Dissecting Circadian Rhythm in Kalanchoë fedtschenkoi, a Model Species for Crassulacean Acid Metabolism Research

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Robert C. Moseley
August 2018

Copyright © 2018 by Robert C. Moseley
All rights reserved.

DEDICATION

I dedicate this work to my mother and father, Nanci and Clifford Moseley, and my fiancé Cierra Cathey, whose loving support shows no bounds.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Xiaohan Yang for his mentorship during my dissertation. I would also like to thank Dr. Scott Lenaghan and Dr. Steve Haase for their time and advising even though they had no obligation. I would like to thank my committee members Drs. Gerald A. Tuskan, Jay Chen, and Margaret E. Staton for their participation in this process, and for providing guidance and critiques to improve the way I approach science. I want to thank the Bredesen Center, specifically Dr. Lee Riedinger, for giving a small school kid a chance, as well as Wanda Davis, Tracey Bucher, and Jessica Garner for all their help over the last five years. I could not have done it without everyone mentioned here.

Lastly, I want to thank my friends and family who supported me through this process, and especially my soon-to-be wife, Cierra Cathey, for putting up with me.

ABSTRACT

The circadian clock plays a major role in the temporal regulation of an organism's metabolism and physiology. This role is even more important in plants due to their sessile lifestyle as it provides a fitness advantage by synchronizing internal metabolic and physiological processes with periodic environmental stimuli. A prime example of the importance of the circadian clock in plants can be seen in plants with crassulacean acid metabolism (CAM) photosynthesis. In comparison with C_3 [C three] or C_4 [C four] photosynthesis, CAM enhances water-use efficiency via inverted day/night stomatal movement pattern (i.e., stomata close during day and open at night). Also, CAM improves photosynthetic efficiency through a carbon concentration mechanism centered on a temporal separation of carbon dioxide fixation relative to C_3 photosynthesis. These advantages of CAM have garnered increased interest in understanding the molecular mechanism of CAM pathway. However, little is known about the regulatory networks controlling the temporal phases of CAM. The goal of this research is to better understand how the circadian clock regulates CAM-related processes and how it possibly influenced the evolution of CAM.

Through the use of various omics resources and computational methods and tools, the rhythmic transcriptomes of the model obligate CAM species, *Kalanchoë fedtschenkoi* and the model C_3 species, *Arabidopsis thaliana* were examined and compared. Similarities and differences in rhythmicity of gene expression were observed between these two species, leading to testable hypotheses on the circadian clock's role in CAM and its evolution. Furthermore, several candidate clock genes in *K. fedtschenkoi* were predicted as new members of the core clock network and the circadian output network. Evidence in support of the core clock network impacting the inversion of stomatal movement in *K. fedtschenkoi* was generated through use of gene regulation inference and an expanded list of stomata-related genes. Together, the results from this research provide a reduction in the hypothesis space for investigating the circadian regulation of CAM.

TABLE OF CONTENTS

INTRODUCTION....1
The circadian clock in plants....1
The circadian clock and crassulacean acid metabolism....3
Overview of studies linking the circadian clock and CAM....4
Scientific contribution and outline of dissertation....6
References....7
Appendix....11
CHAPTER I INVESTIGATING DIEL RE-SCHEDULING OF GENE EXPRESSION BETWEEN CAM AND C_3 SPECIES....14
Abstract....15
Introduction....15
Methods....16
Estimation of time delay between gene expression profiles....16
Defining convergent diel rescheduling of gene expression....16
Results....17
Convergent evolution of genes involved various processes....17
Discussion....18
Conclusion....19
References....20
Appendix....22
CHAPTER II CONSERVATION AND DIVERSIFICATION OF CIRCADIAN RHYTHMICITY BETWEEN CAM AND C_3 PHOTOSYNTHESIS PLANTS....24
Abstract....25
Introduction....25
Results....26
Identifying rhythmic genes in *A. thaliana* and *K. fedtschenkoi*....26
Functional comparison of rhythmic genes between *A. thaliana* and *K. fedtschenkoi*....27
Rhythmic ortholog groups shared between *A. thaliana* and *K. fedtschenkoi* (Type 1)....28
Rhythmic *A. thaliana* genes with arrhythmic *K. fedtschenkoi* orthologs (Type 2)....29
Rhythmic *K. fedtschenkoi* genes with arrhythmic *A. thaliana* orthologs (Type 3)....29
Rhythmic genes only in *A. thaliana* (Type 4)....30
Rhythmic genes only in *K. fedtschenkoi* (Type 5)....30
Core clock genes in *A. thaliana* and *K. fedtschenkoi*....30
Discussion....31
Conclusion....34
Methods....34
K. fedtschenkoi and *A. thaliana* time-course data....34
Ranking rhythmic genes in *K. fedtschenkoi* and *A. thaliana*....35
Functional analysis of rhythmic genes *K. fedtschenkoi* and *A. thaliana*....35
Identification of orthologous genes....36

References 37
Appendix 41
CHAPTER III COMPARATIVE GENOMICS ANALYSIS PROVIDES NEW INSIGHT INTO MOLECULAR BASIS OF STOMATAL DEVELOPMENT AND MOVEMENT IN *KALANCHOË FEDTSCHENKOI* 59
Abstract 60
Introduction 60
Methods 62
Ortholog groups and phylogenetic analysis 62
Gene ontology analysis of protein sequences 62
Data-mining using paperBLAST 62
Analysis of time-course gene expression data 63
Comparative analysis of gene expression 63
Results 64
Identification of stomata-related genes in 13 plant species 64
Differential enrichment of gene expression in dusk and dawn 65
Comparison of studies 66
Phylogenetic investigation 66
Discussion 67
Conclusion 70
References 71
Appendix 75
CHAPTER IV INFERENCE OF GENE REGULATORY NETWORK UNCOVERS THE LINKAGE BETWEEN CIRCADIAN CLOCK AND CRASSULACEAN ACID METABOLISM IN *KALANCHOË FEDTSCHENKOI* 98
Abstract 99
Introduction 99
Methods 101
Kalanchoë fedtschenkoi and *Arabidopsis thaliana* time-course data 101
Identifying candidate clock genes 102
Identifying candidate core clock genes 102
Identify clock-regulators of stomata-related genes 103
Gene Ontology analysis 103
Comparative analysis of gene expression 103
Results 104
Candidate clock-regulated genes in *K. fedtschenkoi* 104
Candidate clock transcription factors 104
Morning phased candidate clock transcription factors 105
Midday phased candidate clock transcription factors 106
Evening phased candidate clock transcription factors 106
Clock regulation of stomata-related genes in *K. fedtschenkoi* 107
Regulation of rescheduled stomata-related genes 108
Discussion 109
Conclusion 110

References 112
Appendix 117
CONCLUSION 128
VITA 129

LIST OF TABLES

Table 1. Time-shifts in diel gene expression profiles between orthologs in *Kalanchoë fedtschenkoi* (Kafe), *Ananas comosus* (Anco), and *Arabidopsis thaliana* (Arth)..... 22
Table 2. Carboxylation genes in *Kalanchoë fedtschenkoi* and their orthologs in *Arabidopsis thaliana*..... 41
Table 3. Decarboxylation genes in *Kalanchoë fedtschenkoi* and their orthologs in *Arabidopsis thaliana*..... 43
Table 4. Type 1 rhythmic orthologous transcription factors in *Kalanchoë fedtschenkoi* and *Arabidopsis thaliana*. 44
Table 5. Circadian parameters of *Kalanchoë fedtschenkoi* core clock orthologs. 45
Table 6. Plant species used to in PaperBLAST search..... 75
Table 7. Results of PaperBLAST search. 76
Table 8. Species in PaperBLAST database with stomata-related proteins. 77
Table 9. *Arabidopsis thaliana* genes with stomata-related annotations or known as key stomatal genes. 78
Table 10. *Arabidopsis thaliana* genes identified as new stomata-related genes. 82
Table 11. Categorization of whether genes in each species were placed in the same ortholog group (OG) as the gene they matched to in the PaperBLAST database..... 86
Table 12. Categorization of genes placed in the same ortholog group as the gene they matched to in the PaperBLAST database based on whether the matched gene was annotated as stomata-related or known as a key stomatal gene. 87
Table 13. Orthologous genes *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Kalanchoë fedtschenkoi* that displayed differential enrichment between dusk and dawn in gene expression. 88
Table 14. Comparison of results between this study and two other studies investigating gene expression rescheduling. 89
Table 15. Genes identified as clock-regulated in *Kalanchoë fedtschenkoi* that are orthologous to circadian-related *Arabidopsis thaliana* genes. 117
Table 16. Candidate clock transcription factors in *Kalanchoë fedtschenkoi*. 118
Table 17. Regulation relationships between clock transcription factors and stomata-related genes in *Kalanchoë fedtschenkoi.* 119
Table 18. Clock-controlled stomata-related genes with rescheduled gene expression from different studies. 120

