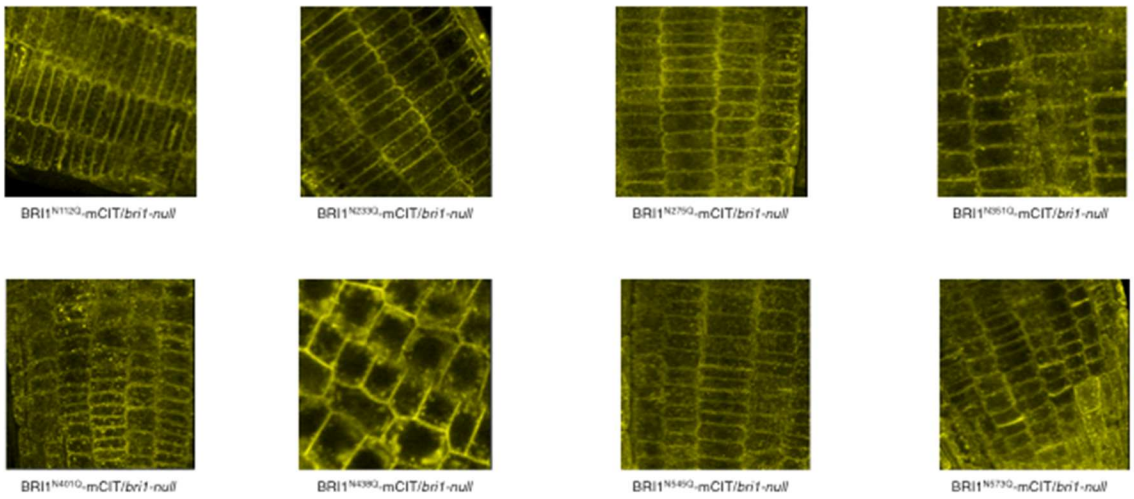


Figure 3.2S. Verification of BRI1 localization for glycosylation mutants by confocal microscopy.
Protein localization and accumulation in root tips from seven day-old seedlings was visualized by confocal microscopy for each glycosylation mutant.

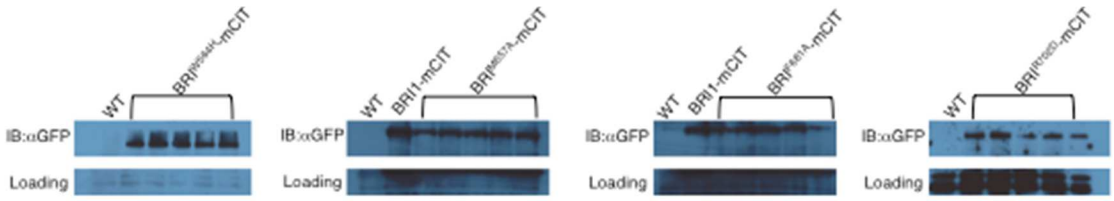


Figure 3.3S. Verification of BRI1 expression by western blot for BR binding and coreceptor binding mutants.
Probing with an anti-GFP antibody, protein accumulation was analyzed by western blot for multiple lines of each BR binding and coreceptor binding mutant.

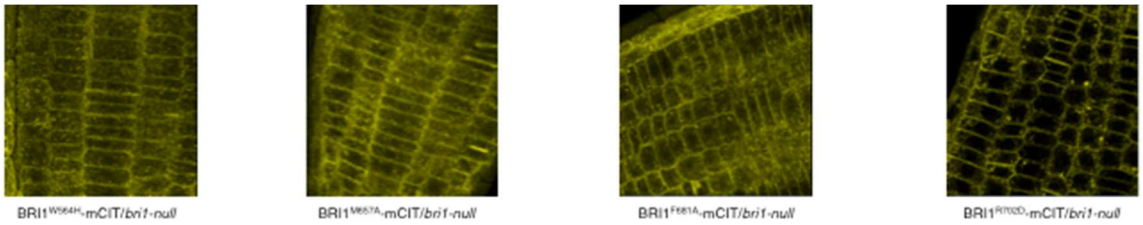


Figure 3.4S. Verification of BRI1 localization for BR binding and coreceptor binding mutants by confocal microscopy.
Protein localization and accumulation in root tips from seven day-old seedlings was visualized by confocal microscopy for each BR binding and coreceptor binding mutant.

Table 3.2S. BL Binding Assay Kd Values.
Using [^{3}H]-BL, specific binding was determined from 100 µg of total protein of microsomal fractions of seven day-old seedlings. Non-specific binding was determined using 100-fold cold inhibitor and subtracted from total binding.

