Title

Nascent RNA sequencing reveals distinct features in plant transcription

Journal

Proceedings of the National Academy of Sciences of the United States of America, 113(43)

ISSN

0027-8424

Authors

Hetzel, Jonathan
Duttke, Sascha H
Benner, Christopher
et al.

Publication Date

2016-10-25

DOI

10.1073/pnas.1603217113

Peer reviewed

Nascent RNA sequencing reveals distinct features in plant transcription

Jonathan Hetzel[a,b,c,1], Sascha H. Duttke[a,b,c,1], Christopher Benner[d,2], and Joanne Chory[a,b,c,2]

[a]Plant Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; [b]Division of Biological Sciences, University of California, San Diego, La Jolla, CA 92093; [c]Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92037; and [d]School of Medicine, University of California, San Diego, La Jolla, CA 92093

Contributed by Joanne Chory, September 7, 2016 (sent for review February 29, 2016; reviewed by Erich Grotewold, Ryan Lister, and Nicholas J. Provart)

Transcriptional regulation of gene expression is a major mechanism used by plants to confer phenotypic plasticity, and yet compared with other eukaryotes or bacteria, little is known about the design principles. 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 maize data to capture characteristics of plant transcription. De novo annotation of nascent transcripts accurately mapped start sites and unstable transcripts. Examining the promoters of coding and noncoding transcripts identified comparable chromatin signatures, a conserved "TGT" core promoter motif and unreported transcription factor-binding sites. Mapping of engaged RNA polymerases showed a lack of enhancer RNAs, promoter-proximal pausing, and divergent transcription in *Arabidopsis* seedlings and maize, which are commonly present in yeast and humans. In contrast, *Arabidopsis* and maize genes accumulate RNA polymerases in proximity of the polyadenylation site, a trend that coincided with longer genes and CpG hypomethylation. Lack of promoter-proximal pausing and a higher correlation of nascent and steady-state transcripts indicate *Arabidopsis* may regulate transcription predominantly at the level of initiation. Our findings provide insight into plant transcription and eukaryotic gene expression as a whole.

plant transcription | nascent transcripts | RNA polymerase pausing | 5′GRO-seq | GRO-seq

Gene expression is a hallmark of life and subject to adaptation in changing environments. Steady-state transcript levels are a result of transcription initiation, elongation, and termination, followed by maturation and decay. Much has been learned about transcriptional mechanisms using yeast and animal models. In contrast, owing to technical difficulties created by plant cell extracts, there remains a large gap in knowledge in plant transcription. Plants and animals diverged more than 1.6 billion years ago. Studying plant transcription therefore not only contributes to a better understanding of the world's largest food source but also the evolution of eukaryotic gene expression.

The signals initiating transcription 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) (1, 2). At the core promoter, located approximately ±50 bp relative to the TSS, basal TFs cooperate with conserved DNA sequence motifs to orchestrate recruitment of the RNA polymerase (RNAP) (1, 3). Transcription has been studied extensively in a number of species (1–3) but not in plant model systems. Studies focusing on promoter-enriched sequences were hindered by the lack of precise TSSs (4, 5) but have improved dramatically through techniques such as paired end analysis of transcription start sites (3PEAT) (6) and cap analysis gene expression (CAGE) (7), but both methods are affected by RNA processing and transcript stability.

To comprehensively study global transcription it is essential to map all transcripts, regardless of RNA stability. Nascent RNA sequencing by global nuclear run-on sequencing (GRO-seq) (8), precision nuclear run-on sequencing (PRO-seq) (9), or native elongating transcript sequencing (NET-seq) (10) highlighted the abundance of unstable transcripts in some eukaryotes such as yeast and mammals (11), and yet these methods have been difficult to perform in plants. GRO-seq was recently used in maize seedlings and provided important insight into monocot transcription (12) but with limited TSS data and the omission of sarkosyl during the run-on reaction. Sarkosyl is required to block RNAP initiation, unhindered elongation, and efficient pause release (13, 14). We thus sought to optimize traditional GRO-seq for plants using *Arabidopsis* as a model with the aim to make it readily available to the community.

Here, we report an adapted GRO-seq method (8), as well as a new version of HOMER (15), to facilitate analysis of plant next-generation sequencing (NGS) data. In this study, we focus on 7meG-capped transcripts as generated by RNAP II from 6-day-old *Arabidopsis* seedlings to identify transcripts encoding protein-coding genes, microRNAs (miRNAs), and other noncoding RNAs. De novo annotation of nascent transcripts revealed many unstable noncoding transcripts, although these transcripts were underrepresented in *Arabidopsis* compared with mammals. Motif analysis identified previously unreported promoter motifs and revealed comparable structures for promoters of coding and noncoding transcripts. Nascent RNA sequencing highlighted the lack of divergent transcripts and promoter-proximal pausing but prominent 3′ pausing that was also apparent in maize. Together, these data affirm distinct features of plant transcription and demonstrate remarkable diversity in the regulation of eukaryotic transcription.

Significance

Transcription is a fundamental and dynamic step in the regulation of gene expression, but the characteristics of plant transcription are poorly understood. We adapted the global nuclear run-on sequencing (GRO-seq) and 5′GRO-seq methods for plants and provide a plant version of the next-generation sequencing software HOMER (homer.ucsd.edu/homer/plants) to facilitate data analysis. Mapping nascent transcripts in *Arabidopsis thaliana* seedlings enabled identification of known and novel transcripts and precisely mapped their start sites, revealing distinct characteristics in plant transcription. Our modified method to map engaged RNA polymerases and nascent transcripts in primary tissues paves the way for comparative and response studies.