LIST OF FIGURES

Figure 1. Overviews of circadian system, the core circadian oscillator, and crassulacean acid metabolism. 11
Figure 2. Overview of major themes within the dissertation. 13
Figure 3. Examples of convergent change in diel expression profile in CAM species. 23
Figure 4. Rhythmically expressed genes in *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi*. 46
Figure 5. Gene ontology categories overrepresented among *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi* rhythmic genes. 47
Figure 6. Gene co-expression networks of the rhythmic gene sets in *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi* rhythmic gene sets. 49
Figure 7. Categories of rhythmic orthologous genes in *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi.* 52
Figure 8. Gene co-expression networks of Type 1 ortholog groups containing only one rhythmic gene in each of the two plant species (*Arabidopsis thaliana* and *Kalanchoë fedtschenkoi)*. 54
Figure 9. Gene co-expression networks and phase call distributions of Type 2 and Type 3 ortholog groups described in Figure 8. 56
Figure 10. Diurnal expression profiles of *Arabidopsis thaliana* core clock genes and their *Kalanchoë fedtschenkoi* orthologs. 57
Figure 11. Phase of day for core clock genes in *Kalanchoë fedtschenkoi.* 58
Figure 12. History of number of publications related to crassulacean acid metabolism, stomata, and guard cell research. 90
Figure 13. Description of dawn and dusk phases for each species and the contingency tables used to gene expression enrichment. 91
Figure 14. Overlap of *Arabidopsis thaliana* genes with annotations related to stomata, known as key stomatal genes, and identified in PaperBLAST database as stomata-related. 92
Figure 15. Gene ontology enrichment of new *Arabidopsis thaliana* stomatal genes. 93
Figure 16. Overlap of ortholog groups between new stomata-related genes of *Arabidopsis thaliana* (Arth), *Solanum lycopersicum* (Soly), and *Kalanchoë fedtschenkoi* (Kafe). 94
Figure 17. Ortholog groups displaying enrichment time of each ortholog and phylogenetic tree and expression dynamics of catalase genes between *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Kalanchoë fedtschenkoi*. 95
Figure 18. Catalase 2's (CAT2) roles in C_3 plants and predicted roles in CAM plants. 97
Figure 19. DLxJTK rank ordered *Kalanchoë fedtschenkoi* genes. 121
Figure 20. Several candidate clock transcription factors are predicted to regulate and be regulated by known core clock transcription factors. 122
Figure 21. Likelihood rankings of candidate clock transcription factors. 123
Figure 22. Gene expression of *Kalanchoë fedtschenkoi* candidate transcriptions and their orthologs in *Arabidopsis thaliana*. 124

Figure 23. *Kalanchoë fedtschenkoi* clock transcription factors and stomata-related genes regulatory network..... 125
Figure 24. *Kalanchoë fedtschenkoi* candidate transcription gene expression and phase calls of their respective targets. 126
Figure 25. Enriched functional groups within of clock-controlled stomata-related genes in *Kalanchoë fedtschenkoi*. 127

LIST OF ATTACHMENTS

Table S1. Arabidopsis rhythmic data…….. SUPMAT_Table_1.csv
Table S2. Kalanchoe rhythmic data……..... SUPMAT_Table_2.csv
Table S3. Ortholog group types (gene lists)…… SUPMAT_Table_3.xlsx
Table S4. Ortholog group Type GO enrichments……..... SUPMAT_Table_4. xlsx
Table S5. Local BLAST database information……... SUPMAT_Table_5.csv
Table S6. Type 5 ortholog group BLAST results……. SUPMAT_Table_6. xlsx
Table S7. Type 5 ortholog group Kalanchoe specific genes…... SUPMAT_Table_7. xlsx
Table S8. Type 5 ortholog group CAM specific genes…….…... SUPMAT_Table_8. xlsx
Table S9. All filtered PaperBLAST output……. SUPMAT_Table_9.txt
Table S10. DLxJTK output……... SUPMAT_Table_10.csv
Table S11. DLxJTK TF information……. SUPMAT_Table_11.csv
Table S12. Likelihood rankings……..… SUPMAT_Table_12. xlsx
Table S13. LEM output of clock regulation of stomata genes... SUPMAT_Table_13. xlsx
Table S14. Remaining EC genes and targets data……. SUPMAT_Table_14.csv

Figure S1. Heatmaps from visual inspection of Arabidopsis and Kalanchoe……
……..... SUPMAT_Figure_1.docx
Figure S2. Phylogenetic trees of selected ortholog groups ……...
……..….... SUPMAT_Figure_2.docx
Figure S3. Kaladp0674s0182 (CTF8) Pfam search results……
…….....…SUPMAT_Figure_3.docx

INTRODUCTION

The circadian clock in plants

The term *circadian* comes from the Latin word *circa*, meaning "around", and *diēm*, meaning "day" and the term *circadian clock* is used to describe the endogenous timekeeping network that allows organisms to synchronize their physiology with their changing environment and to perform biological processes at the most relevant times of the day and year (Nohales and Kay 2016). The circadian clock found in all organisms is thought to have evolved in response to the predictable day-night cycle of light and temperature due to the 24-h rotation of the Earth (Millar 2016). Clocks in eukaryotes have independent evolutionary origins but rely on transcription- and translation-feedback loops (Bell-Pederson et al. 2005), termed the core clock or core circuit. Further research over the past several decades has added additional interlocked feedback loops and clock components to the circadian clock network, creating a more complex concept of the circadian clock (Nohales and Kay 2016). Currently, this paradigm of interlocked transcriptional/translational feedback loops underlines the molecular mechanisms of the clock in all eukaryotes studied to date (McClung 2006).

The overall network architecture is conserved across kingdoms, but the nature of the individual clock components may differ (McClung 2006). Plants are a good example of this, when the sequencing of *Arabidopsis* genome was completed in 2000 (Initiative 2000), no identifiable orthologs to most known animal clock proteins were found. This finding indicates that the clock mechanism in *Arabidopsis*, at least in terms of its configuration, is novel in comparison with animals. Since then, genes involved in the circadian clock have been identified in a lot of sequenced plant genomes. Numerous reviews have discussed the conservation of the paradigm of interlocked feedback loops in plants and have highlighted the prevalence of repressor elements in plant clock systems (McClung 2006; Song et al. 2010; McClung 2013; Nohales and Kay 2016).

In plants, the circadian clock is ever more important due to the sessile life of plants. The clock provides a fitness advantage by allowing coordinated responses to multiple and often simultaneous environmental challenges that a plant might face, and hence increases fitness (Hotta et al. 2007). For instance, plants cannot move away from or retreat from pathogens or herbivores who pose a threat. Instead, the clock has been recruited to temporally regulate defense responses to biotic stresses, a concept included in the optimal defense hypothesis, or ODH (Ming et al. 2015). ODH posits that plants use the circadian clock to temporally restrict both the basal expression and the induction of defense pathways to the time of day when the threat posed by pathogens and herbivores is maximal (Greenham and McClung 2015). Indeed, a study by (Arimura et al. 2008) in lima beans reported that herbivory damage at night is more effective than daytime damage at stimulating accretion of defense compounds, such as jasmonic acid, and emission of defense-associated volatiles.

The process by which the circadian clock regulates biological processes can be divided into: 1) input pathways which transduce the environmental signals to synchronize the internal clock with local time; 2) the central oscillator which provides the rhythmicity of the clock; and 3) the output pathways which couple the activity of the oscillator to the

observable rhythms (Hotta et al. 2007; Nohales and Kay 2016) (Figure 1). Light and temperature are great signals to convey time to plants and, indeed, are needed to properly synchronize the clock with the surrounding environment. Light's role in the pacing of the clock is pervasive, and in turn, its perception and signaling are regulated by the clock, a mechanism termed circadian gating. Multiple levels of regulation, from transcription to post-translation, have been found to be affected by light quality and intensity and several clock and light-signaling components have been linked to clock sensitivity to light (Nohales and Kay 2016). Additionally, photoreceptor expressions are subordinate to circadian control (Fankhauser and Staiger 2002; Nohales and Kay 2016). As for temperature, several clock genes have been observed to be transcriptionally sensitive to changes in temperature (Gould et al. 2006; Rugnone et al. 2013). Even though temperature input can reset and entrain the clock, the central oscillator is buffered against temperature changes, an ability called temperature compensation. This ability allows the oscillator to maintain a relatively constant pace over a range of temperatures (e.g., 12-27°C), buffering the effects of minor and physiologically irrelevant temperature changes (Gould et al. 2006). While the influence of light and temperature on the clock is known, how this information is broadcasted to and integrated by the central oscillator is still not fully understood.