	Kd (nM)
BRI1^{W564H}-mCITRINE	24.6 ± 3.7
BRI1^{M657A}-mCITRINE	21.0 ± 2.7
BRI1^{F681A}-mCITRINE	15.4 ± 2.0

Table 3.3S. BRI1 mutagenesis primers.
List of primers used for the generation of the various BRI1 mutants

Phos-CAAGGCTCCGTTTCTGGCTTC	N112Q Forward
Phos-GATGTGTGAGTTTGAGAGAAA	N112Q Reverse
Phos-CAGGTCTCTTCCAATACACTT	N154Q Forward
Phos-AAGAAACTTCAGACCGGAGCA	N154Q Reverse
Phos-CAATTCTCCACTGGGATTCCT	N233Q Forward
Phos-GTTGGAGGAAACATCGAGAAA	N233Q Reverse
Phos-CAGATCTCTAGTAACCAATTC	N275Q Forward
Phos-CAACAACTTGAGCTCTGTGCA	N275Q Reverse
Phos-CAGTTCTCTGGCGAGTTACCG	N351Q Forward
Phos-GTTACTCGACAACGCGAGTGA	N351Q Reverse
Phos-CAATTCTCCGGTCCGATTCTC	N401Q Forward
Phos-GTTGGAGCTGAGATCTAACGT	N401Q Reverse
Phos-CAGTGTTCTGAGCTGGTTTCG	N438Q Forward
Phos-GCTTAAAGTCGGTGGAATCTT	N438Q Reverse
Phos-CAGAATTCATTCTCCGGGAAC	N545Q Forward
Phos-GCTTAGCTTGAGGATAGCGAG	N545Q Reverse
Phos-CAAGGAACGATTCCGGCGGCG	N573Q Forward
Phos-GAAGAGATTGGTGTTGAGATC	N573Q Reverse
Phos-CAUCTTGATCTCAACACC	W564H Forward
Phos-GATTAAGCTTCTGCAGTC	W564H Reverse
Phos-GUCGTTTATATCAAAAAC	Y597V Forward
Phos-CCTCTTACCGGCGATGAA	Y597V Reverse
Phos-GTTATCAAAAACGATGGG	Y599V Forward
Phos-AACGTACCTCTTACCGGC	Y599V Reverse
Phos-GCGTTTCTGGACATGTCT	M657A Forward

Table 3.3S. BRI1 mutagenesis primers.
List of primers used for the generation of the various BRI1 mutants, continued

Phos-CATCGAACCATTGTTATC	M657A Reverse
Phos-GATATTCTCAATTTGGGT	F681A Forward
Phos-CAGATAAGGCATCGAACC	F681A Reverse
Phos-GACGGTTTAAACATTCTT	R702D Forward
Phos-TAGATCACCTACCTCATC	R702D Reverse
Phos-GACCTTACGGAAATCGAT	M727D Forward
Phos-AGTAAGAGCTGACATAGC	M727D Reverse

PERSPECTIVES

This dissertation has focused on the development and applications of a method to study nascent transcripts in plants. Plant biology has largely fallen behind in the study of transcription due to technical difficulties in handling plant extracts, but we are fortunate in how powerful modern genomic techniques are and how quickly we can catch up with other systems. These first studies are an important first step to demonstrating the power of the technique and providing insight into the fundamental steps of gene expression in the plant kingdom.

Using GRO-seq we have identified general characteristics of plant transcription that differ substantially from mammalian and other model systems. We have demonstrated that plants in general do not appear to use divergent transcription or promoter proximal pausing, and these results imply that gene expression is primarily regulated at the level of initiation in plants. The significance of this finding is rather important given the debate between the role of transcription and translation in regulating the overall level of proteins. More recent statistical analysis shows a much stronger connection to transcriptional regulation rather than translational and post-translational regulation [127-129]. With plants showing an even higher connection between nascent transcription and steady state RNA levels compared to humans, we can say that plants are almost certainly placing the emphasis of their regulation on the basic recruitment of the transcriptional machinery for initiation.

One important caveat to the plant GRO-seq work identifying the basic characteristics of plants transcription has all been carried out in seedling tissue for *Arabidopsis* and corn. I believe the next important step for understanding plant transcription will be to start looking at specific tissue types and multiple developmental time points to see if the general characteristics hold true. If aspects such as promoter

pausing and divergent transcription actually differ based on developmental time points it would imply that these characteristics of transcription are regulated by other factors beyond the genomic sequence which would open up very interesting new areas to explore.

In the more general study of plant transcription, the identification of 3' polymerase pausing is an exciting new feature of gene regulation that is seen more prominently in *Arabidopsis* compared to any other model system. It is not the first time this phenomenon has been observed, but it is definitely pronounced in plant systems and therefore may be a more integral form of transcription regulation. As we have not been able to identify a transcription factor, histone mark, DNA sequence, or any other regulatory element associated with the 3' polymerase accumulation, we are left with more questions than answers at this point and can only say that pausing correlates with gene length and the exclusion of CpG methylation. However, with the identification of this feature of plant transcription we hope that it will open up a new area of study and eventually determine the function of this 3' polymerase accumulation.

Alongside the use of normal GRO-seq, we have been successful with the application of 5' GRO-seq. With 5' GRO-seq we now have a much better annotation of TSSs across the *Arabidopsis* genome and have confirmed a technique to study changes in TSSs with environmental responses. While this was not explored in this work, it is an area that needs further attention to identify additional types of regulation of transcription. This technique also addresses one of the major weaknesses in plant biology that can be seen when cloning genes for study where most researchers will include the 5'UTR up until the next gene to ensure the whole promoter was captured simply because they lack a strong annotation of the TSS. Now when designing constructs for genes of interest, it can be done with more precision and an actual understanding of the full length of the

transcript. While this development may not seem critical, a complete and accurate annotation of a gnome of interest improves the overall quality of scientific endeavors based on these general resources.

