the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2024) **118,** 1241–1257 doi: 10.1111/tpj.16654

A normalization method that controls for total RNA abundance affects the identification of differentially expressed genes, revealing bias toward morning-expressed responses

Kanjana Laosuntisuk[1], Amaranatha Vennapusa[2], Impa M. Somayanda[3], Adam R. Leman[4], SV Krishna Jagadish[3,5] and Colleen J. Doherty[1,*]

[1]*Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, North Carolina, USA,*
[2]*Department of Agriculture and Natural Resources, Delaware State University, Dover, Delaware, USA,*
[3]*Department of Plant and Soil Science, Texas Tech University, Lubbock, Texas 79410, USA,*
[4]*Department of Science and Technology, The Good Food Institute, Washington, District of Columbia 20090, USA, and*
[5]*Department of Agronomy, Kansas State University, Manhattan, Kansas 66506, USA*

Received 27 October 2023; revised 12 January 2024; accepted 18 January 2024; published online 30 January 2024.
*For correspondence (e-mail colleen_doherty@ncsu.edu).

SUMMARY

RNA-Sequencing is widely used to investigate changes in gene expression at the transcription level in plants. Most plant RNA-Seq analysis pipelines base the normalization approaches on the assumption that total transcript levels do not vary between samples. However, this assumption has not been demonstrated. In fact, many common experimental treatments and genetic alterations affect transcription efficiency or RNA stability, resulting in unequal transcript abundance. The addition of synthetic RNA controls is a simple correction that controls for variation in total mRNA levels. However, adding spike-ins appropriately is challenging with complex plant tissue, and carefully considering how they are added is essential to their successful use. We demonstrate that adding external RNA spike-ins as a normalization control produces differences in RNA-Seq analysis compared to traditional normalization methods, even between two times of day in untreated plants. We illustrate the use of RNA spike-ins with 3′ RNA-Seq and present a normalization pipeline that accounts for differences in total transcriptional levels. We evaluate the effect of normalization methods on identifying differentially expressed genes in the context of identifying the effect of the time of day on gene expression and response to chilling stress in sorghum.

Keywords: gene expression, RNA-Seq, normalization methods, *Sorghum bicolor*, abiotic stress responses, diel transcriptional changes.

Linked article: This paper is the subject of a Research Highlight article. To view this Research Highlight article visit https://doi.org/10.1111/tpj.16791.

INTRODUCTION

RNA-Sequencing (RNA-Seq) is a high-throughput technology for genome-wide transcriptional analysis. RNA-Seq has been widely used in various research areas, including plant biology, to examine many aspects of RNA biology, including differentially expressed genes (DEGs), transcriptome assembly, alternative splicing, variant discovery, cis-regulatory elements, and roles of non-coding RNAs (Stark et al., 2019). As RNA-Seq was first introduced over a decade ago, many advanced RNA-Seq methods have been developed from the standard protocols to answer in-depth issues in molecular biology, especially for determining gene expression at the transcript level. Short-read sequencing is widely used to evaluate DEGs in response to experimental conditions. To prepare samples for short-read sequencing, total RNA is extracted from tissues collected from living organisms. Total RNA comprises 80–90% of ribosomal RNA (rRNAs), while messenger RNA (mRNAs) that include most protein-coding genes make up only 3% (O'Neil et al., 2013). It is necessary to enrich mRNA to increase sequencing efficiency. As RNA-Seq has a complicated workflow, it is easy to introduce biases unintentionally. There are two types of variation: within-sample variation and between-sample variation. GC content, gene length, and contamination are sources of within-sample variation that affect the detection of different genes in the same sample (Evans et al., 2018).

Sample preparation and analysis strategies have been developed to target this variation. For example, 3′ RNA-seq reduces the gene length bias by replacing the mRNA isolation and fragmentation steps with cDNA synthesis using oligo-dT primers (Moll et al., 2014).

Identifying biologically relevant between-sample variation is the goal of most RNA-Seq experiments. However, sample preparation techniques can also cause variation between samples. The total number of reads, which reflects sequencing depth, is a critical factor that affects the comparison of gene expression between samples (Evans et al., 2018). Therefore, it is essential to normalize for read depth, so several normalization approaches have been developed to account for biases in the datasets. The built-in normalization method in the two popular differential expression (DE) tools, DESeq2 and EdgeR, normalize gene expression using the distribution of read counts to account for sequencing depth and RNA abundance (Risso et al., 2014a; Robinson & Oshlack, 2010). Critically, these methods depend on the assumption that most genes do not change in expression between the samples. This fundamental assumption only holds true in some cases (Coate & Doyle, 2015). For example, transcriptional amplification in tumor cells globally increases the expression of existing genes rather than turning on the expression of new genes (Lin et al., 2012). Several studies show that experimental conditions alter overall transcription in the cells. In yeast, growth rates are highly correlated with total transcript abundance (Athanasiadou et al., 2016; Brauer et al., 2008; Yu et al., 2021), and nutrient limitation also causes transcriptional reprogramming (Lippman & Broach, 2009; Yu, Campbell, et al., 2020). These cases violate the assumption that the expression level of most genes does not change between samples analyzed by RNA-Seq. Therefore, using the distribution-based normalization in these cases would inappropriately shift the expression levels of genes to force similar total read counts and thus would result in incorrect identification of DEGs (Athanasiadou et al., 2016; Lovén et al., 2012). This means that the total RNA levels must be similar between samples for these distribution-based normalization methods to be appropriate. However, many experimental conditions can change the transcription level or RNA stability, affecting the total RNA abundance, thus invalidating the assumption of consistent total RNA levels.

A related challenge in using distribution-based normalization protocols for RNA-Seq data can happen if genes with a drastic increase or decrease in expression level alter the proportional shares of mRNA in the RNA pool. For example, photosynthesis-related genes, which are about 3000 genes in plants (Wang et al., 2017), are only up-regulated during the day to perform photosynthesis, and they decrease in expression at night. Assuming that the total RNA pool remains consistent (as most genes are not DEGs), these genes will become a majority in the RNA pool during the day, and normalizing to the same pool size will artificially reduce the abundance of other genes, even those that do not change in expression, leading to incorrect identification of DEGs. Some DEG-identifying algorithms, such as EdgeR, address this for some proportion of the genes (M. D. Robinson & Oshlack, 2010), but with plants, significant transcriptional changes are frequently induced where photosynthetic genes are a major part of the transcriptome and vary significantly even between different times of day.

Fortunately, there is a simple solution to control for changes in mRNA levels between samples and allow for accurate comparisons. Artificial RNA spike-ins, first developed to account for technical variation in microarray data, can be used as an external control to normalize the total RNA levels (Jiang et al., 2011). These external RNA spike-ins have been proposed since the first high-throughput transcriptional analysis approaches, microarrays, and qRT-PCR (Czechowski et al., 2005; Girke et al., 2000; Hilson et al., 2004). They are often used in RNA-seq studies in mammalian and yeast systems (Brauer et al., 2008; Byrne et al., 2017; Kroustallaki et al., 2019; Lun et al., 2017; Wang et al., 2021; Wilson et al., 2019). However, the use of external RNA spike-ins as a normalization factor is not common in plant RNA-Seq analysis studies.

Here, we demonstrate that spike-in normalization is important in identifying DEGs with RNA-Seq. With a synthetic RNA-Seq dataset, we demonstrate that using spike-ins in normalization resulted in better accuracy, specificity, and sensitivity in DEG calling. We demonstrated that adding external RNA spike-ins significantly affects DEG identification between sorghum samples that differ only by the time of day they are collected under normal and chilling stress conditions. We also demonstrate that external RNA spike-ins affect DEG identification between control and chilling stress. Traditional normalization tended to bias toward identifying morning up-regulated genes; however, spike-in normalization reveals novel evening up-regulated genes in sorghum under normal and chilling stress conditions. Our study highlights the importance of spike-in normalization in maintaining wanted variation caused by the experimental conditions and accurately identifying differentially expressed genes.

RESULTS AND DISCUSSION

External RNA spike-ins can correct for changes in global transcriptional abundance and can capture changes in the composition of differentially expressed genes

Previous studies have demonstrated that RNA spike-ins can accurately correct for changes in global transcription in plants using cDNA arrays, microarrays, and qRT-PCR (Czechowski et al., 2005; Girke et al., 2000; Hilson et al., 2004). Using a synthetic dataset (Figure 1), we demonstrate the effects of a global change in transcription

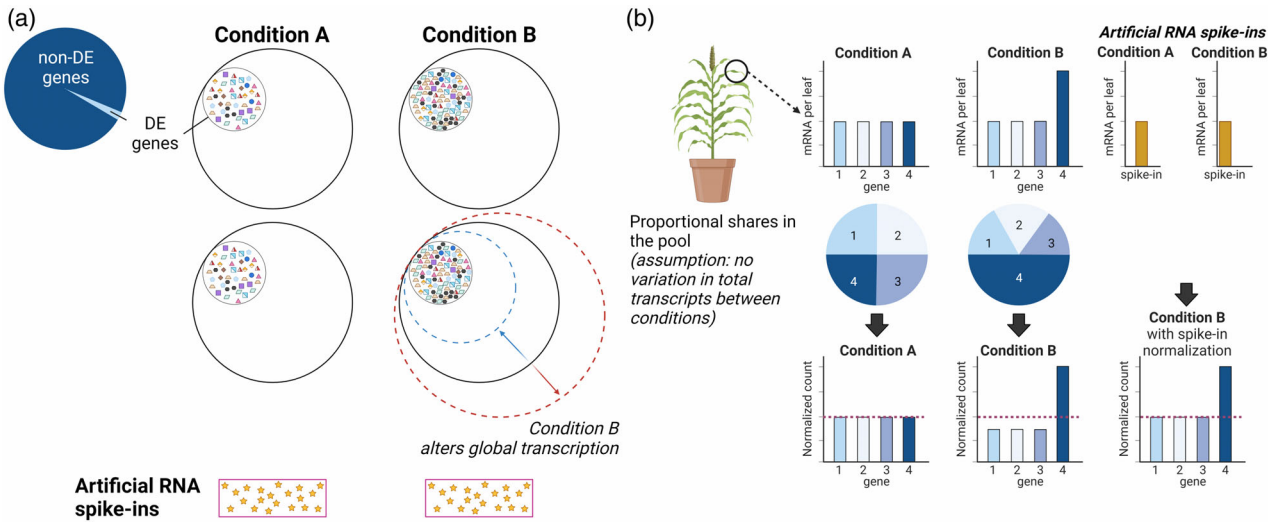

Figure 1. Challenges in RNA-Seq analysis.
(a) Commonly used normalization methods assume that only a small proportion of transcripts are differentially expressed between conditions (small, dashed inner circle in Conditions A and B). It is assumed that most transcripts do not change in expression across experimental conditions (as indicated by the solid outer circle in Conditions A and B), resulting in stable transcript pool. However, some experiments do affect global transcription, either increasing (dashed red circle) or decreasing (dashed blue circle) the size of the RNA pool.
(b) A change in the proportional share of mRNA in the pool could also affect the identification of DEGs. When some genes substantially increase in gene expression (e.g., Gene group 4), commonly used normalization methods that assume no change in global expression would result in artificially reducing the expression of other genes (e.g., groups 1–3) that do not change in expression. The figure was created with Biorender.com.

levels of all genes (Figure 1a) or a portion of genes (Figure 1b) on normalization and DEG calling.