The second component, the central oscillator (Figure 1A), is where the paradigm of interlocked loops comes into play. The first description of the oscillator in plants consisted of a single transcriptional loop (Alabadi et al. 2001). This loop was composed of three genes, circadian clock associated 1 (CCA1) and late elongated hypocotyl (LHY), which are expressed in the morning and timing of CAB expression 1 (TOC1), also known as pseudo-response regulator 1 (PRR1), which is expressed 12 h after dawn. TOC1 acts as a DNA-binding transcriptional repressor of CCA1 and LHY. CCA1 and LHY, in turn, act together to repress TOC1, forming a feedback loop. Several studies have confirmed the relevance of these repressors in precise timekeeping through loss-of-function experiments (Strayer et al. 2000; Alabadi et al. 2002). For example, upon loss of any of the three genes, period length was shortened under free-running conditions, and furthermore, a *cca1:lhy* double mutant exhibited an arrhythmic circadian phenotype. Since this discovery, several additional clock components have been identified and integrated into the clock network, creating a more complex, wired network of interlocked evening and morning feedback loops (Figure 1B).

The extent to which the circadian clock controls biological processes, or output pathways, in plants is still under debate (McClung 2013). Original estimates, based on microarray analyses in *Arabidopsis*, suggested that approximately 5-6% of the transcripts cycled in abundance under continuous light (Harmer et al. 2000), a rather conservative estimate which prompted further inquiry. Several years later, Michael and McClung (2003) suggested that up to one-third of the *Arabidopsis* transcriptome was regulated by the circadian clock. The latest estimate from a study examining rhythmicity in gene expression under multiple photo- and thermo-cycles, and combinations of these photo- and thermo-cycles, is that up to 89% of the *Arabidopsis* transcriptome cycles under at least one diurnal or circadian condition (Michael et al. 2008). Whatever the extent the clock controls biological processes, numerous processes have been documented to be

subordinate to the clock (Figure 1A), such as starch metabolism (Piechulla 1988; Merida et al. 1999; Mallona et al. 2011; Ming et al. 2015), stomatal opening and therefore gas exchange (Hotta et al. 2007), and cotyledon movement (Thain et al. 2000).

The circadian clock and crassulacean acid metabolism

The importance of the circadian clock in temporal regulation of metabolism and physiology is highlighted in plant species that perform crassulacean acid metabolism (CAM) photosynthesis. CAM enhances photosynthetic efficiency through a carbon concentration mechanism and improves water-use efficiency and associated drought tolerance through a reduction of transpirational water loss (Winter et al. 2005; Borland et al. 2009; West-Eberhard et al. 2011) resulting from a temporal separation of CO_2 fixation and an inverted timing of stomatal movement relative to plant species that conduct C_3 or C_4 photosynthesis (Cushman 2001). Carbon fixation in CAM species is commonly described in four temporal phases that integrate patterns of net carbon dioxide (CO_2) uptake and changes in the abundances of key metabolites that define carbon supply and demand over the 24-h cycle (Figure 1C). Specifically, Phase I is characterized by high stomatal conductance and nocturnal CO_2 uptake and assimilation of atmospheric and respiratory CO_2 by phosphoenolpyruvate carboxylase (PEPC) into oxaloacetic acid, which is subsequently reduced to a C_4 acid, usually malate, by NAD(P)-malate dehydrogenase (MDH) and stored in the vacuole. Concomitantly, depletion in storage carbohydrates (e.g., soluble sugars or starch) occurs, which in turn provides the phosphoenolpyruvate (PEP) building blocks to sustain malate formation (Ceusters et al. 2014). Phase I is associated with improved water-use efficiency as the leaf-air vapor pressure deficit is usually lower at night than during the day. Phase II is described as the transition from dark C_4 carboxylation by PEPC to daytime Rubisco-mediated carboxylation as photosynthetically active radiation increases in the early morning. Malate begins to be remobilized from the vacuole and decarboxylated. A blend of atmospheric CO_2 and CO_2 released from the decarboxylation of malate is fixed during this phase. Phase III is defined by daytime C_3 photosynthesis, closed stomata, and major decarboxylation of C_4 acids either by NAD(P)-malic enzyme (ME) or PEP carboxykinase (PEPCK) depending on the species (Ming et al. 2015). Net CO_2 uptake falls to zero and CO_2 from decarboxylation of organic acids during this phase is concentrated behind closed stomata, with concentrations ranging from two-fold to sixty-fold around Rubisco (Luttge 2002). The concentrated CO_2 around Rubisco reduces the oxygenase activity of Rubisco and consequently represses photorespiration in the plant. CO_2 is refixed by Rubisco and assimilated via the Calvin cycle of photosynthesis. Pyruvate or PEP released from the decarboxylation of C_4 acids is converted into storage carbohydrates and reach peak accumulation at the end of the light period. Lastly, Phase IV is characterized by the continuation of partial net CO_2 uptake toward the end of the light period due to the exhaustion of malate stores. This results in a decline in leaf internal partial pressure of CO_2 and an increase in stomatal conductance and transpiration. Rubisco may continue to fix CO_2 during this phase till the end of the light period. C_3 and C_4 carboxylation events can occur in both Phase II and IV, with C_4 carboxylation increasing as the dark period nears in Phase IV and C_3 carboxylation increasing as the light period nears in Phase II.

CAM is well known for its plasticity in the expression of these four phases, but they are generally considered a trademark of CAM expression (Osmond 1978; Cushman 2001; Borland et al. 2014). It is this increased temporal dependence on metabolism and stomatal conductance that has led researchers to believe the clock plays a role in CAM.

While the core processes of CAM, such as carboxylation and decarboxylation, are well understood, little is known about the fundamental signal transduction pathways and regulatory relationships that control the temporal phases of CAM and allow for processes in CAM to synchronize with the 24-h light/dark cycle (Hartwell 2006; Borland et al. 2009; Davies and Griffiths 2012; Dever et al. 2015). There has been much debate on the role the circadian clock plays in CAM-performing plant species and the requirement for a CAM-specific oscillator to allow proper CAM functioning (Cushman 2001; Boxall et al. 2005; Mallona et al. 2011; Davies and Griffiths 2012; Ming et al. 2015; Winter et al. 2015; Abraham et al. 2016; Borland et al. 2016; Hartwell et al. 2016; Yang et al. 2017). An initial theory proposed an underlying oscillator based on malate oscillations and rhythmic changes in the properties of the tonoplast membrane (Wilkins 1992; Hausler et al. 2000; Borland and Dodd 2002; Luttge 2002; Dodd et al. 2003), with support being generated from a computer model demonstrating that this bio-physical oscillator could sustain robust rhythmicity of CAM over a range of temperatures (Luttge 2000). However, this theory was contradicted by (Wyka et al. 2004) when they found that preventing nocturnal malate accumulation in *Kalanchoë daigremontiana* did not cause the predicted phase delays generated by the computer model (Luttge 2000). The current theory stipulates that movement of malate in and out of the vacuole responds to the circadian clock, with the oscillator centered around the CCA1/LHY-TOC1 gene loop (Wyka et al. 2004; Boxall et al. 2005; Hartwell 2005; von Caemmerer and Griffiths 2009). An overview of investigations into how the clock and CAM are coupled is described in the next section.

Overview of studies linking the circadian clock and CAM

The earliest evidence in support of the central clock oscillator controlling CAM processes was reported by (Hartwell 2006) using a phase shifting experiment, which can provide meaningful insight into whether two clock components respond to the same central oscillator. CAM-induced leaves of *Mesembryanthemum crystallinum* were grown at 23°C under light/dark conditions before being subjected to a temperature pulse in the second half of the first subjective night. His results revealed consistent phase advance in the rhythms of several central clock genes and clock-controlled, CAM-related genes following the temperature pulse. Phase advance in malate oscillations was also observed within the leaves. Malate oscillations serve as a good marker for circadian-regulated CO_2 fixation via the CAM pathway as malate oscillations track oscillations in CO_2 fixation by PEPC and malate decarboxylation by malic enzyme (ME). The phase advances in both transcripts and malate oscillations under constant light conditions served as the first evidence that the *CCA1*/*LHY*-*TOC1* gene loop provides temporal information for the synchronization of CO_2 fixation by the CAM pathway (Hartwell 2006).