Author contributions: J.H., S.H.D., C.B., and J.C. designed research; J.H. and S.H.D. performed research; J.H., S.H.D., and C.B. contributed new reagents/analytic tools; J.H., S.H.D., and C.B. analyzed data; and J.H. and S.H.D. wrote the paper.

Reviewers: E.G., The Ohio State University; R.L., The University of Western Australia; and N.J.P., University of Toronto.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE83108).

[1] J.H. and S.H.D. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: cbenner@ucsd.edu or chory@salk.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1603217113/-/DCSupplemental.

Results

Nascent Transcript Profiles 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 *A. thaliana* seedlings (Fig. 1*A* and Fig. S1; see *SI Materials and Methods* for a detailed protocol). GRO-seq captures nascent RNA independent of RNA stability, thereby providing precise maps of engaged RNAP in a strand specific manner (8); 5′GRO-seq specifically enriches for cap-protected 5′ ends, facilitating TSS mapping of nascent transcripts at single-nucleotide resolution (16). Through enzymatic modifications, we enriched nascent transcripts produced by RNAP II (for details, see Fig. S2). We further profiled steady-state transcripts by conventional RNA-seq for comparison with nascent transcript levels. As exemplified for the gene At4g10180, GRO-seq reads align to the full transcript including introns, 5′GRO-seq enriches for 5′ fragments of the gene, and RNA-seq maps the mature, intron-less transcript (Fig. 1*B*). For our analysis, we expanded the HOMER (15) software for plants. In total, we observed active transcription covering ∼40% of the genome by GRO-seq and 28% by RNA-seq in 6-d-old *Arabidopsis* seedlings at 33 million reads. Although this number is in part dependent on sequencing depth (Fig. S3*A*), it notably differs from humans, where ∼75% of the genome was found to be transcribed across different cell lines, with no individual line transcribing more than 57% (17). The ratio of GRO-seq/RNA-seq coverage was 1.39 in *Arabidopsis*, which is significantly smaller than in humans at 1.93 (Wilcoxon *P* value < 0.01; Fig. 1*C*), suggesting there are fewer unstable transcripts and introns in *Arabidopsis*.

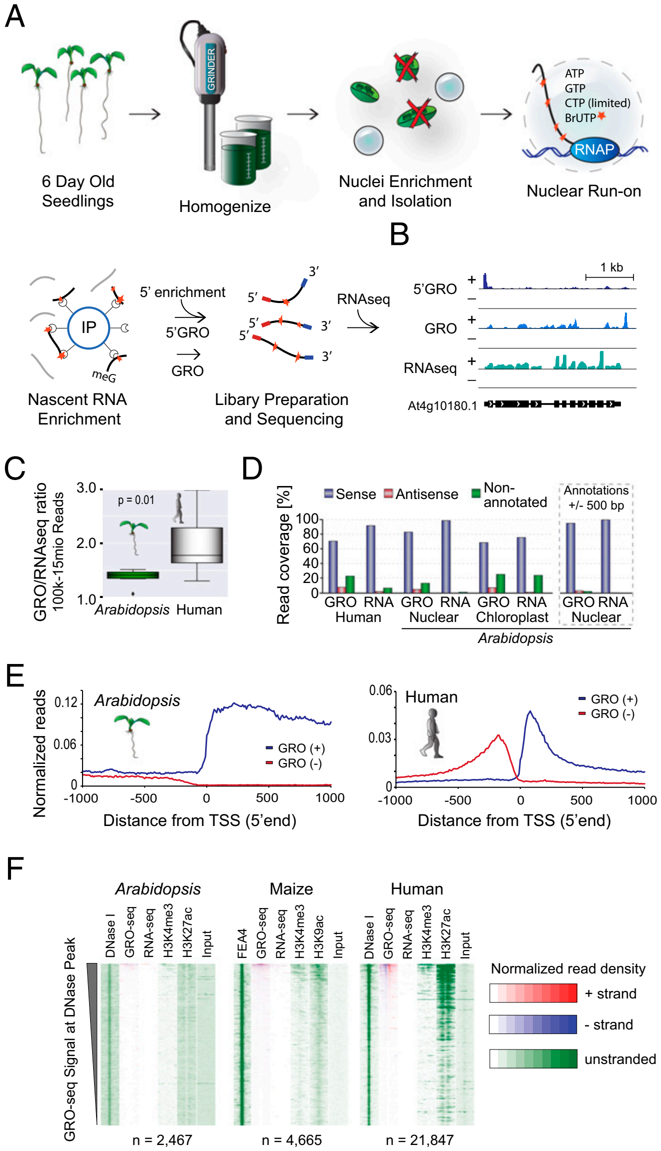


Fig. 1. GRO-seq reveals distinct features in *A. thaliana* transcription. (*A*) GRO-seq method in *Arabidopsis*. (*B*) Browser shot of sample gene At4g10180.1 with normalized read densities along the *y* axis. (*C*) Ratio of nascent/steady-state transcript genome coverage as a function of GRO-seq/RNA-seq coverage for *Arabidopsis* seedlings and human IMR-90 cells (8). The Wilcoxon test was used to calculate *P* value. (*D*) Distribution of RNA-seq and GRO-seq reads relative to annotations or extended annotations (±500 bp) (*Right*) for *Arabidopsis* and human IMR-90 cells. (*E*) Metaplot of GRO-seq signal from annotated genes normalized for reads per bp per gene along *y* axis for *Arabidopsis* and human IMR-90 cells. (*F*) Intergenic sites were defined by DNase-seq peaks [FEA4 ChIP-seq for maize (19)], and heat maps were generated ±1 kb from intergenic sites for signal from DNase-seq, GRO-seq, RNA-seq, H3K4me3, H3K9/27ac, and input in *Arabidopsis*, maize, and IMR-90 cells (8, 12, 18–25). 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. Although 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 ∼76% of the chloroplast as actively engaged with notable bi-directional transcription, demonstrating pervasive transcription in this organelle. This result demonstrates the potential for characterizing prokaryotic or viral polymerases using GRO-seq, even though this was not the goal of our study.