To test the effects of a global change in expression on DEG identification, we created a synthetic RNA-Seq dataset from two experimental conditions, A and B (Figure 2). Each condition had four replicates with varying library sizes, as would be expected from standard library preparation techniques (Figure 2a). In condition B, global transcription is doubled, leading to a doubling of the total reads (Figure 2a). The traditional method employed in DESeq2, known as 'Median of Ratio', calculates the median of the read count ratio in one sample to a geometric mean across all samples (also referred to as a pseudo-reference) (Anders & Huber, 2010). To analyze the effects of normalization, we generated relative log expression (RLE) plots, representing the distribution of read counts after normalization (Gandolfo & Speed, 2018). Traditional normalization scaled transcript abundance between two conditions to the same level, as expected. This indicates that the increased global transcription in condition B is considered noise that needs to be removed (Figure 2a). We tested spike-in normalization using the method Athanasiadou et al. developed (Athanasiadou et al., 2019). In brief, this method utilized a maximum likelihood estimation to determine the calibration constant (?, nu) based on the spike-in read counts (Athanasiadou et al., 2019). Additionally, they computed the library-specific scale factor (?, delta) to account for potential errors introduced during library preparation, assuming that global expression levels between replicates should ideally remain identical (Athanasiadou et al., 2019). The RNA spike-ins were added proportionally to the total sample mRNA according to the experimental protocols for adding spike-ins (Jiang et al., 2011) (Table S1). The RLE plot showed increased transcript abundance in the samples in condition B after spike-in normalization (Figure 2a), suggesting that spike-in normalization can detect a change in global transcript abundance. Even when we control the library size to be consistent across samples, we still observe that only spike-in normalization captured a shift in transcript abundance due to altered global transcription (Figure S1a). While traditional normalization could not distinguish technical variations introduced by sample preparation from biological differences, the spike-in normalization removed only technical variation but preserved a difference in transcript abundance due to the global change in transcription.

To assess the impact of spike-in normalization on DEG identification, we created an RNA-Seq dataset comprising 10,000 genes with random numbers generated based on a negative binomial distribution (Table S2). Under conditions with a substantial increase in gene expression, the traditional normalization adjusted the read count distribution to be uniform across the samples (Figure 2b and Figure S1b–d). In contrast, the spike-in normalization method effectively maintained the unequal distribution between conditions A and B (Figure 2b; Figure S1b–d). For example, in scenarios where 25% and 50% of genes were up-regulated in condition B, more

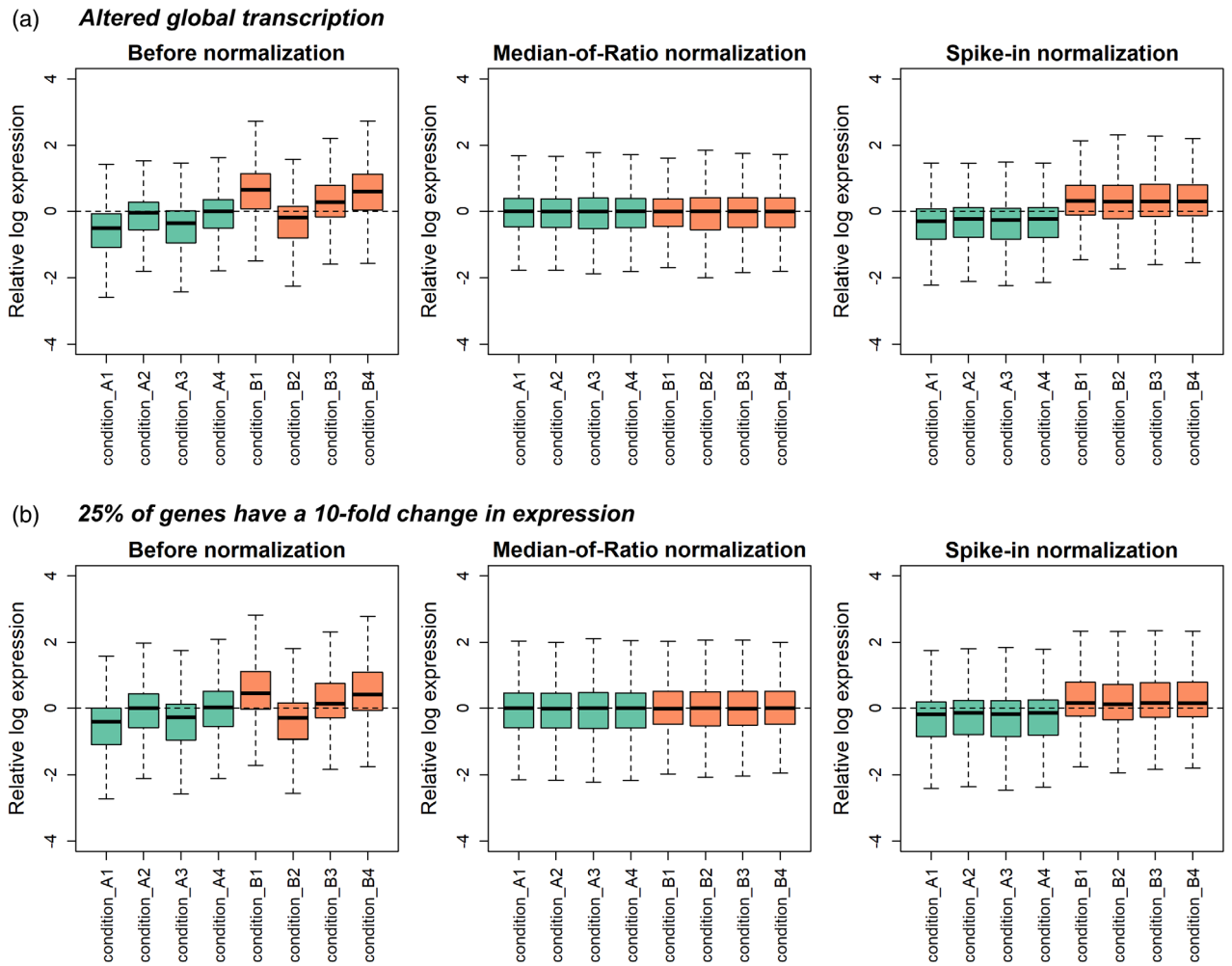

Figure 2. External RNA spike-ins captured differences in transcript abundance due to altered global transcription and altered proportional shares of transcripts. RLE plots of synthetic gene read counts before and after Median-of-Ratio and spike-in normalizations in DESeq2 under altered global transcription (a) and with a drastic change in gene expression in condition B (b) with varied library size across samples.

DEGs identified by the spike-in normalization method overlapped with the list of genes we manipulated to be DEGs than those identified by the traditional method (Table S3).

To further assess the performance of the normalization methods, we constructed a confusion matrix, which allowed us to compare the predicted DEGs (after normalization) with the expected DEGs (genes for which we manually adjusted the expression). The spike-in normalization method exhibited more true positives and true negatives, leading to higher accuracy, sensitivity, and specificity (Table S3), suggesting that external RNA spike-ins significantly enhance DEG identification when drastic changes in gene expression affect the proportional shares of mRNA in the pool. This demonstrates the effectiveness and robustness of the spike-in normalization method in improving the accuracy and reliability of DEG analysis in diverse experimental conditions.

Using external RNA spike-ins leads to identifying novel PM up-regulated genes in sorghum

Through evaluation of a synthetic RNA-Seq dataset, we observed that spike-in normalization significantly enhances the accuracy of DEG identification. However, to validate the performance of spike-in normalization in a real RNA-Seq dataset, we explored its effectiveness in understanding the impact of time of day and chilling stress on gene expression in sorghum leaves. We utilized 3′ RNA-Seq to analyze the transcriptomic changes in sorghum leaves under control and chilling stress conditions during both morning and evening time points. The utility of 3′ RNA-Seq in studying transcriptional responses has been demonstrated in various plant species, including maize, rice, *Brachypodium distachyon*, *Setaria viridis*, switchgrass, apples, tomatoes, and Alpine orchid (*Gymnadenia rhellicani*) (Eveland et al., 2008; Israeli et al., 2019; Kellenberger

et al., 2019; Kremling et al., 2018; Palmer et al., 2019; Silva et al., 2019; Yu, Hao, et al., 2020). One of the advantages of 3′ RNA-Seq is its ability to minimize gene length bias by capturing only one read per transcript (Ma et al., 2019). The success of 3′ RNA-Seq relies heavily on the quality of the reference genome (Ma et al., 2019; Tandonnet & Torres, 2017). Encouragingly, we found that 88% to 92% of the reads uniquely mapped to the sorghum BTx623 reference genome (Table S4). The reads predominantly aligned toward the 3′ end of the gene body, which is consistent with our expectations and the characteristics of 3′ RNA-Seq (Figure S2a), suggesting that the sorghum genome quality is sufficient for 3′ RNA-Seq applications.

Artificial RNA spikes (SIRV set 3, Lexogen, USA) were incorporated into the plant samples during the RNA extraction process. The SIRV set 3, comprising ERCC spike-in controls (Lemire et al., 2011) and Lexogen's Spike-In RNA variants (SIRVs) controls (Paul et al., 2016), is a well-established tool commonly used in mammalian and yeast studies (Blevins et al., 2019; Nadal-Ribelles et al., 2019; Topal et al., 2019) to account for variation in total RNA content between samples. While in mammalian cells, spike-in standards are added proportionally to the cell count (Lovén et al., 2012), determining the ideal method for adding RNA spike-ins to plant samples presents unique challenges. Unlike mammalian cells, accurately assessing the total cell number in plant tissue is more complicated. Nonetheless, several approaches have been considered, such as normalizing to the plant, leaf, tissue weight, or total DNA content. In chilling stress, our focus, changes in cell expansion, cell cycle progression, and potential endoreduplication issues (Ashraf & Rahman, 2019; Louarn et al., 2010; Pirrello et al., 2018; Rymen et al., 2007; Zhao et al., 2014) have been observed. Therefore, we opted to normalize on a per-leaf basis, adding 90 pg of spike-in controls to each sorghum leaf sample, where each sample originated from a single leaf. While chilling stress resulted in a noticeable reduction in leaf size, no size difference was observed between morning (AM) and evening (PM) samples collected on the same day. Therefore, to initially assess the efficacy of spike-ins, we focused on identifying DEGs between the AM and PM samples in either control or chilling conditions so that the leaf size would not be a factor (Figure 3). Notably, the number of spike-in reads correlated with total read counts (Figure S2b), and the proportion of spike-in reads to total reads remained consistent across samples, representing approximately 0.04–0.08% (Figure S2c). A one-way ANOVA was performed to evaluate treatments' impact on the spike-in proportion, resulting in a *P*-value (0.336) higher than 0.05 (Figure S2c), indicating no significant difference in the spike-in proportion between treatments, as expected.

We employed RLE plots to visualize the transcript distribution in our data (Figure 3). As expected, before normalization, the boxes on the plots exhibited noticeable deviations from the center (Figure 3a). The PCA plot illustrated that the biological replicates of each treatment were not well clustered (Figure 3a), indicating the need for normalization before proceeding with the differential gene expression analysis. The traditional normalization approaches completely removed the variation between biological and technical replicates, aligning all samples to the same level, as anticipated, based on their underlying assumption (Figure 3b). In contrast, reads normalized using the normalization factors derived from the spike-in read count reduced the variation within biological replicates while retaining the variation between treatments (Figure 3c). The PM samples generally displayed a higher average read count than the AM samples under control and chilling stress conditions (Figure 3c). Despite differences in leaf size, we noticed little difference in the average read count when comparing AM control to AM chilling or PM control to PM chilling (Figure 3c). Additionally, including spike-ins reduced the within-treatment variation between samples in the same condition compared to the traditional method (Figure 3c). This suggests that the use of spike-in normalization not only impacted between-treatment total read levels but also reduced variation within samples of the same condition.

The PCA plots for both traditional and spike-in-based normalization demonstrated that the first principal component distinctly separated samples based on the time of day, while the second component effectively distinguished samples according to the temperature condition (Figure 3b,c). These findings suggest that the time of day exerts a more significant influence on gene expression variation in leaves of this sorghum genotype than the differences between the control and chilling temperatures. Consequently, this highlights the critical importance of evaluating stress responses multiple times throughout the day to obtain a comprehensive and accurate representation of gene expression changes (Blair et al., 2019; Bonnot et al., 2023; Fowler et al., 2005; Fowler & Thomashow, 2002; Grinevich et al., 2019).