Recently, (Dever et al. 2015) demonstrated the link between CCA1/LHY-TOC1 oscillator and CAM pathway through silencing of light-period, light-activated CAM

enzymes, NAD-ME and pyruvate orthophosphate dikinase (PPDK) in *M. crystallinum*. NAD-ME is one of three enzymes involved in the decarboxylation of malate and PPDK is responsible for gluconeogenic recovery of PEP from pyruvate, the by-product of CO_2 fixation by Rubisco. While the silencing of PPDK showed no loss in rhythmicity of the core clock gene TOC1, the NAD-ME mutant line resulted in arrhythmia of TOC1 expression, as well as reduced PEPC phosphorylation in both light/dark and constant light conditions. This indicates that failure to decarboxylate malate in the light could be feeding back to perturb the core clock oscillator, suggesting that maintenance of rhythmicity in CAM may require cross talk between the metabolism of CAM and the CCA1/LHY-TOC1 oscillator (Dever et al. 2015).

The CAM-inducible *M. crystallinum* was the focus of yet another study looking at circadian control of CAM-related genes. (Cushman et al. 2008) performed a large-scale mRNA expression profiling study on samples taken from leaves conducting C_3 photosynthesis and CAM over a 24-h period and under constant light and temperature conditions. The study found that out of the 8,455 genes identified, 2,343 showed a significant change in steady-state mRNA abundance at one or more time-points. Of these, 858 and 599 exhibited a greater than two-fold increase or decrease in mRNA abundance, respectively. Upon functional characterization, they found many of these genes encoded products functioning in various processes, including CAM-related C_4 acid carboxylation and decarboxylation. Additionally, peak expression for genes with CAM-related functions were shifted 4-8 h and displayed putative circadian clock-regulated expression patterns with a majority exhibiting peak expression during the subjective evening.

Outside of *M. crystallinum*, additional support for the coupling between CAM and the core clock was found in analysis of the *Ananas comosus* genome, where an enrichment of *cis*-regulatory elements of circadian clock regulatory genes was found in diurnally expressed photosynthetic genes relative to non-photosynthetic gene copies (Ming et al. 2015). Specifically, promoter regions for beta-carbonic anhydrase (ß-CA) genes responsible for the first step in CO_2 fixation in C_4 and CAM photosynthesis, were enriched for CCA1-binding motifs. These motifs allow for binding of CCA1 and LHY to modify transcription and when the promoter regions of ß-CA in C_3 and C_4 monocots were examined, only one gene in sorghum (no known photosynthetic function) contained the CCA1-binding site. This finding suggests that ß-CA genes in pineapple are under circadian control to synchronize expression with stomatal opening at night to maximize CO_2 fixation (Ming et al. 2015), and further supported the notion that CCA1/LHY-TOC1 gene loop plays a role in synchronizing CAM.

Two follow-up studies in *A. comosus* were conducted to investigate circadian gene expression, with one surveying global gene expression in pineapple leaf tissues spatially and temporally (Wai et al. 2017) and another examining transcription factor (TF) and transcription co-regulator (TC) expression in various pineapple tissues (Sharma et al. 2017). The first study used a clustering approach to examine differential expression between photosynthetic and non-photosynthetic leaf tissues. Most CAM genes were localized in clusters with strong diurnal expression patterns. Furthermore, the clusters were enriched with core circadian clock genes and *cis*-regulatory sequences, as well as immediate downstream processes associated with carbon fixation, stomatal movement,

and carbohydrate transport (Wai et al. 2017). The second study used phylogeny and comparative expression dynamics to identify clock genes. They used HAYSTACK (Mockler et al. 2007), an algorithm for predicting rhythmic expression through model fitting of gene expression to oscillating functions, to predict rhythmic TF/TCs. They found that a set of 68 TF/TCs, whose rhythmic expression was tightly coupled between photosynthetic and non-photosynthetic leaf tissues types, were enriched in homologous genes that form the core circadian clock in *A. thaliana* (Sharma et al. 2017). These studies combined infer direct relationships between the circadian clock and CAM processes, as well as the conservation of the core circadian clock between CAM and C_3 plant species.

Scientific contribution and outline of dissertation

There has been a resurgence in CAM research due to plants which conduct CAM having enhanced photosynthetic efficiency through reduced photorespiration and improved WUE through the inverted day/night stomatal conductance pattern relative to C_3 and C_4 plants. Understanding the molecular mechanisms that govern CAM can lead to engineering of CAM pathways into non-CAM plants, enabling agricultural crops and bioenergy crops with these advantageous properties. Empowering these crops would allow for growth on marginal land or endow the crops with a drought tolerance or avoidance mechanism. The circadian clock is a prime candidate as a regulator of CAM processes and fully understanding its role in CAM is paramount to achieving the goal of engineering CAM properties into non-CAM plants.

Thus, the goal of this dissertation is to expand the knowledge of how the circadian clock and CAM are linked. Four objectives will be completed to accomplish this goal: 1) Create a comprehensive catalog of rhythmic genes in a model CAM and C_3 species, 2) investigate the conservation and diversification of rhythmic gene expression between the model CAM and C_3 species, 3) expand the list of genes involved in stomata-related processes within a CAM species and investigate the molecular mechanism of inverted stomatal movement, and 4) infer regulatory relationships between the circadian clock and stomata-related genes in a CAM species. The dissertation is divided into four chapters, with each chapter written as a separate manuscript for publication. Figure 2 outlines the framework of the dissertation as well as the general theme of each chapter. Chapter 1 will detail the comparative analysis of gene expression performed in (Yang et al. 2017) and contextualize in the scope of the dissertation. Chapter 2 will describe the comparative analysis of rhythmic transcriptomes between the model obligate CAM species, *K. fedtschenkoi*, and model C_3 species, *A. thaliana*, using a new periodicity detection method. This work will highlight the conserved and diverged control of the circadian clock between these two photosynthetic types and shed light on how the clock could have played a role in the evolution of CAM. Chapter 3 will present results from investigating the molecular mechanisms behind the inversion of day/night stomatal movement pattern in CAM species in comparison with C_3 species. Additionally, new genes involved in stomata-related processes across 13 plant species will be presented. Lastly, Chapter 4 will present the findings of inferring circadian regulation of stomatal movement in *K. fedtschenkoi* using a new inference method.

References

Abraham PE, Yin H, Borland AM, Weighill D, Lim SD, De Paoli HC, Engle N, Jones PC, Agh R, Weston DJ et al. 2016. Transcript, protein and metabolite temporal dynamics in the CAM plant Agave. *Nat Plants* **2**: 16178.

Alabadi D, Oyama T, Yanovsky M, Harmon FG, Mas P, Kay SA. 2001. Reciprocal Regulation Between TOC1 and LHY/CCA1 Within the Arabidopsis Circadian Clock. *Science* **293**: 880-883.

Alabadi D, Yanovsky MJ, Mas P, Harmer SL, Kay SA. 2002. Critical Role for CCA1 and LHY in Maintaining Circadian Rhythmicity in Arabidopsis. *Current Biology* **12**: 757-761.

Arimura G, Kopke S, Kunert M, Volpe V, David A, Brand P, Dabrowska P, Maffei ME, Boland W. 2008. Effects of feeding Spodoptera littoralis on lima bean leaves: IV. Diurnal and nocturnal damage differentially initiate plant volatile emission. *Plant Physiol* **146**: 965-973.

Bell-Pederson D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, Zoran MJ. 2005. CIRCADIAN RHYTHMS FROM MULTIPLE OSCILLATORS: LESSONS FROM DIVERSE ORGANISMS. *Nature Reviews: Genetics*: 1-13.

Borland AM, Dodd AN. 2002. Carbohydrate partitioning in crassulacean acid metabolism plants: reconciling potential conflicts of interest. *Functional Plant Biology* **29**: 707-716.

Borland AM, Griffiths H, Hartwell J, Smith JA. 2009. Exploiting the potential of plants with crassulacean acid metabolism for bioenergy production on marginal lands. *J Exp Bot* **60**: 2879-2896.

Borland AM, Guo HB, Yang X, Cushman JC. 2016. Orchestration of carbohydrate processing for crassulacean acid metabolism. *Curr Opin Plant Biol* **31**: 118-124.

Borland AM, Hartwell J, Weston DJ, Schlauch KA, Tschaplinski TJ, Tuskan GA, Yang X, Cushman JC. 2014. Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci* **19**: 327-338.