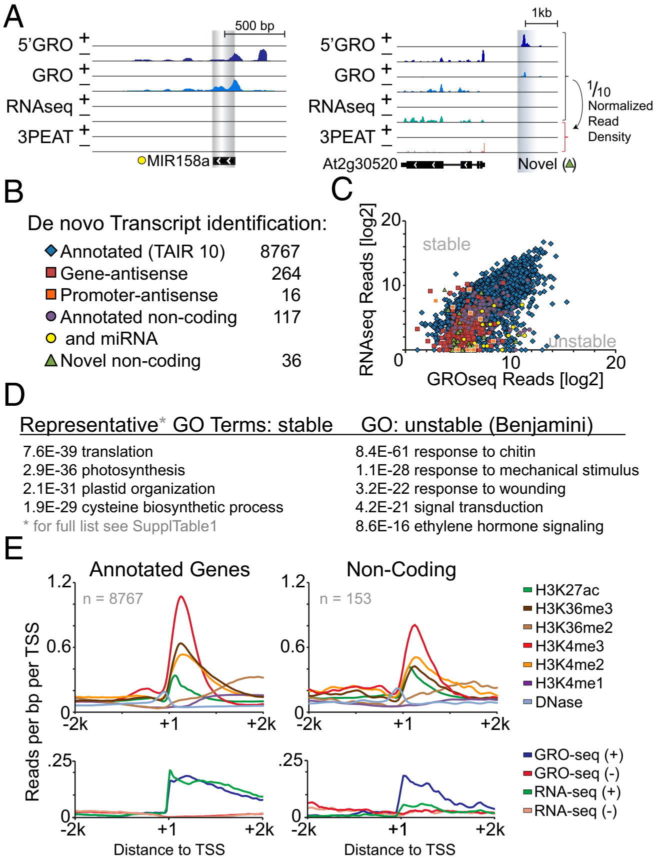
GRO-seq revealed 83% and 68% of engaged RNAPs occupy the sense strand of the *Arabidopsis* nuclear and chloroplast genomes, respectively. On the nuclear genome, ∼4% occupied the antisense strand of genes, and 13% mapped to unannotated regions. These numbers were significantly higher for the chloroplast genome, with 7% and 25% mapping to antisense genic and unannotated regions (Fig. 1*D*). By comparison, 98% of the RNA-seq mapped to the sense strand of nuclear genes. Together, these findings suggest that nuclear RNAPs are heavily engaged on the sense strand in *Arabidopsis*, particularly compared with humans (8). Indeed, expanding the annotations by 500 bp to either side increases the number of nuclear engaged RNAP to 95%, suggesting the majority of nonannotated transcription occurs directly adjacent to the annotated TSS and transcription termination site.

Notably, *Arabidopsis* seedlings lack significant divergent transcription as well as promoter-proximal pausing (Fig. 1*E*). To more thoroughly investigate these findings, we removed promoters within 1 kb of each other to prevent signal overlap appearing as promoter antisense transcription. Replotting the *Arabidopsis* and maize GRO-seq data (14) revealed striking directionality (Fig. S3*B*). It is important to note that the run-on reactions in maize were performed in the absence of sarkosyl, which blocks the initiation but not elongation of RNAP complexes (13) and strips off DNA-associated proteins such as histones (14). The prevalence of promoter-proximal pausing can thus not be ruled out in maize. However, the apparent lack of promoter-proximal pausing in both plant species argues that transcription is predominantly regulated at the level of initiation.

To investigate the presence of enhancer RNAs in plants, we mapped intergenic open chromatin regions using published DNaseI hypersensitivity (DNase-seq) data for *Arabidopsis* (18) and FASCIATED EAR 4 (FEA4) ChIP-seq (19) peaks for maize (due to a lack of DNase-seq data). In total, 2,467 putative intergenic enhancers were identified in *Arabidopsis* and 4,665 in maize compared with 21,847 in the human lung fibroblast IMR-90 cell line. Each site was sorted based on their GRO-seq signals, and heat maps were generated for ±1 kb from the intergenic site of

DNaseI chromatin accessibility. Very little GRO-seq, RNA-seq, or enhancer-associated chromatin marks (H3K9/27ac) were found in *Arabidopsis* and maize compared with humans, and both plants lack the distinctive bidirectional transcription common at mammalian enhancers (Fig. 1*F*) (8, 12, 18–25). Given these data, it appears that if plants have enhancer elements, they rarely, if at all, produce transcripts and therefore differ from mammalian enhancers.