We conducted DEG analysis between dawn and dusk in both the control (control_AM vs control_PM) and chilling temperature (chilling_AM vs chilling_PM) conditions to assess the effect of time of day on gene expression in sorghum under two temperature conditions. Employing the RNA spike-ins for normalization identified a significantly higher number of up-regulated genes in the evening (Figure 3d,e). This increase in the identification of evening up-regulated genes by normalization methods that included spike-ins compared to traditional normalization methods was robust across several spike-in and traditional normalization methods (Figure S3a,b). To compare the effects of adding RNA spike-ins, we focused on comparing the DESeq normalization without using RNA spike-ins (Median of Ratios) (Anders & Huber, 2010) to the RNA spike-in adjusted

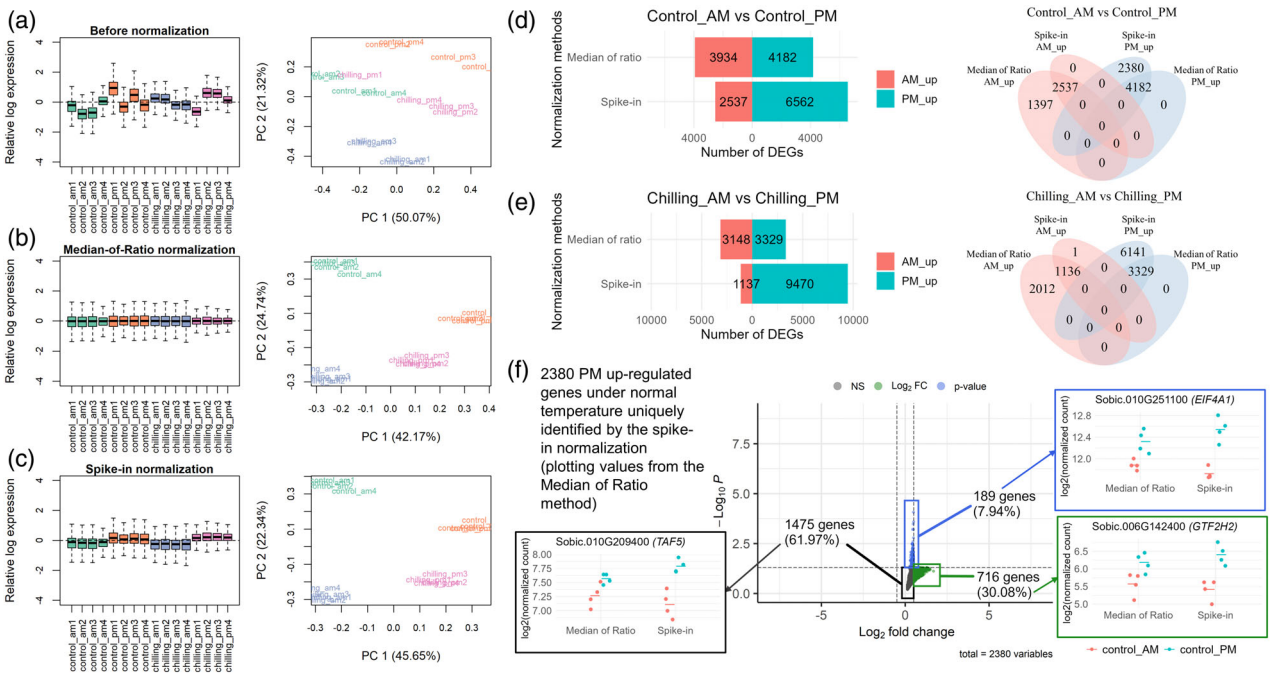

Figure 3. External RNA spike-in normalization preserves variation between experimental conditions.
(a–c) RLE and PCA plots of unnormalized read counts (a), read counts after the Median-of-Ratio normalization (b), and read counts after the spike-in normalization method in DESeq2 (c).
(d, e) Bar charts and Venn diagrams show the number of morning-upregulated (Salmon) and evening-upregulated (Teal) genes under normal (d) and chilling stress (e) from the Median of Ratio and spike-in normalization methods. Genes with FDR <0.05 and the absolute $\log_2$ fold-change >0.5 were identified as differentially expressed genes (DEGs).
(f) Volcano plot showing the 2380 genes identified as DEGs by spike-in method under the control condition. These genes are plotted with their values after the Median of Ratio normalization to visualize why they were not identified as DEGs by the Median of Ratio method. Dot plots represent an example gene for each category not identified as DEGs by the Median of Ratio due to failed $\log_2$ fold-change (blue), failed *P*-value (green), and failed $\log_2$ fold-change and *P*-value (black).

(Athanasiadou et al., 2019). The DEGs exclusively detected by the Median of Ratio method were all morning up-regulated (Figure 3d,e). Conversely, the spike-in normalization uniquely identified novel evening up-regulated genes in both temperature conditions (Figure 3d,e). This observation suggests that the traditional normalization approach, which adjusts the transcript abundance distribution to the same level, introduces a bias toward detecting morning up-regulated genes (Figure 3d,e). In contrast, normalizing the data with external RNA spike-ins successfully maintained the asymmetric transcript distribution in total reads between morning and evening samples (Figure 3d,e). This preservation of asymmetric distribution facilitated the detection of more evening up-regulated genes (Figure 3c), further emphasizing the advantages of utilizing RNA spike-ins for accurate normalization and more comprehensive gene expression analysis in our study.

Genes uniquely identified by the addition of RNA spike-ins are due to differences in both assigned log fold changes and significance

To investigate the reasons behind the differences in the DEGs identified by the Median of Ratio and RNA spike-in normalization methods, we visually examined the $\log_2$ fold-change and *P*-values of the AM vs. PM DEGs exclusive to each method (Figure 3f; Figure S3c–e). Under the control conditions, we found that 1397 unique DEGs from the Median of Ratio method were up-regulated in the morning, while all 2380 unique DEGs from the spike-in method were up-regulated in the evening (Figure 3d). Examining why the DEGs identified by the RNA spike-in normalization method were not identified in the Median of Ratio method showed that of these 2380 genes, a few were not identified by the Median of Ratio method because they did not make the fold-change cutoff (7.94%) or the *P*-value cutoff (30.08%). However, most of the DEGs uniquely identified (61.97%) failed to make either cutoff when using the Median of Ratios (Figure 3f). A similar distribution was observed for the DEGs uniquely identified by the Median of Ratio method. Of these 1397 DEGs, 10.59% passed the *P*-value threshold but did not meet the fold-change cutoff; 24.7% passed the fold-change cutoff but not the *P*-value; and the majority of the DEGs identified only by the Median of Ratio method, 64.71% did not meet either the *P*-value or fold-change cutoff in the RNA-spike in normalization (Figure S3c).

We plotted the expression of selected genes in each category to visualize why the normalization method would make a difference in log fold-change, *P*-value, or both. We observed that both the variation between the biological replicates and the expression level in the PM samples contributed to these differences (Figure 3f; Figure S3c). For example, both an increase in average expression in the evening and a reduction in within-condition variation contributed to the genes uniquely identified by the RNA spike-in normalization (e.g., *GENERAL TRANSCRIPTION FACTOR II H2* (*GTF2H2*), *RNA POLYMERASE RPB8* (*NRPB8*), and *TBP-ASSOCIATED FACTOR 5* (*TAF5*)) (Figure 3f). For some genes (e.g., *NRPB8* and *TAF5*), there also appears to be a reduction in the average expression of the AM timepoint after the RNA spike-in normalization. Examples of genes uniquely identified as DEGs and expressed higher in the AM than in the PM by the Median of Ratio method include *BASIC LEUCINE-ZIPPER 43* (*bZIP43*), *AUXIN RESPONSE FACTOR 16* (*ARF16*), and *PSEUDO-RESPONSE REGULATOR 7* (*PRR7*) (Figure S3c). For these three genes, the RNA spike-in normalization reduced the within-condition variation between samples and increased the expression of each sample in the PM, thus reducing the statistical significance of their difference (*bZIP43*), total change in expression levels (*ARF16)*, or both (*PRR7)*.

The improvement of transcript abundance with the spike-in normalization was also observed in the AM vs. PM DEGs under chilling stress conditions. Over 80% of non-DEGs did not pass both *P*-value and log fold-change cutoffs (Figure S3d,e). Furthermore, some genes in this group showed an opposite expression direction in the chilling stress conditions (Figure S3d,e). For example, Sobic.003G048600 (*GLUTAREDOXIN 4*, *GRX4*) and Sobic.003G257600 (*TAF11*) were significantly up-regulated in the evening with the spike-in normalization method, but their expression increased in the morning with the default normalization (Figure S3d). Sobic.001G537300 (*HOMEOBOX-LEUCIN ZIPPER PROTEIN REVOLUTA*, *REV*) and Sobic.007G186300 (*ANKYRIN REPEAT-CONTAINING PROTEIN 2, AKR2*) were significantly up-regulated in the morning with the default normalization (Figure S3e). However, they tended to be down-regulated in the morning with a spike-in normalization (Figure S3e). In addition, the biological replicates in the evening samples after the RNA spike-in normalization were more tightly grouped with each other than the replicates after the Median of Ratio method (Figure S3d,e). These findings underscore the importance of carefully selecting and applying appropriate normalization methods to accurately interpret and compare gene expression patterns, as the normalization method affects not only the statistical significance but can also alter the log fold-changes differences between treatments.

Functional analysis indicates that normalization with RNA spike-ins changes the enriched cellular functions identified

We demonstrated that the DEGs uniquely identified by the Median of Ratio method exhibited a bias toward morning up-regulated genes. In contrast, the RNA spike-in method revealed more unique evening up-regulated genes. To gain insights into the cellular functions of DEGs detected by both methods and those uniquely identified by each method, we conducted MapMan analysis to categorize the genes (Figure 4). In the control condition, both methods yielded a comparable number of unique DEGs falling into cellular pathways such as cell organization, development, hormones, regulation, and redox (Figure 4a). However, the RNA spike-in method notably showed more significant enrichment of the identified DEGS than the Median of Ratio method in the following pathways: DNA repair, cell cycle, protein targeting, biotic stress response, RNA synthesis, RNA processing, and protein synthesis (Figure 4a). In the context of DEGs between AM and PM under chilling stress conditions, the RNA spike-in method exhibited significant enrichment in almost all cellular pathways compared to the DEGs uniquely detected by the Median of Ratio method (Figure 4b). These results indicate a common function of the DEGs uniquely identified in the RNA spike-in method. Since the RNA spike-in identified genes are all higher expressed in the PM samples, these results support the idea that evening-specific functions and cellular processes may be under-represented or missed using traditional DEG analysis.

We focus on genes in RNA synthesis, a fundamental process in the central dogma of gene expression. In MapMan analysis, we found that the RNA spike-in method identified more unique DEGs relating to RNA and protein synthesis than the Median of Ratio method under control and chilling stress conditions (Figure 4). In the control condition, 534 RNA-related DEGs were identified by both normalization methods, and 128 and 227 genes were unique to the Median of Ratio and RNA spike-in methods, respectively. In the chilling stress condition, there were 402 common DEGs in both methods, and 215 and 530 genes were uniquely found in the Median of Ratio and RNA spike-in methods, respectively. RNA polymerases are essential for RNA production in cells, especially RNA polymerase II which is responsible for mRNA synthesis (Kwapisz et al., 2008; Vannini & Cramer, 2012). Both the Median of Ratio and RNA spike-in methods identified that several nuclear RNA polymerase (*NRP*) genes were up-regulated at night under the control condition, and spike-in normalization method identified additional evening up-regulated *NRP* genes that were not detected as DEGs with the traditional normalization (Figure 4; Table S5). Even more *NRP* genes were up-regulated in the evening under chilling stress with the spike-in normalization. Interestingly,