Boxall SF, Foster JM, Bohnert HJ, Cushman JC, Nimmo HG, Hartwell J. 2005. Conservation and divergence of circadian clock operation in a stress-inducible Crassulacean acid metabolism species reveals clock compensation against stress. *Plant Physiol* **137**: 969-982.

Ceusters J, Borland AM, Taybi T, Frans M, Godts C, De Proft MP. 2014. Light quality modulates metabolic synchronization over the diel phases of crassulacean acid metabolism. *J Exp Bot* **65**: 3705-3714.

Cushman JC. 2001. Crassulacean Acid Metabolism. A Plastic Photosynthetic Adaptation to Arid Environments. *Plant Physiology* **127**: 1439-1448.

Cushman JC, Tillett RL, Wood JA, Branco JM, Schlauch KA. 2008. Large-scale mRNA expression profiling in the common ice plant, Mesembryanthemum crystallinum, performing C3 photosynthesis and Crassulacean acid metabolism (CAM). *J Exp Bot* **59**: 1875-1894.

Davies BN, Griffiths H. 2012. Competing carboxylases: circadian and metabolic regulation of Rubisco in C3 and CAM Mesembryanthemum crystallinum L. *Plant Cell Environ* **35**: 1211-1220.
Dever LV, Boxall SF, Knerova J, Hartwell J. 2015. Transgenic perturbation of the decarboxylation phase of Crassulacean acid metabolism alters physiology and metabolism but has only a small effect on growth. *Plant Physiol* **167**: 44-59.
Dodd AN, Griffiths H, Taybi T, Cushman JC, Borland AM. 2003. Integrating diel starch metabolism with the circadian and environmental regulation of Crassulacean acid metabolism in Mesembryanthemum crystallinum. *Planta* **216**: 789-797.
Fankhauser C, Staiger D. 2002. Photoreceptors in Arabidopsis thaliana: light perception, signal transduction and entrainment of the endogenous clock. *Planta* **216**: 1-16.
Gould PD, Locke JC, Larue C, Southern MM, Davis SJ, Hanano S, Moyle R, Milich R, Putterill J, Millar AJ et al. 2006. The molecular basis of temperature compensation in the Arabidopsis circadian clock. *Plant Cell* **18**: 1177-1187.
Greenham K, McClung CR. 2015. Integrating circadian dynamics with physiological processes in plants. *Nat Rev Genet* **16**: 598-610.
Harmer SL, Hogenesch JB, Straume M, Chang H, Han B, Zhu T, Wang X, Kreps JA, Kay SA. 2000. Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. *Science* **290**: 2110-2113.
Hartwell J. 2005. The co-ordination of central plant metabolism by the circadian clock. *Mechanistic and Functional Studies of Proteins* **33**: 945-948.
Hartwell J. 2006. The circadian clock in CAM plants. *Endogenous Plant Rhythms*: 211-236.
Hartwell J, Dever LV, Boxall SF. 2016. Emerging model systems for functional genomics analysis of Crassulacean acid metabolism. *Curr Opin Plant Biol* **31**: 100-108.
Hausler RE, Baur B, Scharte J, Teichmann T, Eicks M, Fischer KL, Flugge U, Schubert S, Weber A, Fischer K. 2000. Plastidic metabolite transporters and their physiological functions in the inducible crassulacean acid metabolism plant Mesembryanthemum crystallinum. *The Plant Journal* **24**: 285-296.
Hotta CT, Gardner MJ, Hubbard KE, Baek SJ, Dalchau N, Suhita D, Dodd AN, Webb AA. 2007. Modulation of environmental responses of plants by circadian clocks. *Plant Cell Environ* **30**: 333-349.
Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**.
Luttge U. 2000. The tonoplast functioning as the master switch for circadian regulation of crassulacean acid metabolism. *Planta* **2000**: 761-769.
Luttge U. 2002. CO2-concentrating: consequences in crassulacean acid metabolism. *Journal of Experimental Botany* **53**: 2131-2142.
Mallona I, Egea-Cortines M, Weiss J. 2011. Conserved and divergent rhythms of crassulacean acid metabolism-related and core clock gene expression in the cactus Opuntia ficus-indica. *Plant Physiol* **156**: 1978-1989.
McClung CR. 2006. Plant Circadian Rhythms. *The Plant Cell* **18**: 792-803.

McClung CR. 2013. Beyond Arabidopsis: the circadian clock in non-model plant species. *Semin Cell Dev Biol* **24**: 430-436.
Merida A, Rodriquez-Galan JM, Vincent C, Romero JM. 1999. Expression of the Granule-Bound Starch Synthase I (Waxy) Gene from Snapdragon Is Developmentally and Circadian Clock Regulated. *Plant Physiology* **120**: 401-409.
Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM et al. 2008. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet* **4**: e14.
Millar AJ. 2016. The Intracellular Dynamics of Circadian Clocks Reach for the Light of Ecology and Evolution. *Annu Rev Plant Biol* **67**: 595-618.
Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* **47**: 1435-1442.
Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J. 2007. The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis. *Cold Spring Harbor Symposia on Quantitative Biology* **72**: 353-363.
Nohales MA, Kay SA. 2016. Molecular mechanisms at the core of the plant circadian oscillator. *Nat Struct Mol Biol* **23**: 1061-1069.
Osmond CB. 1978. CRASSULACEAN ACID METABOLISM: A CURIOSITY IN CONTEXT. *Annual Review of Plant Physiology* **29**: 379-414.
Piechulla B. 1988. Plastid and nuclear mRNA fluctuations in tomato leaves - diurnal and circadian rhythms during extended dark and light periods. *Plant Molecular Biology* **11**: 345-353.
Rugnone ML, Faigon Soverna A, Sanchez SE, Schlaen RG, Hernando CE, Seymour DK, Mancini E, Chernomoretz A, Weigel D, Mas P et al. 2013. LNK genes integrate light and clock signaling networks at the core of the Arabidopsis oscillator. *Proc Natl Acad Sci U S A* **110**: 12120-12125.
Sharma A, Wai CM, Ming R, Yu Q. 2017. Diurnal Cycling Transcription Factors of Pineapple Revealed by Genome-Wide Annotation and Global Transcriptomic Analysis. *Genome Biol Evol* **9**: 2170-2190.
Song YH, Ito S, Imaizumi T. 2010. Similarities in the circadian clock and photoperiodism in plants. *Curr Opin Plant Biol* **13**: 594-603.
Strayer C, Oyama T, Schultz TF, Raman R, Somers DE, Mas P, Panda S, Kreps JA, Kay SA. 2000. Cloning of the Arabidopsis Clock Gene TOC1, an Autoregulatory Response Regulator Homolog. *Science* **289**: 768-771.
Thain S, Hall A, Millar AJ. 2000. Functional independence of circadian clocks that regulate plant gene expression. *Current Biology* **10**: 951-956.
von Caemmerer S, Griffiths H. 2009. Stomatal responses to CO2 during a diel Crassulacean acid metabolism cycle in Kalanchoe daigremontiana and Kalanchoe pinnata. *Plant Cell Environ* **32**: 567-576.
Wai CM, VanBuren R, Zhang J, Huang L, Miao W, Edger PP, Yim WC, Priest HD, Meyers BC, Mockler T et al. 2017. Temporal and spatial transcriptomic and microRNA dynamics of CAM photosynthesis in pineapple. *Plant J* **92**: 19-30.