While 5' GRO-seq is one of the more powerful techniques presented in this dissertation, it is also one of the most technically challenging as well. Typically, 5' GRO-seq requires a fairly high minimum level of gene expression in order to cleanly call transcripts of interest. As a result, it is necessary to repeat this protocol using different developmental time points as well as different environmental conditions to cover the full range of the *Arabidopsis* transcriptome. As it is currently the most technically challenging of the various GRO-seq variations, it may be a more difficult undertaking than anticipated. Currently myself and other researchers are working on ways to increase the efficiency of the technique and hopefully further work is able to reduce the required level of gene expression. In my hands the best results have come from increasing the number of dephosphorylation steps to reduce the background of non-capped transcription fragments, but there are many avenues to consider and hopefully these technical limitations will be overcome.

One of the important features of 5' GRO-seq is that it allows for the identified of unstable transcripts such as miRNAs that are very difficult to annotate correctly. A simple literature search of some of the most commonly studied *Arabidopsis* miRNAs such as mir158a shows that only the mature miRNA has been identified and studied, while our data clearly shows the primary transcript is significantly longer. The results of the 5' GRO-seq are important for improving the general knowledge within the plant biology field, but also open up the possibility for new and interesting studies that were not possible before as seen with the annotations of miRNAs. We would imagine that the use of the mature RNA is probably sufficient for most studies, but it is possible that

additional regulatory information could be found from a more in depth study of the primary transcript. Again, these tools will help other researchers in the plant biology field and allow for better experimental design.

With the optimization of the GRO-seq method, we have been able to show application of the technique through the characterization of pol IV and pol V transcripts. It is a relatively simple experiment that has been able to answer several decade old questions about the transcripts they produce and their function. With a new reliable method to study the transcripts produced by these two polymerases, studies of how various mutants affect RDDM can move forward well beyond what was possible several years ago.

The validation of known Pol IV transcripts from Zhai et al. [50] supports their model of Pol IV transcription being dependent on very small transcripts that only give rise to one siRNA each. This validation is important because it supports the application of the GRO-seq method for Pol V transcription and actually provides new insights into the Pol IV function. In the Col background we were only able to identify 10 Pol IV targets, which should have been much higher. This result provides evidence that most Pol IV transcripts are immediately being bound by the dicer like complexes and processed into siRNAs. Rapid processing was always suspected given the difficulty of capturing Pol IV transcripts, but now we have an additional line of evidence.

With the identification of Pol V transcripts we have gained insight into RDDM and will hopefully open up the field a bit more. One of the most interesting points was the high number of Pol V transcripts that were not dependent on Pol IV, suggesting that Pol V transcription can be guided by non-CHH methylation. What exactly Pol V is doing in these cases is still unclear. It may be the case that any DNA methylation is able to recruit Pol V transcription, or that we are missing something fundamental about Pol V

activity. Additionally, the lack of strandedness for many Pol V sites indicates that Pol V simply needs to be recruited and may have less specificity compared to other RNA polymerases such as Pol II. The higher error rate of Pol V also supports this model where it is less specific due to the nature of its transcripts being used for scaffolding. Most importantly, the success of this technique demonstrates the additional applications it can be used for.

With the tools to analyze nascent transcription, we are now able to look at an environmental response within minutes and without the complication of filtering out stable transcripts seen in RNA-seq. Most early studies of responses have been on the order of hours and therefore included several thousand gene targets. Using GRO-seq, the total number of target genes is reduced dramatically and we can now identify primary responses, which may be key to manipulating entire pathways rather than just the end targets. Additionally, many of the common pathways studied in plants such as light signaling and hormone signaling are incredibly interwoven because they make use of many of the same downstream targets for phenotypes such as hypocotyl growth. By identifying the earliest responders of pathways we can begin to differentiate what is actually an environmental response and what is simply a target gene that is used by multiple pathways. We have tested several environmental responses including shade, heat, and high temperature and the preliminary results are promising that we will be able to identify new gene targets for manipulating these pathways. By combining GRO-seq, RNA-seq, and proteomic studies, plant biology can move forward with building high-resolution signal transduction pathway models and therefore identify better targets for use in applied studies such as increasing crop yield.

Along with the plant specific work, the robustness of the method developed for *Arabidopsis* opens up the study of other species and whole tissue types. Previous GRO-

seq studies largely focused on cell lines and embryos due to the ease of the nuclei isolation and the relatively cleanliness of the method prep. With the method in this dissertation it is possible to use this technique for species ranging from sea slugs to whole honey bees. We have tested many of these species ourselves and will hopefully complete an in-depth study in future work. The ability to study a wide variety of species means developing an incredibly comprehensive view of transcription across evolutionary branches and the prospect of identifying new regulatory elements, TSSs, and more in a wide variety of model organisms that may be poorly annotated or difficult to work with. As we are continuously drawing on models developed in other model systems, it would be exciting to be able to see how plants compare to other eukaroytes and what makes them unique.

Overall, this work will hopefully provide new tools and a method for future researchers to take advantage of. It is an incredibly exciting time in the world of biology as technology continues to move forward making experiments possible that could only be dreamed of as little as 10 years ago. With the ability to study crops directly and to establish pathways in model systems like *Arabidopsis* we can build models of growth, yield, resistance, and any other traits we desire and leverage this information into the next generation of agricultural endeavors to face the challenges of our booming population. The field still has a substantial amount of work to do given the rapid growth rate and our inability to match it with crop yield, but every study is another step forward, and every new method is a valuable tool to be considered for future work.