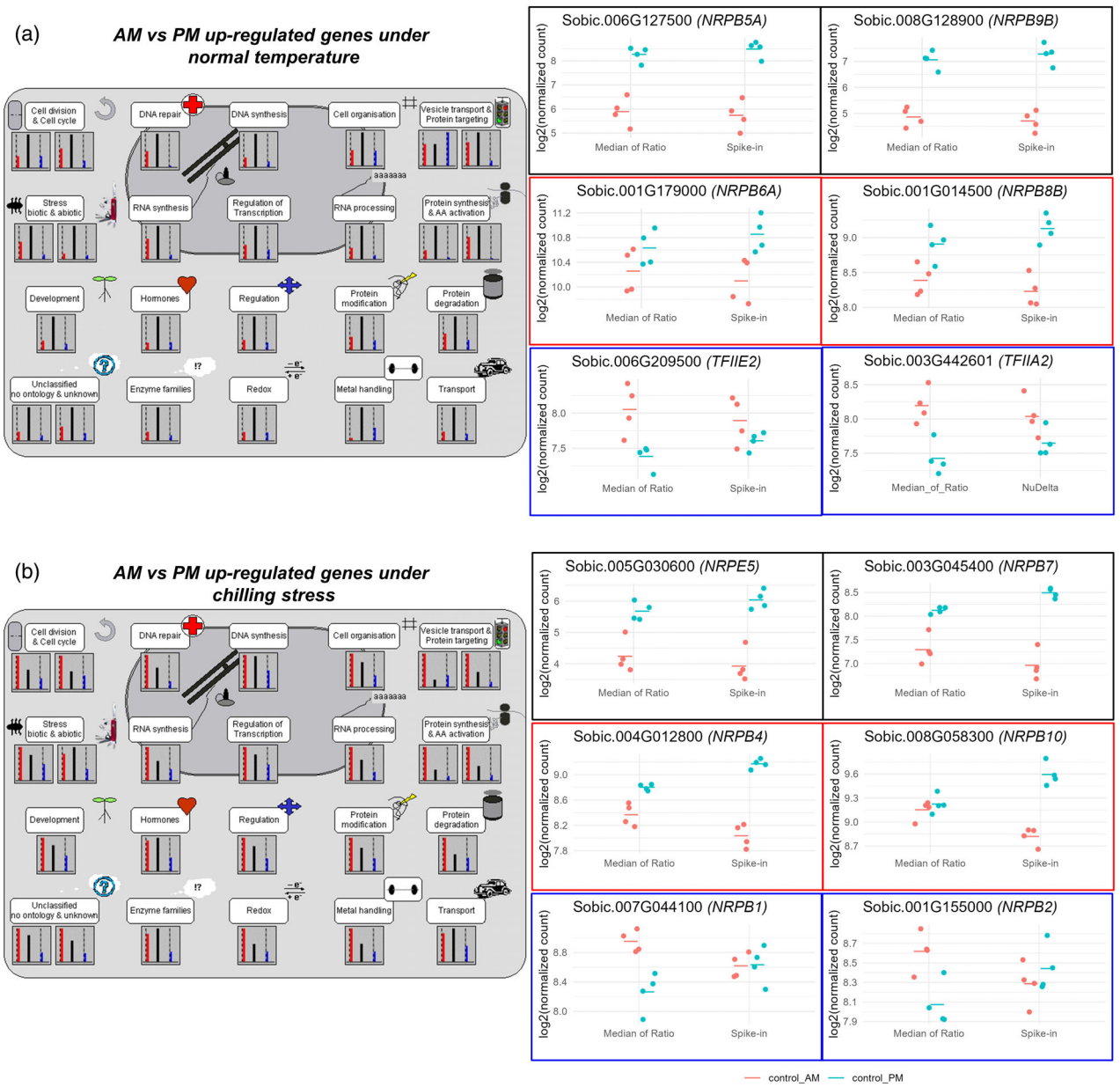

Figure 4. Spike-in normalization-specific DEGs were enriched in various biological pathways, including RNA synthesis.
MapMan classified DEGs between morning and evening in the control (a) and chilling stress conditions (b) identified by the Median of Ratio method only (blue, right bar), by spike-in only (red, left bar), and by both methods (black, center bar) into different cell functional groups. Dot plots represent examples of genes from the RNA synthesis group after different normalization methods.

Sobic.007G044100 (*NRPB1*) and Sobic.001G155000 (*NRPB2*), the largest catalytic subunits in the RNAP II that are commonly shared among RNAP II, IV, and V and function in the catalytic site (Ream et al., 2014) were identified as morning up-regulated genes under chilling stress by the traditional normalization. In contrast, these two genes were not differentially expressed by the spike-in normalization (Figure 4; Table S5). Although only measured at the transcriptional level, this increase in the abundance of RNAP II subunits might indicate that RNAP II is more active, or, if these are subunits shared with RNAP IV, that RNAP IV is more active, supporting why there is an observed increase in global transcripts when using the RNA spike-in method. Several publications indicate that the expression of genes encoding the RNAP subunits is affected by abiotic stress and some are required for abiotic stress tolerance (Borsani et al., 2005; Fernández-Parras et al., 2021; Popova et al., 2013). This provided a possibility that the sorghum produces more RNAP subunits at night to reprogram gene expression in response to temperature changes, and the

RNA spike-in normalization identified additional nighttime up-regulated RNAP coding genes.

Normalization with RNA spike-ins affects the evaluation of the response to chilling stress

Based on our evaluation of the differences between the normalization methods, we evaluated the response to chilling stress compared to the control conditions in sorghum at dawn and dusk using both normalization methods. DE analysis showed that the traditional normalization method detected a comparable number of chilling up- and down-regulated genes in the morning (3411 and 3739 genes, respectively) and evening (2309 and 2611 genes, respectively) (Figure 5a,c). In contrast, RNA spike-in normalization identified more chilling down-regulated genes (4571 down-regulated and 2877 up-regulated genes) in the morning but more chilling up-regulated genes in the evening (3478 up-regulated and 1607 down-regulated genes) (Figure 5a,c). Other spike-in utilized methods exhibited similar results where more chilling down-regulated genes were detected in the morning and chilling up-regulated genes (Figure S4a,b). The biggest factor causing genes to be classified in opposite expression is that the spike-in method shifted the mean of expression downward for the AM chilling samples and upward for the PM chilling samples (Figure 2c). Heatmaps of 1160 genes down-regulated by chilling in the morning that were identified as DEGs by the spike-in normalization clearly showed a difference in the chilling samples. The expression of these genes in chilling conditions after the RNA spike-in normalization was lower than the expression after the default normalization (Figure 5b). Likewise, the difference was distinct in the chilling samples for the 863AM chilling up-regulated DEGs uniquely identified by the Median of Ratio. These genes are identified as more highly expressed under chilling samples by the Median of Ratio method than by the RNA spike-in method. These results suggest that there is a bias in the Median of Ratio normalization for identifying up-regulated genes in the morning.

Furthermore, in the evening (PM) samples, the 867 genes up-regulated under chilling that were uniquely identified by the RNA spike-in method had a strong increase in expression under chilling compared to the expression normalized by the traditional, Median of Ratio method (Figure 5c). In contrast, the expression of the 702 PM DEGs in the evening identified by the Median of Ratio method as down-regulated under the chilling method were both higher under control conditions when normalized by the RNA spike-in method and lower under chilling conditions in the Median of Ratio method analysis (Figure 5d). Thus, in the evening samples, the Median of Ratio method identifies fewer chilling up-regulated genes and more down-regulated genes.

Gene Ontology (GO) enrichment analysis of chilling up- and down-regulated genes in the morning showed that both normalization methods shared several similar GO terms, for example, cell redox homeostasis, protein folding, response to endoplasmic reticulum stress showed up in chilling down-regulated genes, and metabolic process, response to cold, chloroplast organization, and carbohydrate metabolic process showed up in chilling up-regulated genes (Figure S5c,d). In the evening, chilling down-regulated genes were enriched in GO terms related to translation, while chilling up-regulated genes were enriched in photosynthesis and light response (Figure S5e, f). However, there are numerous GO terms that appeared in one normalization method but not in another (Figure S5c–f). For example, photorespiration is identified as enriched in genes down-regulated by chilling in the morning when normalized with RNA spike-ins but not the Median of Ratio method (Figure S5c). Our results indicated that the differences in DEGs captured by the RNA spike-in normalization propagate into unique GO terms in response to chilling stress and could affect our understanding of chilling stress response in plants.

The implication of RNA spike-in controls in RNA-Seq normalization

We have demonstrated that spike-in normalization significantly impacts differential expression analysis, which is a crucial aspect of RNA-Seq studies. Based on our synthetic data, spike-in normalization provides more accurate results, especially in cases where global transcription and proportional shares of the RNA pool are affected by experimental conditions. Several environmental changes have been shown to have effects on global transcription in plants (Branco-Price et al., 2005; Koiwa, 2006; Rymen et al., 2007; Szádeczky-Kardoss et al., 2022; Thatcher et al., 2018; Zhang et al., 2012, 2022). Although artificial spike-in controls have been used in microarray and some RNA-Seq data, their application in plant RNA-Seq data has been relatively limited. Some RNA-Seq protocols incorporate normalization factors such as spike-ins during library preparation to account for variation in library size. However, we observe that adding spike-ins only as a control for library variation has an overall reduction in the DEGs identified in each contrast (e.g., DESeq2 Median of Ratios vs DESeq2 RUV and EdgeR TMM vs EdgeR RUV in Figures S3 and S5) but no effect on the overall proportion of genes identified as DEG in either contrast. Normalization approaches that incorporate RNA spike-ins added during RNA extraction to account for changes in total transcript levels have an effect on the distribution of DEGs identified (Figures S3 and S5). Thus, considering the potential effects of experimental conditions on transcription in organisms, it is essential to consider using spike-ins in RNA-Seq experiments to ensure reliable and interpretable results.

In our sorghum dataset, spike-in normalization revealed that more transcripts were accumulated at night

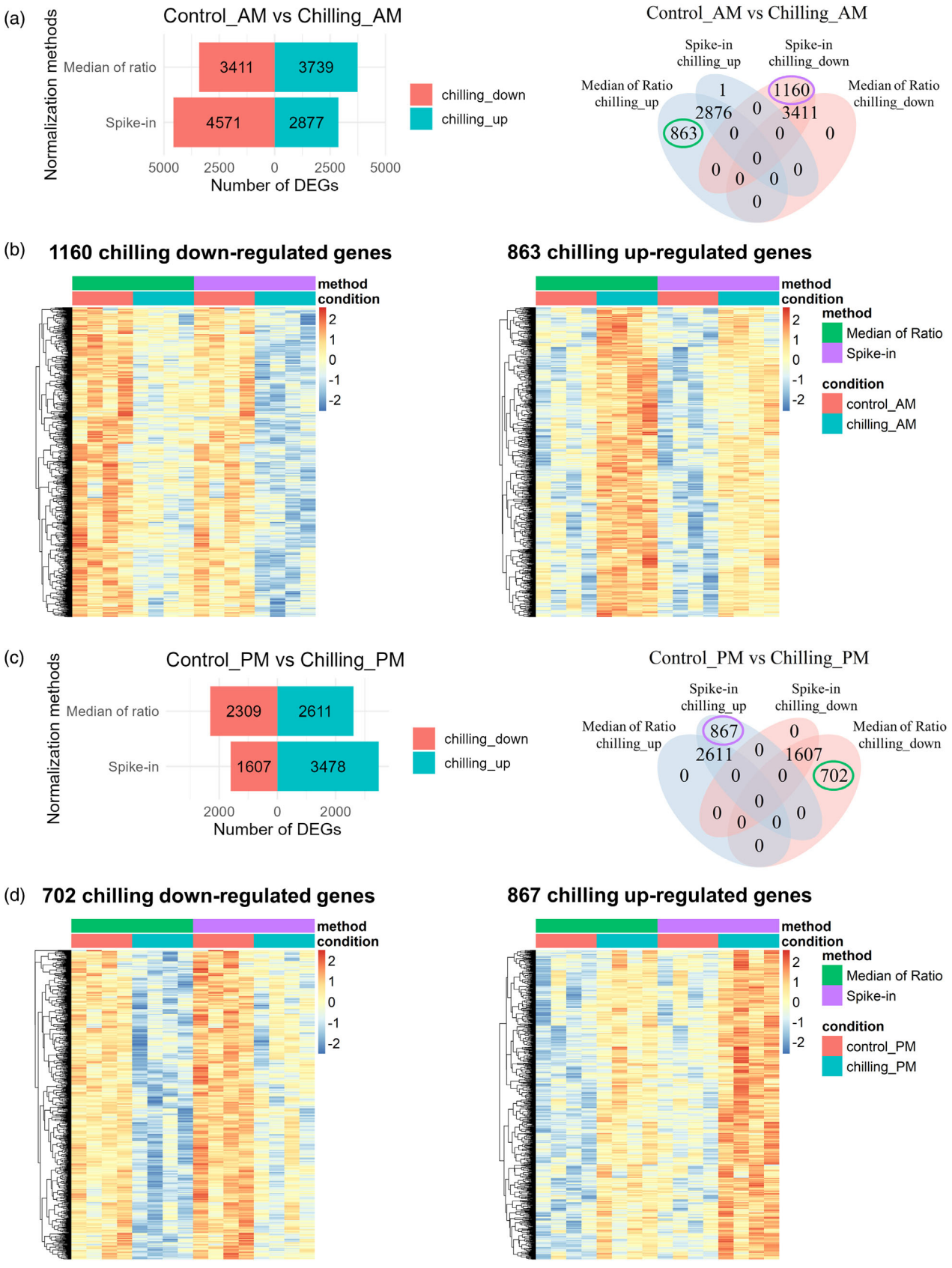

Figure 5. Spike-in normalization revealed contrasting responses to chilling stress at different times of the day. Spike-in normalization identified more genes as down-regulated chilling DEG genes in the morning but more up-regulated genes in response to chilling stress in the evening.
Bar charts and Venn diagrams (a, c) show the number of chilling down-regulated and up-regulated genes in the morning (a) and in the evening (c) from the Median of Ratio and spike-in normalization methods. Genes with FDR <0.05 and the absolute log fold-change >0.5 were identified as differentially expressed genes.
(b, d) Heat maps display normalized expression of chilling down-regulated and up-regulated genes in the morning (a) and in the evening (c) that were uniquely identified by the spike-in (left, green top bar) and Median of Ratio (right, purple top bar) methods.