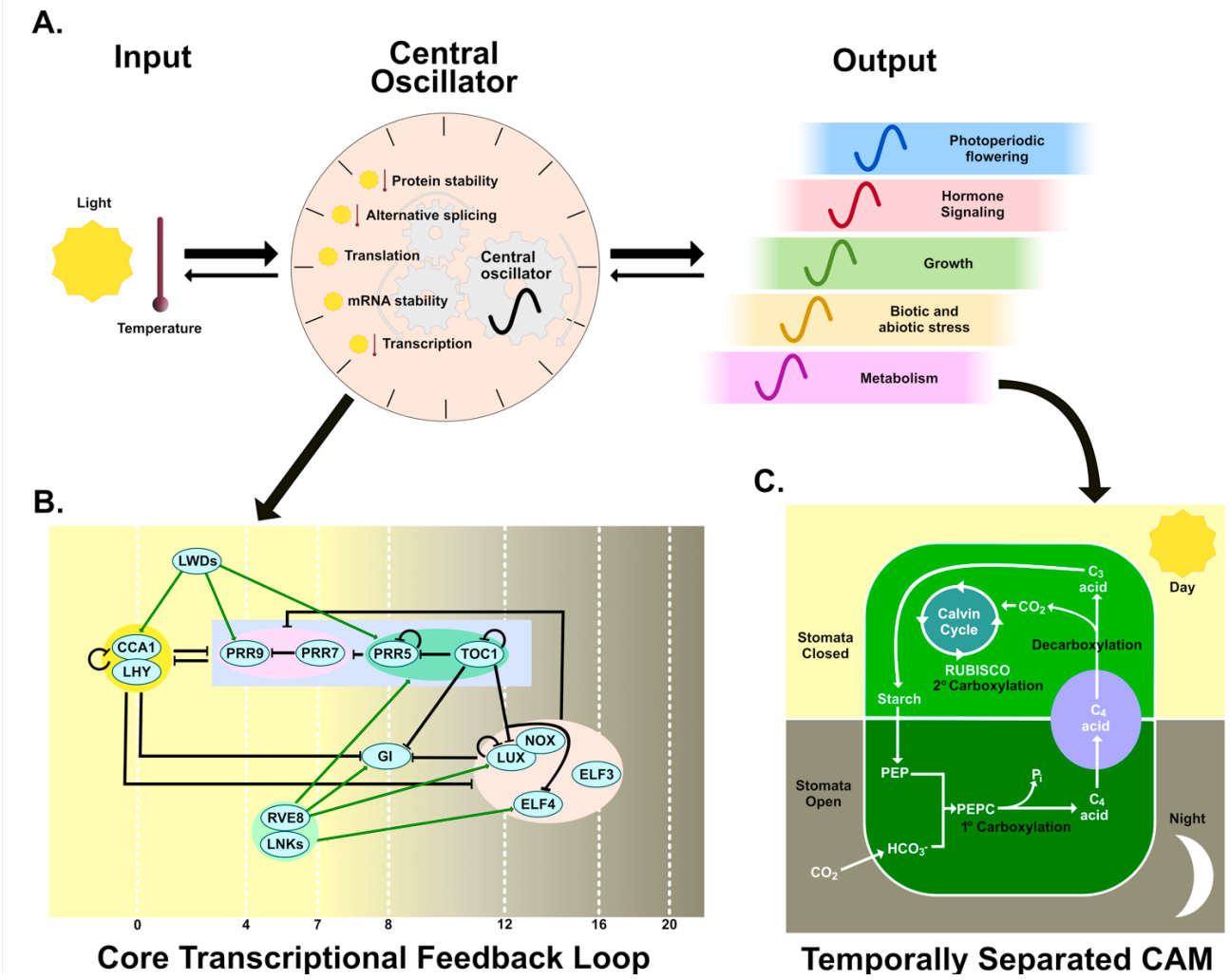
West-Eberhard MJ, Smith JA, Winter K. 2011. Plant science. Photosynthesis, reorganized. *Science* **332**: 311-312.
Wilkins M. 1992. Tansley Review No. 37 Circadian rhythms: their origin and control. *New Phytologist* **121**: 347-375.
Winter K, Aranda J, Holtum JAM. 2005. Carbon isotope composition and water-use efficiency in plants with crassulacean acid metabolism. *Functional Plant Biology* **32**.
Winter K, Holtum JA, Smith JA. 2015. Crassulacean acid metabolism: a continuous or discrete trait? *New Phytol* **208**: 73-78.
Wyka TP, Bohn A, Duarte HM, Kaiser F, Luttge UE. 2004. Perturbations of malate accumulation and the endogenous rhythms of gas exchange in the Crassulacean acid metabolism plant Kalanchoe daigremontiana: testing the tonoplast-as-oscillator model. *Planta* **219**: 705-713.
Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, Liu D, Weigel D, Cheol Yim W, Ha J et al. 2017. The Kalanchoë genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications* **8**: 1-15.

Appendix

Figure 1. Overviews of circadian system, the core circadian oscillator, and crassulacean acid metabolism.

A) Light and temperature are the two main environmental signals that are integrated by the central oscillator within plants. The signals influence the pace of the clock, denoted by the black wave, and entrain it by influencing different molecular processes at the core of the oscillator. Output rhythms, denoted by colored waves, are subordinate to the clock, and proper clock function is important for the arrangement of multiple physiological processes. Traditionally, the circadian system has been seen as a linear pathway; however, recent evidence supports the circadian system as a highly intricate network. The core circadian oscillator, while regulated by external stimuli, also regulates its own sensitivity to external stimuli. Furthermore, output pathways have also been seen to regulate clock function. **B)** The transcriptional feedback loop depicted here is the core of the circadian oscillator. The successive expression of each gene throughout the day is shown from left to right, and the time of activity is expression in hours after first light. The yellow and black areas represent day and night, respectively. Green arrows represent activation of transcription and black bars represent repression of transcription. The blue ovals represent proteins and the pink ovals represent functional groups. At the start of the day, CCA1 and LHY, paralogs, heterodimerize and repress the activation of the PRR-encoding genes, TOC1, GI, and the evening complex (EC) members LUX, NOX, ELF3, and ELF4. PRR9, PRR7, and PRR5, are expression in succession and repress the expression of CCA1 and LHY, as well as their own transcription. Near the end of the light period, TOC1 represses PRR9, PRR7, PRR5, CCA1, and LHY, as well as GI, LUX, and ELF4. After, the EC maintains the repression of GI and represses PRR9 and PRR7. Transcriptional repression is the main feature of the core oscillator in plants, however, transcriptional activators have recently been identified within the oscillator. In the morning, the genes LWD1 and LWD2 promote expression of CCA1, PRR9, PRR7, and TOC1. Later in the day, PRR5, TOC1, and ELF4 expression is promoted by RVE8, LNK1, and LNK2. RVE8 also promotes GI and LUX expression. **C)** Stomata open up at the beginning of the dark period, allowing the uptake of CO_2. CO_2 is then converted into bicarbonate which is then assimilated into an organic acid by PEPC. The organic acid then is stored in the vacuole. As the light period beings, stomata begin to close, and the decarboxylation of the C_4 acid beings. The CO_2 released from decarboxylation is refixed by RUBISCO and assimilated via the Calvin cycle of photosynthesis. Pyruvate of PEP released from the decarboxylation event is converted into storage carbohydrates. The subsequent depletion of storage carbohydrates provides PEP for the continuation of CAM. Images A and B are redrawn images of Figures 1 and 2 in Nohales and Kay, *Nature Structural and Molecular Biology*, 2016, respectively. Image C is a redrawn image of Figure 1 in Borland et al., *Trends in Plant Science*, 2014.

Figure 1. Continued.