Nascent Transcript Identification. Unlike RNA-seq, which measures steady-state levels of RNA species, GRO-seq captures nascent transcripts independent of transcript stability (8, 16). This method can be exemplified by the microRNA MIR158A (Fig. 2*A*) (26). The annotated miRNA used in previous studies is 100 bp, whereas the actual primary transcript as mapped by GRO-seq is more than 1 kb in length and initiates several hundred base pairs upstream of the current annotation (26). Additionally, GRO-seq captures transcripts previously undetected by RNA-seq or 3PEAT (Fig. 2*A*). We therefore created an unbiased atlas of *Arabidopsis* transcription using de novo transcript discovery based on GRO-seq expression and 5′GRO-seq to annotate the 5′ ends of each transcript. This identified 9,200 high-confidence transcripts defined by a continuous transcribed region (>10 reads GRO-seq) with a TSS defined by 5′GRO (>threefold-enriched) for 8,767 annotated protein-coding genes, 264 gene-antisense, 16 promoter-antisense, 117 annotated noncoding, and 36 unannotated intergenic transcripts (Fig. 2*B* and Dataset S1). Intergenic transcripts were unstable and significantly less abundant than described in human cell lines (8, 27, 28). However, gene-antisense RNAs, which were described as modulators of gene expression (29), are more enriched in *Arabidopsis*, suggesting an increase in antisense gene regulation in plants compared with humans (12, 27, 29). Comparison of GRO-seq and RNA-seq transcript levels at each transcript provides an estimate of transcript stability because nascent transcripts are unaffected by degradation. Plotting the de novo transcripts with respect to their GRO-seq and RNA-seq levels revealed a range of stability for annotated protein-coding genes compared with general instability for noncoding annotated RNAs, miRNAs, promoter-antisense transcripts, and unannotated noncoding transcripts (Fig. 2*C*). The transcripts start sites largely agreed with those defined using 3PEAT in roots (7), but noncoding transcripts were more exclusive to 5′GRO-seq (Fig. S4).



De novo Transcript identification:	
◆ Annotated (TAIR 10)	8767
■ Gene-antisense	264
■ Promoter-antisense	16
● Annotated non-coding	117
○ and miRNA	
▲ Novel non-coding	36

Representative* GO Terms: stable	GO: unstable (Benjamini)
7.6E-39 translation	8.4E-61 response to chitin
2.9E-36 photosynthesis	1.1E-28 response to mechanical stimulus
2.1E-31 plastid organization	3.2E-22 response to wounding
1.9E-29 cysteine biosynthetic process	4.2E-21 signal transduction
* for full list see SupplTable1	8.6E-16 ethylene hormone signaling

Fig. 2. Genome-wide identification and characterization of nascent transcripts in *A. thaliana*. (*A*) Example browser shot for noncoding transcripts MIR158a and At2g30520. Relative read densities for 5′GRO/GRO and RNA-seq were scaled by 10% to enable visualization alongside the 3PEAT TSS mapping data at these loci. (*B*) Classification of de novo-identified high-confidence transcripts with HOMER. (*C*) RNA stability plot of RNA-seq reads vs. GRO-seq reads for identified transcript groups in B. (*D*) Representative list of GO terms with Benjamini values (multiple testing-corrected) identified for the most stable and unstable transcripts defined <eightfold more RNA-seq than GRO-seq over exons. Unstable transcripts were defined as having eightfold more GRO-seq than RNA-seq. (*E*) Metaplot at the TSSs of 8,767 annotated genes and 153 noncoding transcripts from *B* with normalized reads per base pair per TSS for chromatin modifications (18, 23, 25) (*Top*) and transcripts (*Bottom*).

Comparison of RNA-seq vs. GRO-seq for exon coverage of human IMR-90 cells revealed a higher variance and lower correlation than in *Arabidopsis* (*Arabidopsis*, $r^2 = 0.57$; Human, $r^2 = 0.32$; Fig. S5*A*), underlining a much tighter correlation between transcription and steady-state RNA levels in *Arabidopsis*. Only exons were used to avoid bias associated with differential intron length between species. Together with the absence of promoter-proximal pausing, this correlation proposes *Arabidopsis* transcription is more predominantly regulated at the level of transcription initiation compared with humans.

To investigate whether RNA stability was associated with biological functions, we performed gene ontology (GO) analysis of the most stable and unstable transcripts (>eightfold enriched; Dataset S2). Stable transcripts were associated with translation, photosynthesis, and metabolic functions, whereas unstable transcripts were enriched for stimulus response genes, signal transduction, and hormones. These findings are consistent with the biological theme that transcripts associated with essential processes are stable, whereas regulated genes tend to be less stable.

Differential analysis of the epigenetic landscape revealed unifying signatures at annotated protein-coding and noncoding of de novo transcripts. H3K4me3, H3K27ac, and H3K36me3 peaked in proximity to the +1 nucleosome and H3K4me2 slightly downstream (23, 25). DNase hypersensitivity overlapped with the promoter region, and H3K4me1 and H3K36me2 increase along the body of the gene (18, 23, 25), as expected (30, 31) (Fig. 2 *E*).

***Arabidopsis* Promoter Structures and Identification of the TGT Core Promoter Motif.** High-resolution nascent TSS data are provided by 5′GRO-seq, enabling investigation of promoter elements in a distance-specific manner. The core promoter region of the 9,200 identified *Arabidopsis* transcripts was remarkably GC-poor, with a strong AT enrichment around −30 bp, suggesting a predominant role of the TATA-box (Fig. 3*A*). This finding contrasts with human core promoters that are ordinarily GC-rich and only slightly enriched for the TATA-box (~10%). De novo motif analysis of initiation sites using HOMER underlined the strong prevalence of an Initiator element (Inr)-like motif (44.8%) and variations thereof (Fig. 3*B*). Notably, the *Arabidopsis* Inr consensus sequences “TYA(+1)YYN” and “TYA(+1)GGG” differ from the traditional Inr “TCAKTY” in *Drosophila* (32). Our analysis further identified an “initiator” that we termed the “TGT motif” at ~4% of *Arabidopsis* TSSs (Fig. 3*B*). Analysis of HeLa (28) and *Drosophila* S2 5′GRO-seq data revealed the TGT motif to be conserved in humans and flies (Fig. 3*C*). Compared with the enriched GO terms of genes associated with Inr-containing promoters, TGT-associated genes were slightly enriched for terms related to in negative regulation,