REFERENCES

1 Dawson, I.G. and Johnson, J.E. (2016) Does Size Matter? A Study of Risk Perceptions of Global Population Growth. *Risk analysis : an official publication of the Society for Risk Analysis*

2 Lee, R. (2011) The outlook for population growth. *Science* 333, 569-573

3 Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H., Shinn, P., Stevenson, D.K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C.C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D.E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W.L., Berry, C.C., and Ecker, J.R. (2003) Genome-wide insertional mutagenesis of Arabidopsis thaliana. *Science* 301, 653-657

4 O'Malley, R.C., Barragan, C.C., and Ecker, J.R. (2015) A user's guide to the Arabidopsis T-DNA insertion mutant collections. *Methods in molecular biology* 1284, 323-342

5 Blumel, M., Dally, N., and Jung, C. (2015) Flowering time regulation in crops-what did we learn from Arabidopsis? *Current opinion in biotechnology* 32, 121-129

6 Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011) Global quantification of mammalian gene expression control. *Nature* 473, 337-342

7 de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M., and Vogel, C. (2009) Global signatures of protein and mRNA expression levels. *Molecular bioSystems* 5, 1512-1526

8 Cheadle, C., Fan, J., Cho-Chung, Y.S., Werner, T., Ray, J., Do, L., Gorospe, M., and Becker, K.G. (2005) Control of gene expression during T cell activation: alternate regulation of mRNA transcription and mRNA stability. *BMC genomics* 6, 75

9 Kadonaga, J.T. (2004) Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* 116, 247-257

10 Juven-Gershon, T. and Kadonaga, J.T. (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Developmental biology* 339, 225-229

11 Ohler, U. and Wassarman, D.A. (2010) Promoting developmental transcription. *Development* 137, 15-26

12 Smale, S.T. and Kadonaga, J.T. (2003) The RNA polymerase II core promoter. *Annual review of biochemistry* 72, 449-479

13 Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C., and Chambon, P. (1980) Promoter sequences of eukaryotic protein-coding genes. *Science* 209, 1406-1414

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

15 Purnell, B.A., Emanuel, P.A., and Gilmour, D.S. (1994) TFIID sequence recognition of the initiator and sequences farther downstream in Drosophila class II genes. *Genes & development* 8, 830-842

16 David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C.J., Bofkin, L., Jones, T., Davis, R.W., and Steinmetz, L.M. (2006) A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences of the United States of America* 103, 5320-5325

17 Yamashita, R., Sathira, N.P., Kanai, A., Tanimoto, K., Arauchi, T., Tanaka, Y., Hashimoto, S., Sugano, S., Nakai, K., and Suzuki, Y. (2011) Genome-wide characterization of transcriptional start sites in humans by integrative transcriptome analysis. *Genome research* 21, 775-789

18 Park, D., Morris, A.R., Battenhouse, A., and Iyer, V.R. (2014) Simultaneous mapping of transcript ends at single-nucleotide resolution and identification of widespread promoter-associated non-coding RNA governed by TATA elements. *Nucleic acids research* 42, 3736-3749

19 Thomas, M.C. and Chiang, C.M. (2006) The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology* 41, 105-178

20 Kwak, H., Fuda, N.J., Core, L.J., and Lis, J.T. (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339, 950-953

21 Churchman, L.S. and Weissman, J.S. (2012) Native elongating transcript sequencing (NET-seq). *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* Chapter 4, Unit 4 14 11-17

22 Core, L.J., Waterfall, J.J., and Lis, J.T. (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322, 1845-1848

23 Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature genetics* 46, 1311-1320

24 Melgar, M.F., Collins, F.S., and Sethupathy, P. (2011) Discovery of active enhancers through bidirectional expression of short transcripts. *Genome biology* 12, R113

25 Lam, M.T., Cho, H., Lesch, H.P., Gosselin, D., Heinz, S., Tanaka-Oishi, Y., Benner, C., Kaikkonen, M.U., Kim, A.S., Kosaka, M., Lee, C.Y., Watt, A., Grossman, T.R., Rosenfeld, M.G., Evans, R.M., and Glass, C.K. (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* 498, 511-515

26 Morton, T., Petricka, J., Corcoran, D.L., Li, S., Winter, C.M., Carda, A., Benfey, P.N., Ohler, U., and Megraw, M. (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *The Plant cell* 26, 2746-2760

27 Molina, C. and Grotewold, E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC genomics* 6, 25

28 Deaton, A.M. and Bird, A. (2011) CpG islands and the regulation of transcription. *Genes & development* 25, 1010-1022

29 Saxonov, S., Berg, P., and Brutlag, D.L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America* 103, 1412-1417

30 Yamamoto, Y.Y., Ichida, H., Abe, T., Suzuki, Y., Sugano, S., and Obokata, J. (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic acids research* 35, 6219-6226

31 Yamamoto, Y.Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., Seki, M., Shinozaki, K., and Abe, T. (2007) Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC genomics* 8, 67

32 Yamamoto, Y.Y., Yoshioka, Y., Hyakumachi, M., and Obokata, J. (2011) Characteristics of core promoter types with respect to gene structure and expression in Arabidopsis thaliana. *DNA research : an international journal for rapid publication of reports on genes and genomes* 18, 333-342