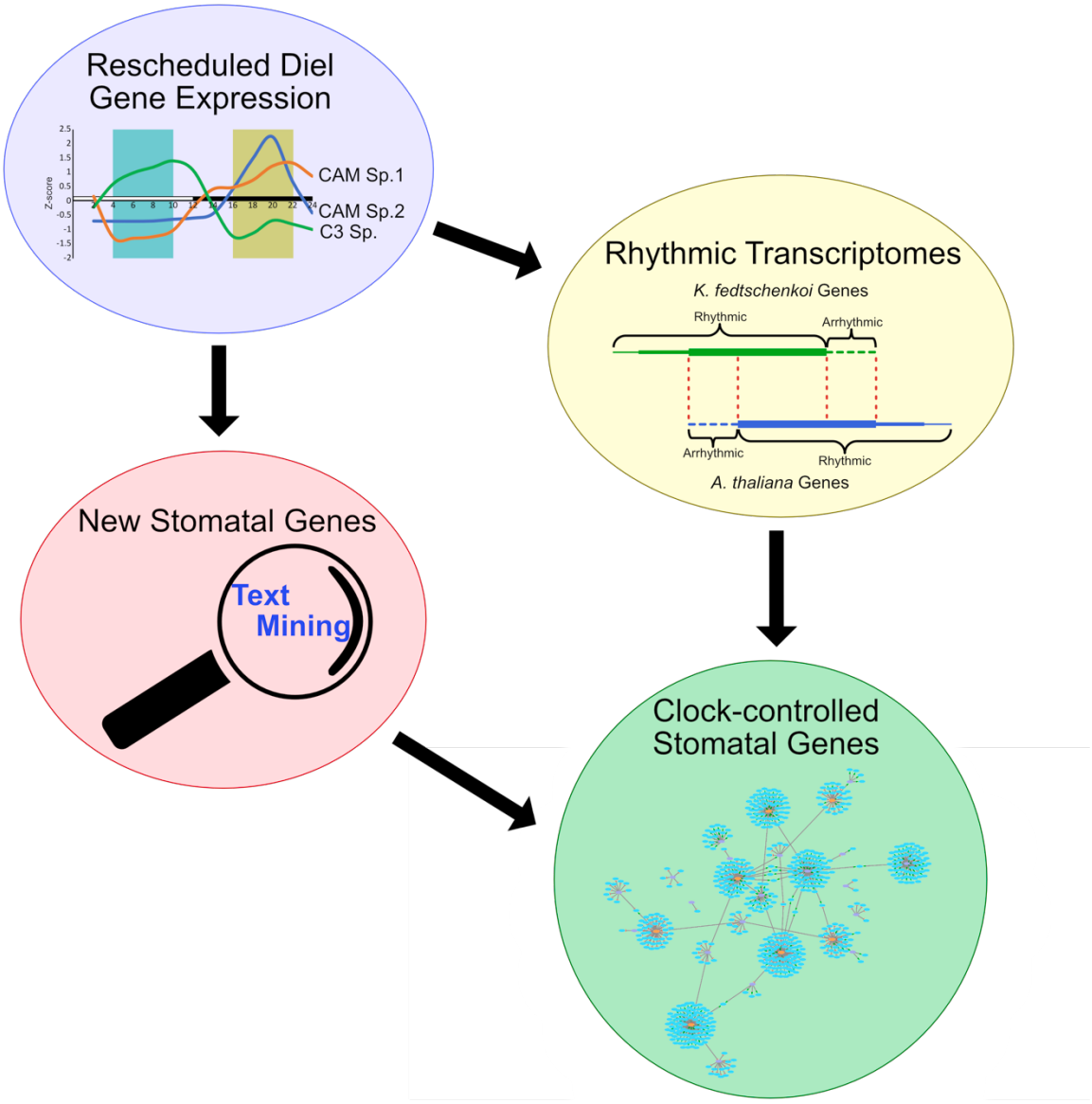


Figure 2. Overview of major themes within the dissertation.

CHAPTER I
INVESTIGATING DIEL RE-SCHEDULING OF GENE EXPRESSION BETWEEN CAM AND C_3 SPECIES

Part of the text, Figure 1, and Table 1 were taken from Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, Liu D, Weigel D, Cheol Yim W, Ha J, Heyduk K, Goodstein D, Guo H, **Moseley R**, et al. 2017. The Kalanchoë genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications* **8**: 1-15.

Moseley's contribution to this publication was comparative analysis of gene expression dynamics between three plant species to estimate temporal shifts between orthologous genes.

Abstract

Crassulacean acid metabolism (CAM) is an adaptation in plants that improves photosynthetic and water-use efficiency and has evolved independently many times in diverse lineages of flowering plants. It is hypothesized that convergent evolution of temporal gene expression underpins the independent emergences of CAM from C_3 photosynthesis. To test this hypothesis, we compared diel gene expression profiles among *Kalanchoë fedtschenkoi* (a CAM species in the dicot lineage), *Ananas comosus* (a CAM species in the monocot lineage) and *Arabidopsis thaliana* (a C_3 species in the dicot lineage). Our comparative analysis revealed signs of convergence in re-scheduling of diel expression of genes involved in nocturnal CO_2 fixation, stomatal movement, heat tolerance, circadian clock, and carbohydrate metabolism in the two CAM species in comparison with the C_3 species. The results presented in the study provide new insights into the evolution of CAM from C_3 photosynthesis.

Introduction

Crassulacean acid metabolism (CAM) is a metabolic adaptation in plants that occur in resource-limited, arid to semi-arid environments. This adaptation involves a carbon-concentrating mechanism that improves photosynthetic CO_2 fixation, as well as an inversion of day/night stomatal movement pattern, relative to C_3 plants, that enhances water-use efficiency and associated drought avoidance/tolerance (West-Eberhard et al. 2011). CAM photosynthesis is generally described in four phases but can be simplified into two phases: 1) nighttime uptake of atmospheric CO_2 through open stomata and subsequent primary fixation of CO_2 into malate through a few enzymatic reactions; and 2) daytime decarboxylation of malate and CO_2 refixation via C_3 photosynthesis, facilitated by Rubisco (Owen and Griffiths 2013; Borland et al. 2014). The division of the two phases is generally derived from peak storage of malate in the vacuole at dawn. CAM is found in over 400 genera across 36 families of vascular plants (Yang et al. 2015). The genesis of CAM is thought to be a result of multiple, independent events of evolution from diverse ancestral C_3 photosynthesis lineages (Silvera et al. 2010). In the lineages in which CAM evolved, the core biochemical features of the CAM pathway are similar, with some variation in the decarboxylase enzyme and in the storage carbohydrates that provide substrate for malate synthesis at night (Christopher and Holtum 1996; Holtum et al. 2005).

The goal of this study was to test the hypothesis that convergent evolution in temporal diel gene expression underpins the multiple and independent emergences of CAM from C_3 photosynthesis. A genome-wide screening for these signatures of convergent change was performed using the following criteria: the signatures are 1) isomorphic in the CAM genomes of distant groups and 2) dimorphic in related C_3 photosynthesis genomes. Various genomic resources were employed from one monocot CAM species, *Ananas comosus* (pineapple) (Ming et al. 2015); one dicot CAM species, *Kalanchoe fedtschenkoi*; and one dicot C_3 species, *Arabidopsis thaliana* (Initiative 2000; Mockler et al. 2007).

Methods

Estimation of time delay between gene expression profiles

The diurnal gene expression data with 2-h intervals for *A. comosus* were obtained from (Ming et al. 2015). The diurnal gene expression data with 4-h intervals for *A. thaliana* were obtained from (Mockler et al. 2007). The diurnal gene expression with 2-h intervals for *K. fedtschenkoi* were generated in this study. To calculate the time-delay between time series gene expression profiles, the diurnal expression data were adjusted to one-h interval time series data by interpolation using the cubic spline function in MATLAB (Mathworks, Inc.). The gene expression data for all three species were transformed to represent the z-score change relative to the respective gene's mean expression, where gene expression measurements at each time point represent the number of standard deviations away from the mean. Pairwise correlation was calculated between each pair of orthologous genes for all possible time delays using the circular cross correlation function (cxcorr) in MATLAB (Mathworks, Inc.). The cxcorr function produces a correlation coefficient between gene 1 and shifted, or lagged, copies of gene 2 as a function of the lag. With each correlation coefficient, a lag value was given. The lag values were then converted into hours, giving an estimate on time delay. The time delay at which the correlation was maximum was selected as the estimated delay between the two genes.

Defining convergent diel rescheduling of gene expression

The diurnal gene expression profiles were compared between CAM and C_3 species to identify convergent shifts of diurnal gene expression in CAM species relative to C_3 species. Exclusively, comparisons were performed of diurnal expression profiles of 9,733 ortholog groups of genes from *K. fedtschenkoi* (eudicot, CAM photosynthesis), *A. comosus* (monocot, CAM photosynthesis), and *A. thaliana* (eudicot, C_3 photosynthesis) with transcripts > 0.01 FPKM in mature leaf samples collected at six or more diel time points. *K. fedtschenkoi* genes were extracted that displayed a time shift of less than or equal to 3 hours between its own gene expression profile and the gene expression profile of the *A. comosus* ortholog, whereas the time shifts between the CAM species (*K. fedtschenkoi* and *A. comosus*) gene expression profiles and their *A. thaliana* ortholog expression profile were equal to or greater than 6 hours.

In conjunction with this time-delay criterion, two other criteria were employed to identify a *K. fedtschenkoi* gene that had undergone convergent evolution of gene expression: 1) its diel gene expression profile was highly correlated (Spearman's rank correlation coefficient, $r>0.8$) with those of at least one of the orthologs in *A. comosus*, but not highly correlated ($r<0.5$) with those of any of the orthologs in *A. thaliana* and 2) it displayed a significant difference (false discovery rate <0.01) in transcript abundance either between midday and midnight or between dawn and dusk (Figure 3A).