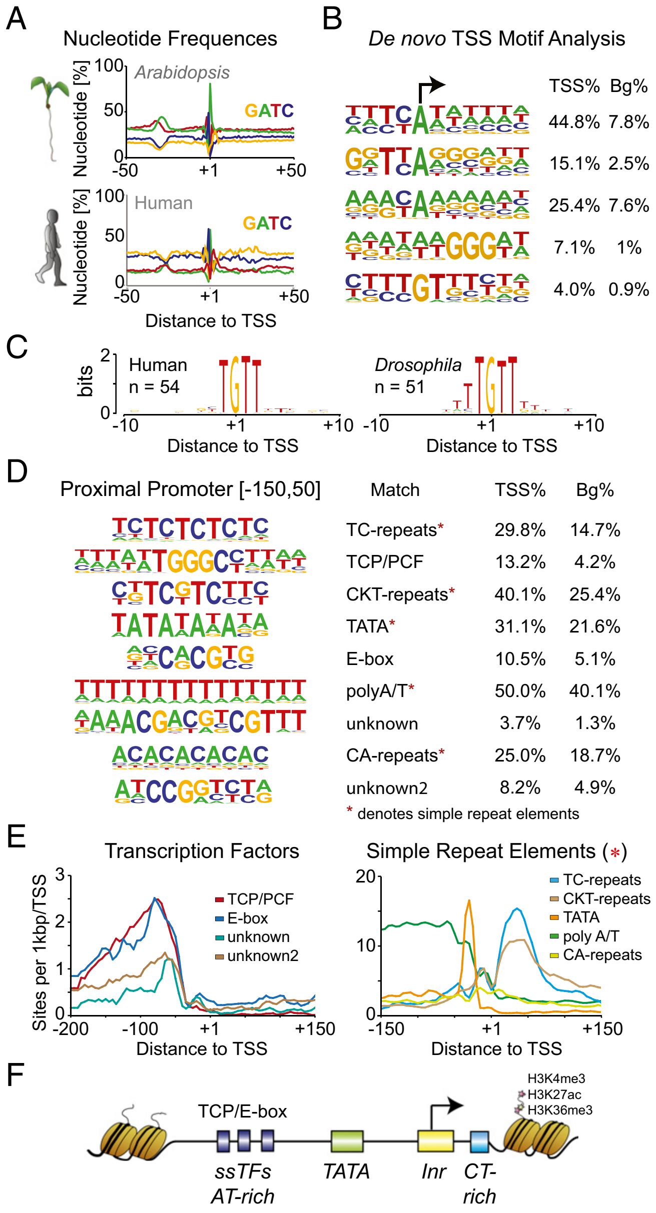


Fig. 3. *Arabidopsis* promoter features and motifs. (*A*) Metaplot of nucleotide frequency with respect to the +1 TSS as defined by 5′GRO-seq for annotated transcripts at the core promoter region of *Arabidopsis* and human HeLa samples (28). (*B*) Position-restricted de novo motif analysis of *Arabidopsis* initiating nucleotides using HOMER. Percentage of motifs at TSSs compared with background levels. (*C*) TGT motif as identified at *Drosophila* S2 and human (HeLa) TSSs. (*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*) Metaplot of the TF binding sites (*Left*) and simple repeat elements (*Right*) in distance to the TSS along the *x* axis. (*F*) Simplified schematic of regulatory sequence features at *Arabidopsis* promoters.

chromatin organization, gene silencing, and dsRNA response (Fig. S5*B*).

Sequence-specific TFs modulate gene expression and commonly bind the proximal promoter region. De novo motif analysis using HOMER highlighted a predominance of simple sequence repeats in *Arabidopsis* but also identified TF binding sites, two of which were unknown (Fig. 3 *D* and *E*). The most prevalent sequence patterns were TC and CKT repeats present downstream of the TSS and simple polyA/T repeats upstream thereof. AT-rich sequences were reported to inhibit nucleosome formation and aid DNA flexibility, thereby facilitating TF recruitment (33). The TATA-box was found in 31% of plant promoters, similar to the 29% reported previously (4). 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 enriched TF motifs were the enhancer box (E-box) and TEOSINTE-BRANCHED CYCLOIDEA/PROLIFERATING CELL FACTOR (TCP/PCF) elements, with 10.5% and 13.2%, respectively. The E-box is targeted by the basic helix–loop–helix TFs (34), and both the E-box and TCP motif are conserved among plants (35) and critical for development, which is reflected in their prevalence in the promoters of genes regulated during seedling development. Additionally, two unidentified motifs were found that have not been reported previously in *Arabidopsis* (Fig. 3 *D* and *E*). Using the Catalogue of Inferred Sequence Binding Preferences (CIS-BP) database (36), the closest but not exact match for unknown motif 1 was Sterol Uptake Control Protein 2 (UPC2) (NRWACGA), whereas unknown motif 2 matched best to Activator of Stress Genes 1 (ASG1) (WTCCGG), both belonging to the zinc cluster TF family in *Saccharomyces* (37). The factors binding these motifs in *Arabidopsis* remain to be identified. Cognate promoter motifs in TGT- or Inr-containing promoters did not differ notably with the TATA element slightly depleted and TCP/PCF elements enriched in TGT-containing promoters (Fig. S5*C*).

Given our classification of coding, noncoding, and antisense transcripts, we compared the promoters from each class. Although there were some minor differences (Fig. S5*D*), 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. A simplified model for the basal *Arabidopsis* promoter structure at this stage of development is proposed (Fig. 3*F*).