33 Yamamoto, Y.Y., Yoshitsugu, T., Sakurai, T., Seki, M., Shinozaki, K., and Obokata, J. (2009) Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *The Plant journal : for cell and molecular biology* 60, 350-362

34 Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D., Bito, H., Worley, P.F., Kreiman, G., and Greenberg, M.E. (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature* 465, 182-187

35 Haag, J.R. and Pikaard, C.S. (2011) Multisubunit RNA polymerases IV and V: purveyors of non-coding RNA for plant gene silencing. *Nature reviews. Molecular cell biology* 12, 483-492

36 Onodera, Y., Haag, J.R., Ream, T., Costa Nunes, P., Pontes, O., and Pikaard, C.S. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* 120, 613-622

37 Woychik, N.A. and Hampsey, M. (2002) The RNA polymerase II machinery: structure illuminates function. *Cell* 108, 453-463

38 Grummt, I. (2003) Life on a planet of its own: regulation of RNA polymerase I transcription in the nucleolus. *Genes & development* 17, 1691-1702

39 Schramm, L. and Hernandez, N. (2002) Recruitment of RNA polymerase III to its target promoters. *Genes & development* 16, 2593-2620

40 Ream, T.S., Haag, J.R., Wierzbicki, A.T., Nicora, C.D., Norbeck, A.D., Zhu, J.K., Hagen, G., Guilfoyle, T.J., Pasa-Tolic, L., and Pikaard, C.S. (2009) Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Molecular cell* 33, 192-203

41 Tucker, S.L., Reece, J., Ream, T.S., and Pikaard, C.S. (2010) Evolutionary history of plant multisubunit RNA polymerases IV and V: subunit origins via genome-wide and segmental gene duplications, retrotransposition, and lineage-specific subfunctionalization. *Cold Spring Harbor symposia on quantitative biology* 75, 285-297

42 Haag, J.R., Ream, T.S., Marasco, M., Nicora, C.D., Norbeck, A.D., Pasa-Tolic, L., and Pikaard, C.S. (2012) In vitro transcription activities of Pol IV, Pol V, and RDR2 reveal coupling of Pol IV and RDR2 for dsRNA synthesis in plant RNA silencing. *Molecular cell* 48, 811-818

43 Landick, R. (2009) Functional divergence in the growing family of RNA polymerases. *Structure* 17, 323-325

44 Law, J.A. and Jacobsen, S.E. (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics* 11, 204-220

45 Matzke, M.A. and Mosher, R.A. (2014) RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nature reviews. Genetics* 15, 394-408

46 Herr, A.J., Jensen, M.B., Dalmay, T., and Baulcombe, D.C. (2005) RNA polymerase IV directs silencing of endogenous DNA. *Science* 308, 118-120

47 Kanno, T., Huettel, B., Mette, M.F., Aufsatz, W., Jaligot, E., Daxinger, L., Kreil, D.P., Matzke, M., and Matzke, A.J. (2005) Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nature genetics* 37, 761-765

48 Pontier, D., Yahubyan, G., Vega, D., Bulski, A., Saez-Vasquez, J., Hakimi, M.A., Lerbs-Mache, S., Colot, V., and Lagrange, T. (2005) Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes & development* 19, 2030-2040

49 Li, S., Vandivier, L.E., Tu, B., Gao, L., Won, S.Y., Li, S., Zheng, B., Gregory, B.D., and Chen, X. (2015) Detection of Pol IV/RDR2-dependent transcripts at the genomic scale in Arabidopsis reveals features and regulation of siRNA biogenesis. *Genome research* 25, 235-245

50 Zhai, J., Bischof, S., Wang, H., Feng, S., Lee, T.F., Teng, C., Chen, X., Park, S.Y., Liu, L., Gallego-Bartolome, J., Liu, W., Henderson, I.R., Meyers, B.C., Ausin, I., and Jacobsen, S.E. (2015) A One Precursor One siRNA Model for Pol IV-Dependent siRNA Biogenesis. *Cell* 163, 445-455

51 Wierzbicki, A.T., Haag, J.R., and Pikaard, C.S. (2008) Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell* 135, 635-648

52 Wierzbicki, A.T. (2012) The role of long non-coding RNA in transcriptional gene silencing. *Current opinion in plant biology* 15, 517-522

53 Wierzbicki, A.T., Cocklin, R., Mayampurath, A., Lister, R., Rowley, M.J., Gregory, B.D., Ecker, J.R., Tang, H., and Pikaard, C.S. (2012) Spatial and functional relationships among Pol V-associated loci, Pol IV-dependent siRNAs, and cytosine methylation in the Arabidopsis epigenome. *Genes & development* 26, 1825-1836

54 Zhong, X., Hale, C.J., Law, J.A., Johnson, L.M., Feng, S., Tu, A., and Jacobsen, S.E. (2012) DDR complex facilitates global association of RNA polymerase V to promoters and evolutionarily young transposons. *Nature structural & molecular biology* 19, 870-875

55 Eun, C., Lorkovic, Z.J., Sasaki, T., Naumann, U., Matzke, A.J., and Matzke, M. (2012) Use of forward genetic screens to identify genes required for RNA-directed DNA methylation in Arabidopsis thaliana. *Cold Spring Harbor symposia on quantitative biology* 77, 195-204

56 Dvir, A., Conaway, J.W., and Conaway, R.C. (2001) Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Current opinion in genetics & development* 11, 209-214