than at dawn in this genotype, irrespective of temperature conditions, suggesting that the regulation of transcription machinery by the circadian clock possibly occurs in plants, similar to mammalian studies (Koike et al., 2012; Wang et al., 2018). For example, studies in mice using ChIP-Seq of RNA pol II in preinitiation and elongation states have indicated that RNA pol II activity depends on the time of day (Koike et al., 2012). The nascent transcription peaks at night (CT14.5). At the same time, the active form of RNAPII with phosphorylation of Serine 5 at the C terminus domain (CTD) highly accumulates on the genome in the early morning (CT0.6), suggesting that nascent transcription starts at night. Other histone modifications, such as promoter-enriched H3K27ac and the elongation marks, H3K36me3 and H3K79me2, also peak after the nascent transcription peak at night. Interestingly, the amount of nascent transcripts accumulates highly around CT15, but it is not highly correlated with the accumulation of mature mRNAs, with the mean circadian phase around CT7. This suggests that post-transcriptional modifications also play a role in driving the rhythmicity of mRNA levels.

Various publications have identified genes with rhythmic expression under diurnal conditions in different plant species (Michael et al., 2008). For instance, in Arabidopsis, approximately 89% of transcripts cycle when plants are grown under photocycles, thermocycles, or combined photocycles and thermocycles (Michael et al., 2008). Similarly, sorghum, maize, and foxtail millet show that 52%, 30%, and 43% of genes exhibit rhythmic expression under diurnal conditions (Lai et al., 2020). Furthermore, around 60% and 59% of transcripts show rhythmic expression in poplar (*Populus trichocarpa*) and rice (*Oryza sativa* ssp. japonica) (Filichkin et al., 2011). These findings strongly suggest a circadian regulation of transcription in plants. Additionally, post-transcriptional modifications, such as alternative splicing and polyadenylation, also occur in a time-of-day-dependent manner in plants.

We determined the correlation between the transcript abundance of each gene and total spike-in transcripts to propose candidate pseudo-reference genes if exogenous RNA spike-ins are not available in the dataset (Table S6). These candidates will be useful only for this sorghum cultivar and these growth conditions. However, if additional RNA-Seq experiments include RNA spike-ins, identifying robust markers may be possible for each genotype. Such data would enhance existing approaches, such as ISNorm, which seeks to identify internal genes that can be used, like exogenous RNA spike-ins, to account for global changes in transcription (Lin et al., 2020).

By employing RNA spike-in normalization, we preserved the asymmetric distribution of transcript abundance influenced by the time of day in the differential expression analysis. This analysis using RNA spike-ins resulted in identifying more differentially expressed genes, especially those up-regulated in the evening. It indicates that the current RNA-Seq analysis workflow may not fully capture the variation in total RNA levels when comparing morning and evening samples, potentially leading to an artificial reduction in evening gene expression to align it with the average gene expression in the morning. This, in turn, reduces the ability to identify genes expressed at higher levels at night compared to dawn. However, by utilizing artificial RNA spike-ins and normalization methods that leverage their benefits, we improve the identification of genes expressed during the nighttime. Analyzing both control and chilling stress conditions with the spike-ins reveals that several biological processes, even those previously unknown to have time-of-day-specific regulation, were expressed at higher levels at night. Critically, transcription and translation, which are fundamental components of the central dogma, show significant time-of-day specific regulation, emphasizing the potential need to control for global changes in gene expression.

Our study utilized external RNA spike-ins in these RNA-Seq experiments to examine the effects of the time of day and chilling stress in sorghum at the tissue level. Therefore, we added the spike-ins at the beginning of the RNA extraction and on a per-leaf basis to reflect differences in whole-leaf transcriptome and correct technical variation due to sample preparation. However, the spike-ins could be added based on weight, total RNA, cell number, or other basis, and the choice will significantly impact the genes identified as differentially expressed. For example, our comparison of leaves from two times of day likely reflects similar cell sizes and cells per leaf as these do not change significantly throughout the day. Studies in plants comparing the effects of aneuploidy or polyploidy on transcription have utilized external RNA spike-ins to control for the altered effects (Hou et al., 2018; Robinson et al., 2018). These studies investigated the relative changes in transcription compared to diploids. The researchers added spike-ins based on the total RNA to allow this evaluation of

relative expression modulation (Del Pozo & Ramirez-Parra, 2015; Robinson et al., 2018). However, neither our normalization per tissue nor normalizing per total RNA would fully capture changes in the total RNA per cell that might occur when comparing plants with different cell sizes due to ploidy differences, mutations, or environmental treatments. Endoreduplication can be another challenge in plants, and endoreduplication levels vary between different cell types, and tissues, and in response to different environmental conditions (Pirrello et al., 2018; Wos et al., 2022; Zumel et al., 2023). In these cases, adding spike-ins on a per-cell basis would be optimal, but quantifying cell number is challenging in intact plant tissues. One approach is to utilize RNA spike-ins in RNA-Seq based on the ratio of DNA to RNA to estimate gene expression per unit of DNA (Robinson et al., 2018). However, in comparisons where the DNA copy number changes, adding spike-ins per DNA/RNA ratio would not be the ideal approach (Hou et al., 2018). The genome-normalized expression method was developed to address this situation. This method estimates gene expression per genome and per cell by normalizing mRNA level by the copy number of its gene in the genome (Coate & Doyle, 2010; Hou et al., 2018; Shi et al., 2021; Yang et al., 2021) this method uses external RNA spike-ins in RNA-Seq (which estimate transcript-normalized expression) can estimate relative transcriptome size to determine the effect of ploidy on transcriptome and gene expression (Coate & Doyle, 2010; Hou et al., 2018; Robinson et al., 2018; Shi et al., 2021; Yang et al., 2021).

The transcription level per cell could be particularly important for the rapidly growing field of single-cell RNA-Seq. This technique has revealed that mRNA levels per cell are highly variable. The mRNA levels per cell are a critical factor because, in single-cell sequencing, the transcript level also affects the technical noise level (Brennecke et al., 2013; Grün et al., 2014). In mammalian single-cell sequencing studies, RNA spike-ins have demonstrated the presence of hypertranscription states, where cells undergo global transcriptional upregulation (Kim et al., 2023; Percharde et al., 2017). RNA spike-ins have also been successfully used to control for differences in transcriptome size, for example, in maturing oocytes (Wu, 2022) and pluripotent states (Shao et al., 2022). Further complicating single-cell RNA-Seq analysis, in mammalian cells, significant variation in the number of mRNA per cell is observed due to bursts of active transcription periods (Raj et al., 2006). Global, cell-wide constraints affect transcription bursts in *E. coli* and mammalian cells more than in yeast (Sanchez & Golding, 2013). However, in clonal yeast, single-cell sequencing with spike-in addition demonstrated that transcriptome size heterogeneity occurs even between clonal cells (Nadal-Ribelles et al., 2019). Although this technique can determine transcripts per cell, plant single-cell sequencing studies primarily use global normalization methods and have not fully taken advantage of RNA spike-ins. Using RNA spike-ins, Song et al. corrected for capture efficiency and observed a cell-type-dependent increase in the transcription level between diploid and tetraploid cells (Song et al., 2020). Given the variability observed between mRNA levels per transcript in other systems, it is likely that quantitative comparisons of expression between cells will require the use of RNA spike-ins. Incorporating RNA spike-ins in plant single-cell sequencing would correct for technical variation in sample preparation between cells and allow for accurate comparison across cell types and cells with varied total RNA (Lun et al., 2017).

Ultimately, the decision about how to normalize with spike-ins will depend on the experimental question and needs to be carefully considered. Although this requires some thought and effort, we propose that the value of adding spike-ins and the improvement it will provide to the transcriptomic analysis in plants make it well worth the effort.

MATERIALS AND METHODS

Plant growth conditions and sample collection

Sorghum bicolor cultivar BTx623 was grown in a controlled environment chamber (Conviron model PGR15; Winnipeg, MB, Canada) facility at the Department of Agronomy, Kansas State University. The chambers were maintained at a 12 h photoperiod, with 800 μmol m^{-2} sec^{-1} light intensity at 5 cm above the canopy and 60% relative humidity (RH). The daytime/nighttime temperatures for the controlled and chilling stress conditions were 30/20°C and 20/10°C, respectively, in the growth chambers with the 12/12 h of light/dark cycles. The chambers were programmed to reach the daytime (0800–1700 h) target temperatures of 30 and 20°C, by following a gradual increase from 20 to 30°C (control) and 10 to 20°C (chilling stress), with a 3 h transition (0500–0800 h). Similarly, the nighttime (2000–0500 h) target temperatures of 20 and 10°C were obtained by a gradual decrease in temperature from 30 to 20°C (control) and 20 to 10°C (chilling stress) respectively, with a 3 h transition (0500–0800 h). The second fully expanded leaf with a visible collar from the top on the 15-day-old seedlings was collected at 1 h after the light was on as the morning samples and at 1 h before the light was off as the evening samples. The fresh samples were flash-frozen in liquid nitrogen and stored at −80°C.

RNA extraction and library preparation

Total RNA was extracted from leaf tissue using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). In this study, 5.94 μl of 15.15 pg/μl Spike-in RNA variant controls (SIRVs) set 3 (Lexogen, Vienna, Austria) were added after tissue grinding to obtain 90 pg of SIRV per leaf. DNase I treatment was performed on RNeasy spin columns during a washing step. Total RNA was measured by Nanodrop Lite spectrophotometer (Thermo Scientific, Waltham, MA, USA). In this study, 300 ng of total RNA was used in the QuantSeq 3′mRNA-Seq Library Prep Kit for Illumina (FWD) (Lexogen, Vienna, Austria) according to a manufacturer's instruction. In brief, oligo (dT) primers were used to isolate mRNA and to start the first-strand cDNA synthesis. After that, RNA was removed, and the second-stranded cDNA synthesis was performed using

random primers containing Illumina-compatible linker sequences. cDNA libraries were purified with magnetic beads. Library amplification with 18 rounds of PCR was performed using i7 index primers. Libraries were measured and quality checked by an Agilent 2200 TapeStation. Libraries were made from 10 nM pooled samples and sequenced by the Illumina NovaSeq 6000 platform with 100-bp single reads at NC State University Genomic Sciences Laboratory (Raleigh, NC, USA).