RNAP 3′ Accumulation. Although promoter-proximal pausing in *Arabidopsis* seedlings was not readily apparent, we noticed a sharp accumulation of RNAP adjacent to the 3′ polyadenylation site (PAS) (38) (Fig. 4*A*). Analysis of published RNAP II ChIP-seq (40) also showed a clear increase in 3′ paused polymerase (Fig. 4*B*), whereas the RNA-seq signal approaches zero as expected; 3′ pausing at a lesser extent was previously described (8, 39), but a mechanism remains elusive. 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* and *E*). A breakdown of genes by length revealed several distinct characteristics between the GRO-seq signal, RNA-seq signal, and amount of 3′ RNAP accumulation (Fig. 4*D*). Longer genes show a higher accumulation in the GRO-seq reads compared with RNA-seq. Reanalyzing maize GRO-seq data (12) showed similar strong 3′ accumulation of nascent RNA that increased with gene length (Fig. S6*A*), suggesting that the higher level of 3′ pausing for longer genes is a characteristic of plant transcription.

In addition to gene length, CpG methylation was associated with 3′ pausing. In *Arabidopsis*, CpG methylation was excluded from the 3′ pausing sites and promoter regions of transcribed genes, whereas CpG methylation is only excluded from the promoter region in mammalian systems (22) (Fig. 4*E*). In aggregate, long genes showed an average maximum of CpG body methylation at ~40% in the middle of the gene, which drops below 10% at their 5′ and 3′ ends. However, short genes rarely showed over 15% body methylation and drop comparatively minimally at the 3′ end, demonstrating a connection between CpG methylation and 3′ pausing. In contrast, CHG methylation 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*. Although a comparable MethylC-sequencing

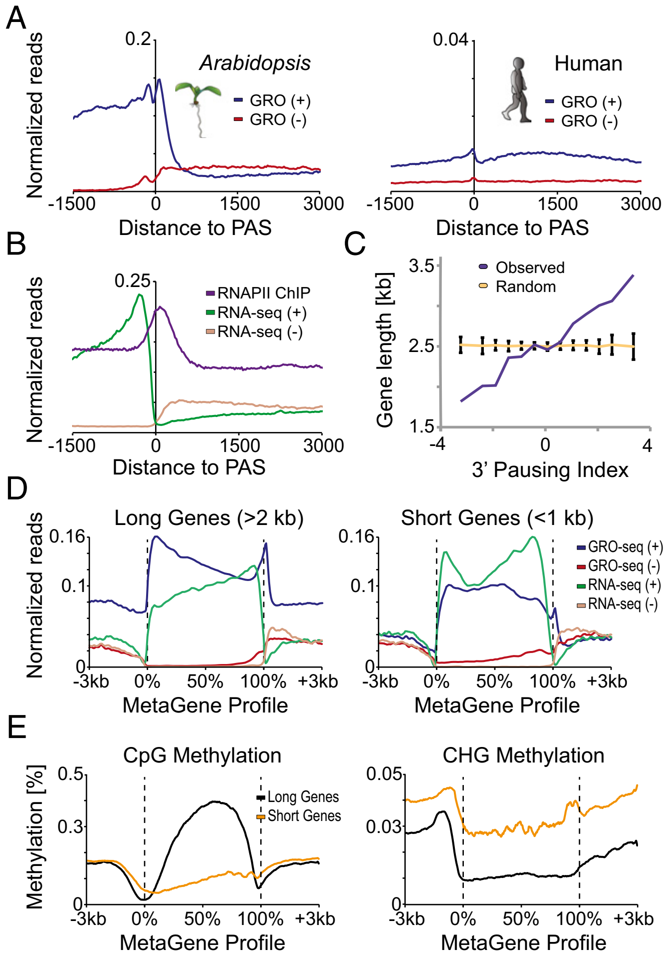


Fig. 4. *Arabidopsis* shows extensive 3′ RNAP accumulation in proximity to the PAS followed by rapid transcription termination. (*A*) Metaplot of GRO-seq signal from TAIR10 annotated genes for *Arabidopsis* and human samples (8). Reads were normalized as reads per base pair per gene. (*B*) RNAP II accumulation as shown by ChIP-Seq (purple) (40) with RNA-seq (+) (green) and (-) (tan) signal around the PAS. (*C*) The average gene length as a function of 3′ pausing index for expressed genes (gene body fragments per kilobase of transcript per million mapped reads: >5; at least 10 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 with the reads from 0 to −250. Randomized data are shown in tan, and SDs were calculated based on 1,000 randomizations. (*D*) Metaplots anchored by TSS (0%) and PAS (100%) of GRO-seq and RNA-seq for genes >2.0 kb (*Left*) and <1 kb (*Right*) in total length. Reads were normalized as per base pair per gene. (*E*) CpG methylation (*Left*) and CHG methylation (*Right*) (50) were plotted as percentage methylation along the normalized gene body.

dataset did not exist in maize, we were able to reanalyze methyl-DNA immunoprecipitation (24) to show methylation is generally excluded from both the 3′ and 5′ ends of genes as seen in *Arabidopsis* (Fig. S6*B*). These data demonstrate a connection between 3′ pausing of engaged RNAP and CpG methylation as another distinct characteristic of plant transcription.