57 Roeder, R.G. (2003) Lasker Basic Medical Research Award. The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen. *Nature medicine* 9, 1239-1244

58 Danino, Y.M., Even, D., Ideses, D., and Juven-Gershon, T. (2015) The core promoter: At the heart of gene expression. *Biochimica et biophysica acta* 1849, 1116-1131

59 Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., and Jacquier, A. (2009) Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* 457, 1038-1042

60 Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009) Divergent transcription: a new feature of active promoters. *Cell cycle* 8, 2557-2564

61 Erhard, K.F., Jr., Talbot, J.E., Deans, N.C., McClish, A.E., and Hollick, J.B. (2015) Nascent transcription affected by RNA polymerase IV in Zea mays. *Genetics* 199, 1107-1125

62 Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G.K., Khatun, J., Williams, B.A., Zaleski, C., Rozowsky, J., Roder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Bar, N.S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S.,

Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O.J., Park, E., Persaud, K., Preall, J.B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S.E., Hannon, G., Giddings, M.C., Ruan, Y., Wold, B., Carninci, P., Guigo, R., and Gingeras, T.R. (2012) Landscape of transcription in human cells. *Nature* 489, 101-108

63 Borner, T., Aleynikova, A.Y., Zubo, Y.O., and Kusnetsov, V.V. (2015) Chloroplast RNA polymerases: Role in chloroplast biogenesis. *Biochimica et biophysica acta* 1847, 761-769

64 Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322, 1851-1854

65 Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008) Divergent transcription from active promoters. *Science* 322, 1849-1851

66 Sullivan, A.M., Arsovski, A.A., Lempe, J., Bubb, K.L., Weirauch, M.T., Sabo, P.J., Sandstrom, R., Thurman, R.E., Neph, S., Reynolds, A.P., Stergachis, A.B., Vernot, B., Johnson, A.K., Haugen, E., Sullivan, S.T., Thompson, A., Neri, F.V., 3rd, Weaver, M., Diegel, M., Mnaimneh, S., Yang, A., Hughes, T.R., Nemhauser, J.L., Queitsch, C., and Stamatoyannopoulos, J.A. (2014) Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell reports* 8, 2015-2030

67 Pautler, M., Eveland, A.L., LaRue, T., Yang, F., Weeks, R., Lunde, C., Je, B.I., Meeley, R., Komatsu, M., Vollbrecht, E., Sakai, H., and Jackson, D. (2015) FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. *The Plant cell* 27, 104-120

68 Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R., Farnham, P.J., Hirst, M., Lander, E.S., Mikkelsen, T.S., and Thomson, J.A. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology* 28, 1045-1048

69 Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57-74

70 Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A.H., Thomson, J.A., Ren, B., and Ecker, J.R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462, 315-322

71 Luo, C., Sidote, D.J., Zhang, Y., Kerstetter, R.A., Michael, T.P., and Lam, E. (2013) Integrative analysis of chromatin states in Arabidopsis identified potential regulatory

mechanisms for natural antisense transcript production. *The Plant journal : for cell and molecular biology* 73, 77-90

72 Wang, X., Elling, A.A., Li, X., Li, N., Peng, Z., He, G., Sun, H., Qi, Y., Liu, X.S., and Deng, X.W. (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *The Plant cell* 21, 1053-1069

73 Zhu, B., Zhang, W., Zhang, T., Liu, B., and Jiang, J. (2015) Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures. *The Plant cell* 27, 2415-2426

74 Liang, G., He, H., and Yu, D. (2012) Identification of nitrogen starvation-responsive microRNAs in Arabidopsis thaliana. *PloS one* 7, e48951

75 He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., and Kinzler, K.W. (2008) The antisense transcriptomes of human cells. *Science* 322, 1855-1857

76 Swiezewski, S., Liu, F., Magusin, A., and Dean, C. (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* 462, 799-802

77 Costas, C., de la Paz Sanchez, M., Stroud, H., Yu, Y., Oliveros, J.C., Feng, S., Benguria, A., Lopez-Vidriero, I., Zhang, X., Solano, R., Jacobsen, S.E., and Gutierrez, C. (2011) Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nature structural & molecular biology* 18, 395-400

78 Xu, L., Zhao, Z., Dong, A., Soubigou-Taconnat, L., Renou, J.P., Steinmetz, A., and Shen, W.H. (2008) Di- and tri- but not monomethylation on histone H3 lysine 36 marks active transcription of genes involved in flowering time regulation and other processes in Arabidopsis thaliana. *Molecular and cellular biology* 28, 1348-1360

79 Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012) Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. *The Plant cell* 24, 2719-2731

80 Parry, T.J., Theisen, J.W., Hsu, J.Y., Wang, Y.L., Corcoran, D.L., Eustice, M., Ohler, U., and Kadonaga, J.T. (2010) The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes & development* 24, 2013-2018

81 Duttke, S.H., Lacadie, S.A., Ibrahim, M.M., Glass, C.K., Corcoran, D.L., Benner, C., Heinz, S., Kadonaga, J.T., and Ohler, U. (2015) Human promoters are intrinsically directional. *Molecular cell* 57, 674-684

82 Hamilton, T.L., Stoneley, M., Spriggs, K.A., and Bushell, M. (2006) TOPs and their regulation. *Biochemical Society transactions* 34, 12-16