RNA-Sequencing data processing

Raw reads were quality checked by FastQC (version 0.11.8) (Andrews, 2010). Adapter trimming was performed by BBDuk (in BBMap version 38.34) with the following parameters: *k = 13, ktrim = r, useshortkmers = t, mink = 5, qtrim = r, trimq = 10, minlength = 20* (Bushnell, 2014; 'BBMap Guide' 2016). FastQC was used again to check the reads after trimming. Reads were mapped to Sorghum bicolor cultivar BTx623 reference genome (Phytozome 12, genome version 3.1.0 and annotation version 3.1.1) (McCormick et al., 2018) and SIRV sequences (Lexogen, USA) using STAR (version 2.5.3) with default parameters (Dobin et al., 2013). SAMtools (version 1.9) was used to sort and index BAM files. Read count tables were generated using HTSeq-count in the HTSeq package (version 0.11.2) (Anders et al., 2014). Low read counts were filtered out using the *filterByExpr()* command in EdgeR, resulting in 17,651 genes for further analysis (Robinson et al., 2010). With the sequencing read depth we used, about 60% of spike-in transcripts were detected with 3′ RNA-Seq, with 58 out of 92 ERCC and 42 out of 68 SIRV transcripts being identified. The non-detected ERCC transcripts mostly had concentrations lower than 2 amoles/ul (Figure S5a), suggesting that our sequencing approach would likely miss transcripts with concentrations below this limit. This provides a valuable quantitative limit of detection not commonly available in traditional RNA-Seq studies.

Normalization prior to differential expression analysis

Normalization was performed before a differential expression analysis in DESeq2 and EdgeR (Figure S5b). Traditional normalization in DESeq2 was a Median of Ratio method (Anders & Huber, 2010). In DESeq2, we used the command *DESeq()* to perform normalization and differential expression analysis at the same time. EdgeR has a Trimmed Mean of M values (TMM) method as a default normalization (Robinson & Oshlack, 2010). The command *calcNormFactors(method = 'TMM')* was run to perform TMM normalization in EdgeR.

To normalize gene read counts based on spike-in reads in DESeq2, we used the method developed by Athanasiadou et al. (2019), and this method is referred to as 'spike-in normalization' throughout this manuscript. To distinguish it from other spike-in methods tested in Figures S2, S4, and S5, it is referred to as 'DESeq2 Spike-in (Nu*Delta)' in these figures. This method estimates calibration constants based on the abundance of spike-in controls and library-specific correction factors accounting for unwanted variations. The maximum likelihood calibration constant (v_j, Nu) was estimated from three factors: (1) the proportion of spike-in counts across all libraries contributed by the reference spike-in, the molecules per sample for the reference spike-in and the size of spike-in library j (Athanasiadou et al., 2019). Then the nominal abundance of transcripts in library j is the transcript counts divided by v_j (Athanasiadou et al., 2019). The library-specific scaling factor (δj, Delta) is based on the assumption that the transcript abundance should be identical across technical replicates in each treatment (Athanasiadou et al., 2019). δj is the scaling factor for library j in condition l is the exponential of β_j where β_j is the difference between the mean of log(nominal abundance of transcript i in library j) and the mean of log transformation of nominal abundance of transcript i among libraries in condition l (Athanasiadou et al., 2019). In DESeq2, v_j and δj were used as a size factor given by $v_j\delta_j$ divided by a geometric mean of $v\delta$ (Athanasiadou et al., 2019). We also tested other methods utilizing spike-in read counts in DESeq2 were from Brennecke et al. (2013). The Median of Ratio method was used to calculate size factors from the spike-in read count table by running *estimateSizeFactors()* and *sizeFactors()* commands (Brennecke et al., 2013). The size factor values were then added to the gene read count dataset prior to DE analysis This method is referred to as 'DESeq2 Spike-in size factor' in Figures S3–S5.

We utilized a spike-in-based normalization in EdgeR. We applied a $\log_2$ transformation to total spike-in reads to obtain normalization factors of each library and stored them as offsets to normalized gene counts during the DE analysis (Lun et al., 2017). This is identified as EdgeR $\log_2$ spike-in in Figures S3–S5. In contrast, we used the spike-ins to correct for library size only (ignoring changes in global gene expression). We did this using Removed Unwanted Variation (RUV) to calculate unwanted variation factors based on spike-in reads by running the command *RUVg(k = 1)* (Risso et al., 2014b). The factors from the RUV method were applied in both DESeq2 and EdgeR analysis as one factor in the design matrix along with the treatment factors (Risso et al., 2014b).

The distribution of transcripts before and after normalization was visualized as an RLE plot using the *plotRLE()* function in the EDASeq package (Risso et al., 2011). The PCA plot was created by the *plotPCA()* function in the EDASeq package (Risso et al., 2011).

Differential expression analysis

DESeq2 and EdgeR were conducted on normalized read counts (Love et al., 2014; Robinson et al., 2010) to obtain differentially expressed genes between times of day and between temperature conditions in R version 4.2.3. Genes identified as differentially expressed had an absolute $\log_2$ fold change higher than 0.5 and an FDR less than 0.05. Visualization, including volcano plots, dot plots, Venn diagrams, and heatmap plots, was created by EnhancedVolcano (version 1.4.0) (Blighe et al., 2022), ggplot2 (version 3.4.2), VennDiagram (version 1.7.3) (Chen & Boutros, 2011), and pheatmap (version 1.0.12) packages in R version 4.2.3.

Functional analysis with MapMan and GO enrichment analysis

MapMan analysis (version 3.5.1R2) was performed on the DEGs obtained from the Median of Ratio and spike-in, Nu*Delta, methods in DESeq2 (Thimm et al., 2004). The sorghum locus IDs that begin with 'Sobic' from the current annotation version (3.1.1) were converted to the locus IDs beginning with 'Sb' to be compatible with the sorghum database in MapMan. The ID conversion list was downloaded from the SorGSD database (Liu et al., 2021).

The R package 'topGO' (version 2.48.0) was used to determine enriched biological process GO terms (Alexa & Rahnenfuhrer, 2022). Sorghum annotated GO terms were from AgriGO v2.0 (Tian et al., 2017). The algorithm 'weight01' with Fisher statistics was used to calculate the significant levels.

Scripts used for all analyses

All scripts used for the analysis are available at https://github.com/kanjanal aosuntisuk/sorghum_spikein.

ACCESSION NUMBERS

Sequencing data can be found in the National Center for Biotechnology Information Sequence Read Archive (SRA) database (BioProject number PRJNA1033776).

ACKNOWLEDGEMENTS

We thank the reviewers for their suggestions to improve the manuscript and discussion.

AUTHOR CONTRIBUTIONS

ARL and CJD conceived the project and designed the experiments. AV, I.MS, and SVKJ grew plants and collected plant tissues. KL prepared samples for RNA sequencing, analyzed sequencing data, developed the normalization pipelines, generated the synthetic data sets and graphs, and performed the qRT-PCR. All authors discussed the results and reviewed the manuscript.

FUNDING INFORMATION

We would like to acknowledge funding for this project by Defense Advanced Research Projects Agency (DARPA) D19AP00026 and National Science Foundation (NSF) 2210293 to Colleen J. Doherty and funding from the Development and Promotion of Science and Technology Talents Project (DPST), Thailand, to K.L.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Table S1. Read count in the synthetic RNA-Seq dataset for demonstrating the effect of global transcription change.

Table S2. Read count in the synthetic RNA-Seq dataset for demonstrating the effect of proportional changes of the RNA pool with similar library size across samples.

Table S3. Number of DEGs and confusion matrix tables under the effect of proportional changes of the RNA pool.

Table S4. Number of reads after pre-processing (adapter trimming and rRNA removal), genome alignment, and gene assignment.

Table S5. Genes encoding the subunits of RNA polymerases were differentially expressed between AM and PM in control and chilling stress conditions according to the Median of Ratio and RNA spike-in normalization methods.

Table S6. Pearson's correlation values between sorghum transcript counts and total spike-in counts.

Figure S1. RNA spike-ins captured differences in transcript abundance due to altered global transcription and altered proportional shares of transcripts.

RLE plots of synthetic gene read counts before and after Median of Ratio and spike-in normalizations in DESeq2 under altered global transcription (a) and (b, c, and d) the drastic change in gene expression in condition B (b, c, and d) with consistent library size (a, b, and c) and varied library size across samples (d).

Figure S2. Details of sorghum 3′ RNA-Seq libraries and the relationship between gene and spike-in read counts.

(a) Read distribution on the gene body. The gene body coverage was calculated by the RSeQC package (version 2.6.6) (Wang et al., 2012). 3′ RNA-Seq contributed to reads mapped to the 3′ end of the transcripts.

(b) Correlation between total spike-in reads and total filtered gene reads. Pearson correlation coefficient and *P*-value are shown in the plot.

(c) The proportion of spike-in reads to total reads. The boxplot shows the median of spike-in proportion in four treatments. One-way ANOVA indicated that there was no significant difference (*P*-value >0.05) in the mean of the proportions between treatments.

Figure S3. RNA spike-in normalization approaches identified more evening-upregulated genes in both control and chilling stress conditions.

(a, b) The number of DEGs between morning and evening under control (a) and chilling temperatures (b) from different normalization approaches in DESeq2 and EdgeR (See Materials and Methods for details about each normalization). Traditional normalization methods in DESeq2 and EdgeR are the Median of Ratio and TMM, respectively. RUV-based approaches use spike-ins to normalize for library size but do not account for changes in global transcription. Genes with FDR <0.05 and the absolute $\log_2$ fold-change >0.5 were identified as differentially expressed genes. Bold *y*-axis labels highlighted methods utilizing external RNA spike-ins as normalization factors and italic labels indicated analyses in DESeq2.

(c) Volcano plot displayed the $\log_2$ fold-change and *P*-value of 1370 genes identified as DEGs up-regulated in the AM by the traditional, Median of Ratio method in control conditions. Their $-\log_{10}P$ and $\log_2$ fold-change values are plotted for their normalization with the spike-in normalization to visualize why they were not detected with the spike-in approach. Dot plots represent the example of genes that were not differentially expressed by the Median of Ratio method due to failed $\log_2$ fold-change (blue), failed *P*-value (green), and failed $\log_2$ fold-change and *P*-value (black).

(d) Volcano plot displayed the $\log_2$ fold-change and *P*-value of 6142 genes uniquely identified as evening up-regulated genes with the spike-in normalization under the chilling stress condition. Their $-\log_{10}P$ and $\log_2$ fold-change values after the Median of Ratio normalization are plotted to visualize why these were not identified by the Median of Ratio approach. Dot plots represent the example of genes that were not differentially expressed by the Median of Ratio method due to failed $\log_2$ fold-change and *P*-value.

(e) Volcano plot displayed the $\log_2$ fold-change and *P*-value of 2012 genes uniquely identified as morning up-regulated genes with the spike-in normalization under chilling stress. Their $-\log_{10}P$ and $\log_2$ fold-change values are plotted for the Median of Ratio normalization. Dot plots represent the example of genes that were not differentially expressed by the spike-in normalization due to failed $\log_2$ fold-change and *P*-value.

Figure S4. RNA spike-in normalization approaches identified more chilling down-regulated genes in the morning but more chilling up-regulated genes in the evening.

(a,b) The number of DEGs between control and chilling stress in the morning (a) and in the evening (b) from different normalization approaches in DESeq2 and EdgeR (See Materials and Methods for details about each normalization). Genes with FDR less than 0.05 and an absolute $\log_2$ fold-change higher than 0.5 were identified as DEGs. Bold *y*-axis labels highlighted methods utilizing external RNA spike-ins as normalization factors and italic labels indicated analyses in DESeq2.

(c–f) Heatmaps represent lists of significant GO terms (adjusted *P*-value <0.01) of AM chilling down-regulated genes (c), AM chilling upregulated genes (d), PM chilling down-regulated genes (e), and

PM chilling up-regulated genes (f) from the Median of Ratio and spike-in normalization methods. Numbers inside the boxes are the numbers of genes and colors indicate the $\log_2$ of adjusted *P*-values.

Figure S5. Details of external RNA spike-in controls and spike-in normalization methods used in the sorghum 3′ RNA-Seq dataset.