Discussion

This study has put forward a GRO-seq method for mapping engaged RNAP at a genome-wide level in primary plant tissue. The identification of nascent transcripts and definition of TSSs revealed distinct characteristics of *Arabidopsis* transcription and their connection to other eukaryotic systems. The lack of divergent transcription in *Arabidopsis* and likely maize contrasts with the notion that eukaryotic promoters are inherently divergent (41). Highly directional initiation of transcription was also observed in *Drosophila* (38). Notably, both *Arabidopsis* and *Drosophila* display strong core promoter signatures, suggesting a prominent role for the core promoter and its motifs in mediating transcriptional directionality. *Arabidopsis* core promoters were enriched for distinct Inr-like motifs and the TATA-box with 80% and 30%, respectively. The strong prevalence of these motifs may be due to developmental timing. On the other hand, despite commonly containing more than one copy of the TATA-box binding protein (TBP) gene, plants lack TBP-related factors (42). In bilateral symmetric animals, these factors were shown to support different transcription systems, enabling regulatory diversity through core promoter motif diversity (42, 43). *Arabidopsis*, on the other hand, encodes two additional eukaryotic RNAPs: RNAP IV and RNAP V, which are integral to the repression of a subset of genes and transposons through RNA-directed DNA methylation (44). These additional RNAPs may reflect a different evolutionary approach to increasing the regulatory diversity of the genome.

GRO-seq identified 9,200 transcripts in 6-d-old *Arabidopsis* seedlings, of which only 153 were noncoding transcripts generated by RNAP II. This number is considerably less than in humans (8, 17). Plants lack enhancer RNAs (eRNAs) but notably also the NEGATIVE ELONGATION FACTOR (NELF) involved in promoter-proximal pausing (45). eRNAs were reported to mediate release of NELF-dependent pausing (46). Therefore, given the absence of NELF, potential eRNAs may not have provided the same selective advantages in plants. In contrast, however, Zhu et al. (25) 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. Cell-type specific nuclei obtained using isolation of nuclei tagged in specific cell types (INTACT) (47) or nuclear-localized reporters in combination with fluorescence-activated cell sorting may clarify these results.

Capturing nascent transcripts enabled characterization of basic features. Absence of promoter-proximal pausing, together with a high correlation between nascent and steady-state transcript levels argues that *Arabidopsis* transcription is predominantly regulated at the level of initiation. Li et al. (35) reported transcription to be the most regulated step in human gene regulation. In this light, transcription initiation may be the major step of gene regulation in plants.

RNAP pausing, a major regulator of transcription elongation in mammals (48), was observed predominantly downstream of the PAS in *Arabidopsis* and maize. The underlying mechanism is unknown but is likely a common feature in plant transcription. 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 (39). This idea may hold true in plants based on the higher GRO-seq signal compared with RNA-seq for longer *Arabidopsis* genes, which also show higher amounts of 3′ pausing compared with shorter genes. In addition, DNA methylation was shown to slow down transcription elongation (49), and yet the exact role of gene body methylation in plants is still unclear.

In summary, we have described a method for the analysis of nascent transcripts in primary tissue and provide a high-resolution map of *Arabidopsis* transcripts. GRO-seq opens up avenues to study transcriptional regulation or responses to stimuli at a specific moment in time. We envision that this technical advance will facilitate a better understanding of gene regulation in plants but also eukaryotic transcription in general.

Materials and Methods

A. thaliana (Col-0) 6-d-old seedlings were grown on half Linsmaier and Skoog (LS) medium with 24 h of light at 22 °C. Tissue was mechanically homogenized and nuclei purified by centrifugation. GRO-seq and 5′ GRO-seq are described detailed in *SI Materials and Methods*. Briefly, 5 × 10^6 nuclei were run on, and DNase/proteinase treated, and RNA was extracted using TRIzol. RNA was digested with Terminator 5′–Phosphate-Dependent Exonuclease (Epicentre) before fragmentation. Nascent RNA was enriched twice for 5-bromo-UTP (BrUTP) by immunoprecipitation. After end repair, RNA 5′ pyrophosphohydrolase (RppH) was used for decapping and library prepared using the NEB Next Small RNA Library Prep Set. Data were analyzed using $\text{HOMER}_{\text{plants}}$, accessible at homer.ucsd.edu/homer/plants/.

ACKNOWLEDGMENTS. We thank J. Kadonaga, C. Glass, and Ira Schildkraut for feedback and suggestions. The work was supported by J.C.'s Howard Hughes Medical Institute funding and NIH Grants R01GM094428 and R01GM52413. J.H. was supported by NIH Grant T32GM007240, the Rose Hills Foundation, and the H. A. and Mary K. Chapman Charitable Trust. S.H.D. is the recipient of the University of California at San Diego Molecular Biology/Cancer Center Fellowship and a CRI-Irvington Fellow.