83 Zuo, Y.C. and Li, Q.Z. (2011) Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility. *Genomics* 97, 112-120

84 Anamika, K., Gyenis, A., Poidevin, L., Poch, O., and Tora, L. (2012) RNA polymerase II pausing downstream of core histone genes is different from genes producing polyadenylated transcripts. *PloS one* 7, e38769

85 Chodavarapu, R.K., Feng, S., Bernatavichute, Y.V., Chen, P.Y., Stroud, H., Yu, Y., Hetzel, J.A., Kuo, F., Kim, J., Cokus, S.J., Casero, D., Bernal, M., Huijser, P., Clark, A.T., Kramer, U., Merchant, S.S., Zhang, X., Jacobsen, S.E., and Pellegrini, M. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466, 388-392

86 Mischo, H.E. and Proudfoot, N.J. (2013) Disengaging polymerase: terminating RNA polymerase II transcription in budding yeast. *Biochimica et biophysica acta* 1829, 174-185

87 Zhang, H., Rigo, F., and Martinson, H.G. (2015) Poly(A) Signal-Dependent Transcription Termination Occurs through a Conformational Change Mechanism that Does Not Require Cleavage at the Poly(A) Site. *Molecular cell* 59, 437-448

88 Jonkers, I. and Lis, J.T. (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nature reviews. Molecular cell biology* 16, 167-177

89 Schaukowitch, K., Joo, J.Y., Liu, X., Watts, J.K., Martinez, C., and Kim, T.K. (2014) Enhancer RNA facilitates NELF release from immediate early genes. *Molecular cell* 56, 29-42

90 Berk, A.J. (2000) TBP-like factors come into focus. *Cell* 103, 5-8

91 Duttke, S.H. (2015) Evolution and diversification of the basal transcription machinery. *Trends in biochemical sciences* 40, 127-129

92 Duttke, S.H., Doolittle, R.F., Wang, Y.L., and Kadonaga, J.T. (2014) TRF2 and the evolution of the bilateria. *Genes & development* 28, 2071-2076

93 Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15-21

94 Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nature methods* 9, 357-359

95 Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular cell* 38, 576-589

96 Robinson, J.T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011) Integrative genomics viewer. *Nature biotechnology* 29, 24-26

97 Wang, D., Garcia-Bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., Liu, W., Kaikkonen, M.U., Ohgi, K.A., Glass, C.K., Rosenfeld, M.G., and Fu, X.D. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474, 390-394

98 Huang, L., Jones, A.M., Searle, I., Patel, K., Vogler, H., Hubner, N.C., and Baulcombe, D.C. (2009) An atypical RNA polymerase involved in RNA silencing shares small subunits with RNA polymerase II. *Nature structural & molecular biology* 16, 91-93

99 Law, J.A., Vashisht, A.A., Wohlschlegel, J.A., and Jacobsen, S.E. (2011) SHH1, a homeodomain protein required for DNA methylation, as well as RDR2, RDM4, and chromatin remodeling factors, associate with RNA polymerase IV. *PLoS genetics* 7, e1002195

100 Luo, J. and Hall, B.D. (2007) A multistep process gave rise to RNA polymerase IV of land plants. *Journal of molecular evolution* 64, 101-112

101 Haag, J.R., Brower-Toland, B., Krieger, E.K., Sidorenko, L., Nicora, C.D., Norbeck, A.D., Irsigler, A., LaRue, H., Brzeski, J., McGinnis, K., Ivashuta, S., Pasa-Tolic, L., Chandler, V.L., and Pikaard, C.S. (2014) Functional diversification of maize RNA polymerase IV and V subtypes via alternative catalytic subunits. *Cell reports* 9, 378-390

102 Gao, Z., Liu, H.L., Daxinger, L., Pontes, O., He, X., Qian, W., Lin, H., Xie, M., Lorkovic, Z.J., Zhang, S., Miki, D., Zhan, X., Pontier, D., Lagrange, T., Jin, H., Matzke, A.J., Matzke, M., Pikaard, C.S., and Zhu, J.K. (2010) An RNA polymerase II- and AGO4-associated protein acts in RNA-directed DNA methylation. *Nature* 465, 106-109

103 Mosher, R.A., Melnyk, C.W., Kelly, K.A., Dunn, R.M., Studholme, D.J., and Baulcombe, D.C. (2009) Uniparental expression of PolIV-dependent siRNAs in developing endosperm of Arabidopsis. *Nature* 460, 283-286

104 Zhou, M. and Law, J.A. (2015) RNA Pol IV and V in gene silencing: Rebel polymerases evolving away from Pol II's rules. *Current opinion in plant biology* 27, 154-164

105 Clouse, S.D., Langford, M., and McMorris, T.C. (1996) A brassinosteroid-insensitive mutant in Arabidopsis thaliana exhibits multiple defects in growth and development. *Plant physiology* 111, 671-678

106 Li, J., Nagpal, P., Vitart, V., McMorris, T.C., and Chory, J. (1996) A role for brassinosteroids in light-dependent development of Arabidopsis. *Science* 272, 398-401

107 Kim, T.W. and Wang, Z.Y. (2010) Brassinosteroid signal transduction from receptor kinases to transcription factors. *Annual review of plant biology* 61, 681-704

108 Nam, K.H. and Li, J. (2002) BRI1/BAK1, a receptor kinase pair mediating brassinosteroid signaling. *Cell* 110, 203-212