(a) Concentrations and average observed read counts of detected and undetected ERCC transcripts grouped by experimental conditions. A red vertical line indicated a 2 amol/μl of ERCC controls.

(b) Workflow of differential expression analyses with different normalization approaches (See Materials and Methods for details). The figure was created with Biorender.com.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1033776, https://github.com/kanjanaosuntisuk/sorghum_spikein.

REFERENCES

Alexa, A. & **Rahnenfuhrer, J.** (2022) *TopGO: enrichment analysis for gene ontology*. (version R package version 2.38.1).

Anders, S. & **Huber, W.** (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

Anders, S., **Pyl, P.T.** & **Huber, W.** (2014) HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**(2), 166–169.

Andrews, S. (2010) FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Ashraf, M.A. & **Rahman, A.** (2019) Cold stress response in *Arabidopsis thaliana* is mediated by GNOM ARF-GEF. *The Plant Journal: For Cell and Molecular Biology*, **97**(3), 500–516.

Athanasiadou, R., **Neymotin, B.**, **Brandt, N.**, **Miller, D.**, **Tranchina, D.** & **Gresham, D.** (2016) Growth rate-dependent global amplification of gene expression. *BioRxiv*, 044735. Available from: https://doi.org/10.1101/044735

Athanasiadou, R., **Neymotin, B.**, **Brandt, N.**, **Wang, W.**, **Christiaen, L.**, **Gresham, D.** ***et al.*** (2019) A complete statistical model for calibration of RNA-Seq counts using external spike-ins and maximum likelihood theory. *PLoS Computational Biology*, **15**(3), e1006794.

BBMap Guide. (2016) September 1, 2016. https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/

Blair, E.J., **Bonnot, T.**, **Hummel, M.**, **Hay, E.**, **Marzolino, J.M.**, **Quijada, I.A.** ***et al.*** (2019) Contribution of time of day and the circadian clock to the heat stress responsive transcriptome in *Arabidopsis*. *Scientific Reports*, **9**(1), 4814.

Blevins, W.R., **Carey, L.B.** & **Mar Albà, M.** (2019) Transcriptomics data of 11 species of yeast identically grown in rich media and oxidative stress conditions. *BMC Research Notes*, **12**(1), 250.

Blighe, K., **Rana, S.**, **Turkes, E.**, **Ostendorf, B.**, **Grioni, A.** & **Lewis, M.** (2022) *EnhancedVolcano: publication-ready volcano plots with enhanced Colouring and labeling* (version R package version 1.14.0). https://github.com/kevinblighe/EnhancedVolcano

Bonnot, T., **Impa, S.**, **Krishna Jagadish, S.V.** & **Nagel, D.H.** (2023) Time of day and genotype sensitivity adjust molecular responses to temperature stress in sorghum. *The Plant Journal: For Cell and Molecular Biology*, **116**, 1081–1096. Available from: https://doi.org/10.1111/tpj.16467

Borsani, O., **Zhu, J.**, **Verslues, P.E.**, **Sunkar, R.** & **Zhu, J.-K.** (2005) Endogenous SiRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in *Arabidopsis*. *Cell*, **123**(7), 1279–1291.

Branco-Price, C., **Kawaguchi, R.**, **Ferreira, R.B.** & **Bailey-Serres, J.** (2005) Genome-wide analysis of transcript abundance and translation in *Arabidopsis* seedlings subjected to oxygen deprivation. *Annals of Botany*, **96**(4), 647–660.

Brauer, M.J., **Huttenhower, C.**, **Airoldi, E.M.**, **Rosenstein, R.**, **Matese, J.C.**, **Gresham, D.** ***et al.*** (2008) Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast. *Molecular Biology of the Cell*, **19**(1), 352–367.

Brennecke, P., **Anders, S.**, **Kim, J.K.**, **Kołodziejczyk, A.A.**, **Zhang, X.**, **Proserpio, V.** ***et al.*** (2013) Accounting for technical noise in single-cell RNA-Seq experiments. *Nature Methods*, **10**(11), 1093–1095.

Bushnell, B. (2014) *BBMap: A Fast, Accurate, Splice-Aware Aligner.*" LBNL-7065E. Berkeley, CA (United States): Lawrence Berkeley National Lab. Available from: https://www.osti.gov/biblio/1241166

Byrne, A., **Beaudin, A.E.**, **Olsen, H.E.**, **Jain, M.**, **Cole, C.**, **Palmer, T.** ***et al.*** (2017) Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. *Nature Communications*, **8**(July), 16027.

Chen, H. & **Boutros, P.C.** (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics*, **12**(January), 35.

Coate, J.E. & **Doyle, J.J.** (2010) Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. *Genome Biology and Evolution*, **2**(July), 534–546.

Coate, J.E. & **Doyle, J.J.** (2015) Variation in transcriptome size: are we getting the message? *Chromosoma*, **124**(1), 27–43.

Czechowski, T., **Stitt, M.**, **Altmann, T.**, **Udvardi, M.K.** & **Scheible, W.-R.** (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiology*, **139**(1), 5–17.

Del Pozo, J.C. & **Ramirez-Parra, E.** (2015) Whole genome duplications in plants: an overview from *Arabidopsis*. *Journal of Experimental Botany*, **66**(22), 6991–7003.

Dobin, A., **Davis, C.A.**, **Schlesinger, F.**, **Drenkow, J.**, **Zaleski, C.**, **Jha, S.** ***et al.*** (2013) STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics*, **29**(1), 15–21.

Evans, C., **Hardin, J.** & **Stoebel, D.M.** (2018) Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, **19**(5), 776–792.

Eveland, A.L., **McCarty, D.R.** & **Koch, K.E.** (2008) Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. *Plant Physiology*, **146**, 32–44. Available from: https://doi.org/10.1104/pp.107.108597

Fernández-Parras, I., **Ramírez-Tejero, J.A.**, **Luque, F.** & **Navarro, F.** (2021) Several isoforms for each subunit shared by RNA polymerases are differentially expressed in the cultivated olive tree (Olea Europaea L.). *Frontiers in Molecular Biosciences*, **8**(December), 679292.

Filichkin, S.A., **Breton, G.**, **Priest, H.D.**, **Dharmawardhana, P.**, **Jaiswal, P.**, **Fox, S.E.** ***et al.*** (2011) Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. *PLoS One*, **6**(6), e16907.

Fowler, S.G., **Cook, D.** & **Thomashow, M.F.** (2005) Low temperature induction of *Arabidopsis* CBF1, 2, and 3 is gated by the circadian clock. *Plant Physiology*, **137**(3), 961–968.

Fowler, S. & **Thomashow, M.F.** (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. *The Plant Cell*, **14**(8), 1675–1690.

Gandolfo, L.C. & **Speed, T.P.** (2018) RLE plots: visualizing unwanted variation in high dimensional data. *PLoS One*, **13**(2), e0191629.

Girke, T., **Todd, J.**, **Ruuska, S.**, **White, J.**, **Benning, C.** & **Ohlrogge, J.** (2000) Microarray analysis of developing *Arabidopsis* seeds. *Plant Physiology*, **124**(4), 1570–1581.

Grinevich, D.O., **Desai, J.S.**, **Stroup, K.P.**, **Duan, J.**, **Slabaugh, E.** & **Doherty, C.J.** (2019) Novel transcriptional responses to heat revealed by turning up the heat at night. *Plant Molecular Biology*, **101**(1–2), 1–19.

Grün, D., **Kester, L.** & **van Oudenaarden, A.** (2014) Validation of noise models for single-cell Transcriptomics. *Nature Methods*, **11**(6), 637–640.

Hilson, P., **Allemeersch, J.**, **Altmann, T.**, **Aubourg, S.**, **Avon, A.**, **Beynon, J.** ***et al.*** (2004) Versatile gene-specific sequence tags for *Arabidopsis* functional genomics: transcript profiling and reverse genetics applications. *Genome Research*, **14**(10B), 2176–2189.

Hou, J., **Shi, X.**, **Chen Chen, M.**, **Islam, S.**, **Johnson, A.F.**, **Kanno, T.** ***et al.*** (2018) Global impacts of chromosomal imbalance on gene expression in

Arabidopsis and other taxa. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(48), E11321–E11330.

Israeli, A., Capua, Y., Shwartz, I., Tal, L., Meir, Z., Levy, M. *et al.* (2019) Multiple Auxin-response regulators enable stability and variability in leaf development. *Current Biology*, **29**(11), 1746–1759.e5.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M. *et al.* (2011) Synthetic spike-in standards for RNA-Seq experiments. *Genome Research*, **21**(9), 1543–1551.

Kellenberger, R.T., Byers, K.J.R., De Brito, R.M., Francisco, Y.M., Staedler, A.M., LaFountain, J.S. *et al.* (2019) Emergence of a floral colour polymorphism by pollinator-mediated overdominance. *Nature Communications*, **10**, 63. Available from: https://doi.org/10.1038/s41467-018-07936-x

Kim, Y.-K., Cho, B., Cook, D.P., Trcka, D., Wrana, J.L. & Ramalho-Santos, M. (2023) Absolute scaling of single-cell transcriptomes identifies pervasive Hypertranscription in adult stem and progenitor cells. *Cell Reports*, **42**(1), 111978.

Koike, N., Yoo, S.-H., Huang, H.-C., Kumar, V., Lee, C., Kim, T.-K. *et al.* (2012) Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, **338**(6105), 349–354.

Koiwa, H. (2006) Phosphorylation of RNA polymerase II C-terminal domain and plant osmotic-stress responses. In: Rai, A.K. & Takabe, T. (Eds.) *Abiotic Stress Tolerance in Plants*. Dordrecht: Springer Netherlands, pp. 47–57.

Kremling, K.A.G., Chen, S.-Y., Mei-Hsiu, S., Lepak, N.K., Cinta Romay, M., Swarts, K.L. *et al.* (2018) Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, **555**(7697), 520–523.

Kroustallaki, P., Lirussi, L., Carracedo, S., Panpan You, Q., Esbensen, Y., Götz, A. *et al.* (2019) SMUG1 promotes telomere maintenance through telomerase RNA processing. *Cell Reports*, **28**(7), 1690–1702.e10.

Kwapisz, M., Beckouët, F. & Thuriaux, P. (2008) Early evolution of eukaryotic DNA-dependent RNA polymerases. *Trends in Genetics*, **24**(5), 211–215.

Lai, X., Bendix, C., Yan, L., Zhang, Y., Schnable, J.C. & Harmon, F.G. (2020) Interspecific analysis of diurnal gene regulation in Panicoid grasses identifies known and novel regulatory motifs. *BMC Genomics*, **21**(1), 428.

Lemire, A., Lea, K., Batten, D., Jian Gu, S., Whitley, P., Bramlett, K. *et al.* (2011) Development of ERCC RNA spike-in control mixes. *Journal of Biomolecular Techniques*, **22**(Suppl), S46.

Lin, C.Y., Lovén, J., Rahl, P.B., Paranal, R.M., Burge, C.B., Bradner, J.E. *et al.* (2012) Transcriptional amplification in tumor cells with elevated C-Myc. *Cell*, **151**(1), 56–67.

Lin, L., Song, M., Jiang, Y., Zhao, X., Wang, H. & Zhang, L. (2020) Normalizing single-cell RNA sequencing data with internal spike-in-like genes. *NAR Genomics and Bioinformatics*, **2**(3), lqaa059.

Lippman, S.I. & Broach, J.R. (2009) Protein kinase a and TORC1 activate genes for ribosomal biogenesis by inactivating repressors encoded by Dot6 and its homolog Tod6. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(47), 19928–19933.

Liu, Y., Wang, Z., Xiaoyuan, W., Zhu, J., Luo, H., Tian, D. *et al.* (2021) SorGSD: updating and expanding the sorghum genome science database with new contents and tools. *Biotechnology for Biofuels*, **14**(1), 165.

Louarn, G., Andrieu, B. & Giauffret, C. (2010) A size-mediated effect can compensate for transient chilling stress affecting maize (Zea Mays) leaf extension. *The New Phytologist*, **187**(1), 106–118.

Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, **15**(12), 550.

Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B. *et al.* (2012) Revisiting global gene expression analysis. *Cell*, **151**(3), 476–482.