1. Smale ST, Kadonaga JT (2003) The RNA polymerase II core promoter. *Annu Rev Biochem* 72:449–479.
2. Dvir A, Conaway JW, Conaway RC (2001) Mechanism of transcription initiation and promoter escape by RNA polymerase II. *Curr Opin Genet Dev* 11(2):209–214.
3. Danino YM, Even D, Ideses D, Juven-Gershon T (2015) The core promoter: At the heart of gene expression. *Biochim Biophys Acta* 1849(8):1116–1131.
4. Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* 6:25.
5. Yamamoto YY, et al. (2009) Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *Plant J* 60(2):350–362.
6. Morton T, et al. (2014) Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures. *Plant Cell* 26(7):2746–2760.
7. Mejía-Guerra MK, et al. (2015) Core promoter plasticity between maize tissues and genotypes contrasts with predominance of sharp transcription initiation sites. *Plant Cell* 27(12):3309–3320.
8. Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* 322(5909):1845–1848.
9. Kwak H, Fuda NJ, Core LJ, Lis JT (2013) Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* 339(6122):950–953.
10. Churchman LS, Weissman JS (2012) Native elongating transcript sequencing (NET-seq). *Curr Protoc Mol Biol* Chapter 4:Unit 4.14.1–4.14.17.
11. Seila AC, Core LJ, Lis JT, Sharp PA (2009) Divergent transcription: A new feature of active promoters. *Cell Cycle* 8(16):2557–2564.
12. Erhard KF, Jr, Talbot JE, Deans NC, McClish AE, Hollick JB (2015) Nascent transcription affected by RNA polymerase IV in Zea mays. *Genetics* 199(4):1107–1125.
13. Hawley DK, Roeder RG (1985) Separation and partial characterization of three functional steps in transcription initiation by human RNA polymerase II. *J Biol Chem* 260(13):8163–8172.
14. Rougvie AE, Lis JT (1988) The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell* 54(6):795–804.
15. Heinz S, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576–589.
16. Lam MT, et al. (2013) Rev-Erbs repress macrophage gene expression by inhibiting enhancer-directed transcription. *Nature* 498(7455):511–515.
17. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.
18. Sullivan AM, et al. (2014) Mapping and dynamics of regulatory DNA and transcription factor networks in A. thaliana. *Cell Reports* 8(6):2015–2030.
19. Pautler M, et al. (2015) FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. *Plant Cell* 27(1):104–120.
20. Bernstein BE, et al. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* 28(10):1045–1048.
21. Consortium EP; ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.
22. Lister R, et al. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322.
23. Luo C, et al. (2013) Integrative analysis of chromatin states in Arabidopsis identified potential regulatory mechanisms for natural antisense transcript production. *Plant J* 73(1):77–90.
24. Wang X, et al. (2009) Genome-wide and organ-specific landscapes of epigenetic modifications and their relationships to mRNA and small RNA transcriptomes in maize. *Plant Cell* 21(4):1053–1069.
25. Zhu B, Zhang W, Zhang T, Liu B, Jiang J (2015) Genome-wide prediction and validation of intergenic enhancers in Arabidopsis using open chromatin signatures. *Plant Cell* 27(9):2415–2426.
26. Liang G, He H, Yu D (2012) Identification of nitrogen starvation-responsive microRNAs in Arabidopsis thaliana. *PLoS One* 7(11):e48951.
27. He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW (2008) The antisense transcriptomes of human cells. *Science* 322(5909):1855–1857.
28. Duttke SH, et al. (2015) Human promoters are intrinsically directional. *Mol Cell* 57(4):674–684.
29. Swiezewski S, Liu F, Magusin A, Dean C (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature* 462(7274):799–802.
30. Costas C, et al. (2011) Genome-wide mapping of Arabidopsis thaliana origins of DNA replication and their associated epigenetic marks. *Nat Struct Mol Biol* 18(3):395–400.
31. Xu L, et al. (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. *Mol Cell Biol* 28(4):1348–1360.
32. Juven-Gershon T, Kadonaga JT (2010) Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol* 339(2):225–229.
33. Zuo YC, Li QZ (2011) Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility. *Genomics* 97(2):112–120.
34. Carretero-Paulet L, et al. (2010) Genome-wide classification and evolutionary analysis of the bHLH family of transcription factors in Arabidopsis, poplar, rice, moss, and algae. *Plant Physiol* 153(3):1398–1412.
35. Li S (2015) The Arabidopsis thaliana TCP transcription factors: A broadening horizon beyond development. *Plant Signal Behav* 10(7):e1044192.
36. Weirauch MT, et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6):1431–1443.
37. de Boer CG, Hughes TR (2012) YeTFaSCo: A database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res* 40(Database issue):D169–D179.
38. Nechaev S, et al. (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science* 327(5963):335–338.
39. Anamika K, Gyenis À, Poidevin L, Poch O, Tora L (2012) RNA polymerase II pausing downstream of core histone genes is different from genes producing polyadenylated transcripts. *PLoS One* 7(6):e38769.
40. Chodavarapu RK, et al. (2010) Relationship between nucleosome positioning and DNA methylation. *Nature* 466(7304):388–392.
41. Preker P, et al. (2008) RNA exosome depletion reveals transcription upstream of active human promoters. *Science* 322(5909):1851–1854.
42. Duttke SH, Doolittle RF, Wang YL, Kadonaga JT (2014) TRF2 and the evolution of the bilateria. *Genes Dev* 28(19):2071–2076.
43. Duttke SH (2015) Evolution and diversification of the basal transcription machinery. *Trends Biochem Sci* 40(3):127–129.
44. Law JA, Jacobsen SE (2010) Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nat Rev Genet* 11(3):204–220.
45. Wu CH, et al. (2003) NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila. *Genes Dev* 17(11):1402–1414.
46. Schaukowitch K, et al. (2014) Enhancer RNA facilitates NELF release from immediate early genes. *Mol Cell* 56(1):29–42.
47. Deal RB, Henikoff S (2011) The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana. *Nat Protoc* 6(1):56–68.
48. Jonkers I, Lis JT (2015) Getting up to speed with transcription elongation by RNA polymerase II. *Nat Rev Mol Cell Biol* 16(3):167–177.
49. Rountree MR, Selker EU (1997) DNA methylation inhibits elongation but not initiation of transcription in Neurospora crassa. *Genes Dev* 11(18):2383–2395.
50. Lei M, et al. (2015) Regulatory link between DNA methylation and active demethylation in Arabidopsis. *Proc Natl Acad Sci USA* 112(11):3553–3557.
51. Dobin A, et al. (2013) STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29(1):15–21.
52. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9(4):357–359.
53. Robinson JT, et al. (2011) Integrative genomics viewer. *Nat Biotechnol* 29(1):24–26.
54. Wang D, et al. (2011) Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. *Nature* 474(7351):390–394.