109 Jaillais, Y., Belkhadir, Y., Balsemao-Pires, E., Dangl, J.L., and Chory, J. (2011) Extracellular leucine-rich repeats as a platform for receptor/coreceptor complex

formation. *Proceedings of the National Academy of Sciences of the United States of America* 108, 8503-8507

110 Li, J., Wen, J., Lease, K.A., Doke, J.T., Tax, F.E., and Walker, J.C. (2002) BAK1, an Arabidopsis LRR receptor-like protein kinase, interacts with BRI1 and modulates brassinosteroid signaling. *Cell* 110, 213-222

111 Li, J. and Chory, J. (1997) A putative leucine-rich repeat receptor kinase involved in brassinosteroid signal transduction. *Cell* 90, 929-938

112 Kinoshita, T., Cano-Delgado, A., Seto, H., Hiranuma, S., Fujioka, S., Yoshida, S., and Chory, J. (2005) Binding of brassinosteroids to the extracellular domain of plant receptor kinase BRI1. *Nature* 433, 167-171

113 Wang, X., Li, X., Meisenhelder, J., Hunter, T., Yoshida, S., Asami, T., and Chory, J. (2005) Autoregulation and homodimerization are involved in the activation of the plant steroid receptor BRI1. *Developmental cell* 8, 855-865

114 Hothorn, M., Belkhadir, Y., Dreux, M., Dabi, T., Noel, J.P., Wilson, I.A., and Chory, J. (2011) Structural basis of steroid hormone perception by the receptor kinase BRI1. *Nature* 474, 467-471

115 Santiago, J., Henzler, C., and Hothorn, M. (2013) Molecular mechanism for plant steroid receptor activation by somatic embryogenesis co-receptor kinases. *Science* 341, 889-892

116 Sun, Y., Han, Z., Tang, J., Hu, Z., Chai, C., Zhou, B., and Chai, J. (2013) Structure reveals that BAK1 as a co-receptor recognizes the BRI1-bound brassinolide. *Cell research* 23, 1326-1329

117 She, J., Han, Z., Kim, T.W., Wang, J., Cheng, W., Chang, J., Shi, S., Wang, J., Yang, M., Wang, Z.Y., and Chai, J. (2011) Structural insight into brassinosteroid perception by BRI1. *Nature* 474, 472-476

118 Geldner, N., Hyman, D.L., Wang, X., Schumacher, K., and Chory, J. (2007) Endosomal signaling of plant steroid receptor kinase BRI1. *Genes & development* 21, 1598-1602

119 Su, W., Liu, Y., Xia, Y., Hong, Z., and Li, J. (2012) The Arabidopsis homolog of the mammalian OS-9 protein plays a key role in the endoplasmic reticulum-associated degradation of misfolded receptor-like kinases. *Molecular plant* 5, 929-940

120 Noguchi, T., Fujioka, S., Choe, S., Takatsuto, S., Yoshida, S., Yuan, H., Feldmann, K.A., and Tax, F.E. (1999) Brassinosteroid-insensitive dwarf mutants of Arabidopsis accumulate brassinosteroids. *Plant physiology* 121, 743-752

121 Tanaka, N., Asami, Yoshida, Matsuo, and Okamoto (2003) Physiological Roles of Brassinosteroids in Early Growth of Arabidopsis: Brassinosteroids Have a Synergistic Relationship with Gibberellin as well as Auxin in Light-Grown Hypocotyl Elongation. *Journal of plant growth regulation*, 259-271

122 Wang, Z.Y., Seto, H., Fujioka, S., Yoshida, S., and Chory, J. (2001) BRI1 is a critical component of a plasma-membrane receptor for plant steroids. *Nature* 410, 380-383

123 Nemhauser, J.L., Maloof, J.N., and Chory, J. (2003) Building integrated models of plant growth and development. *Plant physiology* 132, 436-439

124 Halliday, K., Devlin, P.F., Whitelam, G.C., Hanhart, C., and Koornneef, M. (1996) The ELONGATED gene of Arabidopsis acts independently of light and gibberellins in the control of elongation growth. *The Plant journal : for cell and molecular biology* 9, 305-312

125 Weber, A.N., Morse, M.A., and Gay, N.J. (2004) Four N-linked glycosylation sites in human toll-like receptor 2 cooperate to direct efficient biosynthesis and secretion. *The Journal of biological chemistry* 279, 34589-34594

126 Bosch, D., Castilho, A., Loos, A., Schots, A., and Steinkellner, H. (2013) N-glycosylation of plant-produced recombinant proteins. *Current pharmaceutical design* 19, 5503-5512

127 Li, J.J., Bickel, P.J., and Biggin, M.D. (2014) System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2, e270

128 Jovanovic, M., Rooney, M.S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., Rodriguez, E.H., Fields, A.P., Schwartz, S., Raychowdhury, R., Mumbach, M.R., Eisenhaure, T., Rabani, M., Gennert, D., Lu, D., Delorey, T., Weissman, J.S., Carr, S.A., Hacohen, N., and Regev, A. (2015) Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 347, 1259038

129 Battle, A., Khan, Z., Wang, S.H., Mitrano, A., Ford, M.J., Pritchard, J.K., and Gilad, Y. (2015) Genomic variation. Impact of regulatory variation from RNA to protein. *Science* 347, 664-667