Lun, A.T.L., Calero-Nieto, F.J., Haim-Vilmovsky, L., Göttgens, B. & Marioni, J.C. (2017) Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Research*, **27**(11), 1795–1806.

Ma, F., Fuqua, B.K., Hasin, Y., Yukhtman, C., Vulpe, C.D., Lusis, A.J. *et al.* (2019) A comparison between whole transcript and 3′ RNA sequencing methods using Kapa and Lexogen library preparation methods. *BMC Genomics*, **20**, 9. Available from: https://doi.org/10.1186/s12864-018-5393-3

McCormick, R.F., Truong, S.K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D. *et al.* (2018) The sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal: For Cell and Molecular Biology*, **93**(2), 338–354.

Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D. *et al.* (2008) Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules. *PLoS Genetics*, **4**(2), e14.

Moll, P., Ante, M., Seitz, A. & Reda, T. (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. *Nature Methods*, **11**(12), i–iii.

Nadal-Ribelles, M., Saiful Islam, W., Wei, P.L., Nguyen, M., de Nadal, E., Posas, F. *et al.* (2019) Sensitive high-throughput single-cell RNA-Seq reveals within-clonal transcript correlations in yeast populations. *Nature Microbiology*, **4**, 683–692. Available from: https://doi.org/10.1038/s41564-018-0346-9

O'Neil, D., Glowatz, H. & Schlumpberger, M. (2013) Ribosomal RNA depletion for efficient use of RNA-Seq capacity. *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel . . . [et Al.]* Chapter 4 (July): Unit 4.19.

Palmer, N.A., Basu, S., Heng-Moss, T., Bradshaw, J.D., Sarath, G. & Louis, J. (2019) Fall armyworm (Spodoptera Frugiperda smith) feeding elicits differential defense responses in upland and lowland Switchgrass. *PLoS One*, **14**(6), e0218352.

Paul, L., Kubala, P., Horner, G., Ante, M., Holländer, I., Alexander, S. *et al.* (2016) SIRVs: spike-in RNA variants as external isoform controls in RNA-sequencing. *BioRxiv*, 080747. Available from: https://doi.org/10.1101/080747

Percharde, M., Bulut-Karslioglu, A. & Ramalho-Santos, M. (2017) Hypertranscription in development, stem cells, and regeneration. *Developmental Cell*, **40**(1), 9–21.

Pirrello, J., Deluche, C., Frangne, N., Gévaudant, F., Maza, E., Djari, A. *et al.* (2018) Transcriptome profiling of sorted Endoreduplicated nuclei from tomato fruits: how the global shift in expression ascribed to DNA ploidy influences RNA-Seq data normalization and interpretation. *The Plant Journal: For Cell and Molecular Biology*, **93**(2), 387–398.

Popova, O.V., Dinh, H.Q., Aufsatz, W. & Jonak, C. (2013) The RdDM pathway is required for basal heat tolerance in *Arabidopsis*. *Molecular Plant*, **6**(2), 396–410.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y. & Tyagi, S. (2006) Stochastic MRNA synthesis in mammalian cells. *PLoS Biology*, **4**(10), e309.

Ream, T.S., Haag, J.R. & Pikaard, C.S. (2014) Plant multisubunit RNA polymerases IV and V. In: Murakami, K. & Trakselis, M. (Eds.) *Nucleic acid polymerases. Nucleic acids and molecular biology*, Vol **30**. Berlin, Heidelberg: Springer. Available from: https://doi.org/10.1007/978-3-642-39796-7_13

Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. (2014a) The Role of spike-in standards in the normalization of RNA-Seq. In: Datta, S. & Nettleton, D. (Eds.) *Statistical analysis of next generation sequencing data*. Cham: Springer International Publishing, pp. 169–190.

Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. (2014b) Normalization of RNA-Seq data using factor analysis of control genes or samples. *Nature Biotechnology*, **32**(9), 896–902.

Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. (2011) GC-content normalization for RNA-Seq data. *BMC Bioinformatics*, **12**(December), 480.

Robinson, D.O., Coate, J.E., Singh, A., Hong, L., Bush, M., Doyle, J.J. *et al.* (2018) Ploidy and size at multiple scales in the *Arabidopsis* sepal. *The Plant Cell*, **30**(10), 2308–2329.

Robinson, M.D., McCarthy, D.J. & Smyth, G.K. (2010) EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**(1), 139–140.

Robinson, M.D. & Oshlack, A. (2010) A scaling normalization method for differential expression analysis of RNA-Seq data. *Genome Biology*, **11**(3), R25.

Rymen, B., Fiorani, F., Kartal, F., Vandepoele, K., Inzé, D. & Beemster, G.T.S. (2007) Cold nights impair leaf growth and cell cycle progression in maize through transcriptional changes of cell cycle genes. *Plant Physiology*, **143**(3), 1429–1438.

Sanchez, A. & Golding, I. (2013) Genetic determinants and cellular constraints in Noisy gene expression. *Science*, **342**(6163), 1188–1193.

Shao, R., Kumar, B., Lidschreiber, K., Lidschreiber, M., Cramer, P. & Elsässer, S.J. (2022) Distinct transcription kinetics of pluripotent cell states. *Molecular Systems Biology*, **18**(1), e10407.

Shi, X., Yang, H., Chen, C., Hou, J., Hanson, K.M., Albert, P.S. *et al.* (2021) Genomic imbalance determines positive and negative modulation of gene expression in diploid maize. *The Plant Cell*, **33**(4), 917–939.

Silva, K.J., **Pereira, J.S.**, **Bednarek, R.**, **Fei, Z.** & **Khan, A.** (2019) Differential gene regulatory pathways and Co-expression networks associated with fire blight infection in apple (Malus × Domestica). *Horticulture Research*, **6**(April), 35.

Song, Q., **Ando, A.**, **Jiang, N.**, **Ikeda, Y.** & **Jeffrey Chen, Z.** (2020) Single-cell RNA-Seq analysis reveals ploidy-dependent and cell-specific transcriptome changes in *Arabidopsis* female gametophytes. *Genome Biology*, **21** (1), 178.

Stark, R., **Grzelak, M.** & **Hadfield, J.** (2019) RNA sequencing: the teenage years. *Nature Reviews. Genetics*, **20**(11), 631–656.

Szádeczky-Kardoss, I., **Szaker, H.M.**, **Verma, R.**, **Darkó, É.**, **Pettkó-Szandtner, A.**, **Silhavy, D.** ***et al.*** (2022) Elongation factor TFIIS is essential for heat stress adaptation in plants. *Nucleic Acids Research*, **50**(4), 1927–1950.

Tandonnet, S. & **Torres, T.T.** (2017) Traditional versus 3′ RNA-Seq in a non-model species. *Genomics Data*, **11**, 9–16. Available from: https://doi.org/10.1016/j.gdata.2016.11.002

Thatcher, L.F., **Foley, R.**, **Casarotto, H.J.**, **Gao, L.-L.**, **Kamphuis, L.G.**, **Melser, S.** ***et al.*** (2018) The Arabidopsis RNA polymerase II carboxyl terminal domain (CTD) phosphatase-like1 (CPL1) is a biotic stress susceptibility gene. *Scientific Reports*, **8**(1), 13454.

Thimm, O., **Bläsing, O.**, **Gibon, Y.**, **Nagel, A.**, **Meyer, S.**, **Krüger, P.** ***et al.*** (2004) Mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The Plant Journal*, **37**, 914–939. Available from: https://doi.org/10.1111/j.1365-313x.2004.02016.x

Tian, T., **Liu, Y.**, **Yan, H.**, **You, Q.**, **Yi, X.**, **Zhou, D.** ***et al.*** (2017) AgriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Research*, **45**(W1), W122–W129.

Topal, S., **Vasseur, P.**, **Radman-Livaja, M.** & **Peterson, C.L.** (2019) Distinct transcriptional roles for histone H3-K56 acetylation during the cell cycle in yeast. *Nature Communications*, **10**(1), 4372.

Vannini, A. & **Cramer, P.** (2012) Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Molecular Cell*, **45**(4), 439–446.

Wang, L., **Wang, S.** & **Li, W.** (2012) RSeQC: quality control of RNA-Seq experiments. *Bioinformatics*, **28**(16), 2184–2185.

Wang, P., **Hendron, R.-W.** & **Kelly, S.** (2017) Transcriptional control of photosynthetic capacity: conservation and divergence from *Arabidopsis* to rice. *The New Phytologist*, **216**(1), 32–45.

Wang, X., **Frederick, J.**, **Wang, H.**, **Hui, S.**, **Backman, V.** & **Ji, Z.** (2021) Spike-in normalization for single-cell RNA-Seq reveals dynamic global transcriptional activity mediating anticancer drug response. *NAR Genomics and Bioinformatics*, **3**(2), lqab054.

Wang, Y., **Song, L.**, **Liu, M.**, **Ge, R.**, **Zhou, Q.**, **Liu, W.** ***et al.*** (2018) A proteomics landscape of circadian clock in mouse liver. *Nature Communications*, **9**(1), 1553.

Wilson, M.R., **Reske, J.J.**, **Holladay, J.**, **Wilber, G.E.**, **Rhodes, M.**, **Koeman, J.** ***et al.*** (2019) ARID1A and PI3-kinase pathway mutations in the endometrium drive epithelial transdifferentiation and collective invasion. *Nature Communications*, **10**(1), 3554.

Wos, G., **Macková, L.**, **Kubíková, K.** & **Kolář, F.** (2022) Ploidy and local environment drive intraspecific variation in Endoreduplication in *Arabidopsis arenosa*. *American Journal of Botany*, **109**(2), 259–271.

Wu, D. (2022) Mouse oocytes, a complex single cell transcriptome. *Frontiers in Cell and Developmental Biology*, **10**(March), 827937.

Yang, H., **Shi, X.**, **Chen, C.**, **Hou, J.**, **Ji, T.**, **Cheng, J.** ***et al.*** (2021) Predominantly inverse modulation of gene expression in genomically unbalanced disomic haploid maize. *The Plant Cell*, **33**(4), 901–916.

Yu, R., **Campbell, K.**, **Pereira, R.**, **Björkeroth, J.**, **Qi, Q.**, **Vorontsov, E.** ***et al.*** (2020) Nitrogen limitation reveals large reserves in metabolic and translational capacities of yeast. *Nature Communications*, **11**(1), 1881.

Yu, R., **Vorontsov, E.**, **Sihlbom, C.** & **Nielsen, J.** (2021) Quantifying absolute gene expression profiles reveals distinct regulation of central carbon metabolism genes in yeast. *eLife*, **10**(March), e65722. Available from: https://doi.org/10.7554/eLife.65722

Yu, Y., **Hao, H.**, **Doust, A.N.** & **Kellogg, E.A.** (2020) Divergent gene expression networks underlie morphological diversity of abscission zones in grasses. *The New Phytologist*, **225**(4), 1799–1815.

Zhang, H., **Li, X.**, **Song, R.**, **Zhan, Z.**, **Zhao, F.**, **Li, Z.** ***et al.*** (2022) Cap-binding complex assists RNA polymerase II transcription in plant salt stress response. *Plant, Cell & Environment*, **45**(9), 2780–2793.

Zhang, T., **Zhao, X.**, **Wang, W.**, **Pan, Y.**, **Huang, L.**, **Liu, X.** ***et al.*** (2012) Comparative transcriptome profiling of chilling stress responsiveness in two contrasting rice genotypes. *PLoS One*, **7**(8), e43274.

Zhao, L., **Wang, P.**, **Hou, H.**, **Zhang, H.**, **Wang, Y.**, **Yan, S.** ***et al.*** (2014) Transcriptional regulation of cell cycle genes in response to abiotic stresses correlates with dynamic changes in histone modifications in maize. *PLoS One*, **9**(8), e106070.

Zumel, D., **Diéguez, X.**, **Werner, O.**, **Moreno-Ortiz, M.C.**, **Muñoz, J.** & **Ros, R.M.** (2023) High Endoreduplication after drought-related conditions in haploid but not diploid mosses. *Annals of Botany*, mcad159. Available from: https://doi.org/10.1093/aob/mcad159