Results

Convergent evolution of genes involved various processes

Fifty-four *K. fedtschenkoi* genes were identified as candidates for involvement in the convergent shift in diurnal gene expression profiles specific to the two CAM species relative to *A. thaliana* (Table 1). Several genes involved in CO_2 fixation and stomatal movement were identified to have undergone a shift in expression in both CAM species, relative to their *A. thaliana* ortholog. The key enzyme for nocturnal CO_2 fixation in CAM species is phosphoenolpyruvate carboxylase (PEPC) and PEPC kinase (PPCK) is an essential protein kinase that regulates PEPC activity via phosphorylation in the dark in response to the circadian clock (Hartwell 2006; Borland et al. 2014; Yang et al. 2015). The gene expression patterns of the two most highly expressed PPCK genes in both CAM species were highly correlated, with only a 1.5-hour time shift between them, whereas both showed an ~11-hour time shift relative to their best matched ortholog in *A. thaliana* (Figure 3B). The time of max gene expression of the PPCK gene was shifted from daytime in *A. thaliana* to nighttime in the two CAM species (Figure 3B), which is consistent with PPCK activation of PEPC-mediated nighttime CO_2 fixation. Blue light signaling is a key environmental signal that regulates stomatal opening (Kinoshita et al. 2001) and the blue light photoreceptor, phototropin 2 (PHOT2), displayed a 1-hour shift in gene expression relative to its *A. comosus* ortholog's gene expression, whereas the *K. fedtschenkoi* and *A. comosus* orthologs' gene expression profiles displayed an 11-hour shift with their *A. thaliana* ortholog's gene expression profile (Figure 3C). The time of max gene expression shifted from dawn in *A. thaliana* to dusk in the two CAM species (Figure 3C) and suggest that PHOT2 might contribute to the inversion of stomatal movement seen in CAM species relative to stomatal movement in C_3 species. The convergent changes in diurnal gene expression profiles of PPCK and PHOT2 provide evidence for CAM evolving due to temporal rewiring of gene expression.

Being able to sense and respond to the daily fluctuating temperatures a plant faces is important for maintaining proper fitness of the plant (Yamori and Hikosaka 2014). Particular to CAM species, due to closed stomata during the majority of day (Dever et al. 2015), internal leaf temperatures can rise significantly (Krause et al. 2016), which can impair photosynthesis via thermal instability of Rubisco (Salvucci and Crafts-Brandner 2004). Two genes involved in temperature responses were found to have shifted gene expression in both CAM species relative to *A. thaliana*, a heat-shock protein (HSP70) and a low temperature-inducible gene (SRC2). HSP70 has been reported to function as nano-compartments in which Rubisco sub-units can fold in isolation, unimpaired by

aggregations (Lui et al. 2010). HSP70 only had a 1-hour difference in peak gene expression between orthologs in the two CAM species, whereas the *A. thaliana* ortholog was shifted ~10-hours in peak gene expression (Figure 3D). The HSP70 *A. thaliana* ortholog was phased to the evening and shifted to the morning in both CAM species (Figure 3D). SRC2 has been linked to NADPH oxidase (NOX)-mediated reactive oxygen species (ROS) production in response to cold temperatures (Kawarazaki et al. 2013), and ROS has been reported to be involved in expression of genes involved in drought tolerance (Maruyama et al. 2009). Both SRC2 genes in the two CAM species had peak gene expression at the same time of day, whereas the *A. thaliana* ortholog's peak gene expression was shifted by 10-hours (Figure 3E). These results suggest that these genes have undergone convergent changes in diurnal gene expression during the evolution of CAM.

Key physiological and biochemical characteristics of CAM, such as CO_2 fixation and PEPC phosphorylation, are well established as outputs of the circadian clock (Dever et al. 2015; Hartwell et al. 2016). Therefore, a key regulator of the diurnal re-scheduling of gene expression could be the circadian clock. Best hit orthologs of *A. thaliana* circadian clock genes in both CAM species did not display convergent changes in gene expression profiles.

Discussion

In this study, the genome-wide comparisons of CAM species and non-CAM species revealed convergent changes in the diurnal re-scheduling of gene expression. In the context of convergent re-scheduling of gene expression, 54 genes were identified based on the criteria defined in the results section. Several of these rewired genes are involved in CO_2 fixation, stomatal movement, and response to temperature, to name a few (Table 1). These results provide strong support for the hypothesis that convergent evolution in diurnal gene expression underpins the multiple and independent emergences of CAM from C_3 photosynthesis.

Differences in gene expression within and between species are caused by mutations in *cis*-regulatory elements (e.g., promoters, enhancers) and *trans*-regulatory factors (e.g., transcription factors, noncoding RNAs) (Carrol 2008; Stern and Orgogozo 2008). For instance, the binding of a transcription factor (TF) to its associated *cis*-regulatory element located in the promoter of a gene can alter expression of the respective gene. It was proposed that rewiring of gene expression of circadian clock genes could have played a role in the convergent change of gene expression in CAM species as a majority of the known circadian clock genes are TFs. However, none of the genes that were orthologous to *A. thaliana* clock genes in both CAM species were found to have temporally altered gene expression in comparison with C_3 species. However, this does not rule out several possibilities of the circadian clock still playing a role in the convergent change of gene expression in CAM species. For example, three scenarios that could be at play are 1) Genes regulated by certain circadian clock genes in *A. thaliana* could be regulated by a different circadian clock gene in *K. fedtschenkoi*, therefore causing a temporal shift in gene expression, 2) genes that are not involved in circadian rhythms in *A. thaliana* could have taken on this function in *K. fedtschenkoi*, and therefore

entrained the expression of genes they regulated and 3) there are unknown circadian clock genes between *A. thaliana* and *K. fedtschenkoi* that could have temporally different gene expression profiles in comparison with *A. thaliana*.

Conclusion

Convergent changes in the diel re-scheduling of gene expression contributes to the convergent evolution of CAM from C_3 photosynthesis. With regards to the mechanism behind the convergent change in gene expression, the circadian clock was proposed as a key regulator, however, no temporal rewiring was seen in circadian clock gene expression profiles. Several possibilities where presented were the circadian clock could still be implicated in the evolution of CAM. Further studies in *K. fedtschenkoi* are needed to determine what or if any of these scenarios could have taken place. Investigating the rhythmic transcriptome of *K. fedtschenkoi* and comparing it the rhythmic transcriptome of *A. thaliana* could provide insights on the conserved and diverged regulation of plant physiology by the clock between species. Additionally, investigating the regulatory network of the circadian clock in *K. fedtschenkoi* could help better understand how the clock regulates biological process in *K. fedtschenkoi*.

References

Borland AM, Hartwell J, Weston DJ, Schlauch KA, Tschaplinski TJ, Tuskan GA, Yang X, Cushman JC. 2014. Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci* **19**: 327-338.

Carrol S. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**: 25-36.

Christopher J, Holtum J. 1996. Patterns of carbohydrate partitioning in the leaves of crassulacean acid metabolism species during deacidification. *Plant Physiol* **112**: 393-399.

Dever LV, Boxall SF, Knerova J, Hartwell J. 2015. Transgenic perturbation of the decarboxylation phase of Crassulacean acid metabolism alters physiology and metabolism but has only a small effect on growth. *Plant Physiol* **167**: 44-59.

Hartwell J. 2006. The circadian clock in CAM plants. *Endogenous Plant Rhythms*: 211-236.

Hartwell J, Dever LV, Boxall SF. 2016. Emerging model systems for functional genomics analysis of Crassulacean acid metabolism. *Curr Opin Plant Biol* **31**: 100-108.

Holtum JA, Smith JAC, Neuhaus HE. 2005. Intracellular transport and pathways of carbon flow in plants with crassulacean acid metabolism. *Functional Plant Biology* **32**: 429-449.

Initiative TAG. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**.

Kawarazaki T, Kimura S, Lizuka A, Hanamata S, Nibori H, Michikawa M, Imai A, Abe M, Kaya H, Kuchitsu K. 2013. A low temperature-inducible protein AtSRC2 enhances the ROS-producing activity of NADPH oxidase AtRbohF. *Molecular Cell Research* **1833**: 2775-2780.

Kinoshita T, Doi M, Suetsugu N, Kagawa T, Wada M, Shimazaki K. 2001. phot1 and phot2 mediate blue light regulation of stomatal opening. *Nature* **414**: 656-660.

Krause H, Winter K, Krause B, Virgo A. 2016. Protection by light against heat stress in leaves of tropical crassulacean acid metabolism plants containing high acid levels. *Plant Biology* **43**: 1061-1069.

Lui C, Young A, Starling-Windhot A, Bracher A, Saschenbrecker S, Rai B, Rai K, Berninghausen O, Mielke T, Hartl F et al. 2010. Coupled chaperone action in folding and assembly of hexadecameric Rubisco. *Nature* **463**: 197-202.

Maruyama K, Takeda M, Kidokoro S, Yamada K, Sakuma Y, Urano K, Fujita M, Yoshiwara K, Matsukura S, Morishita Y et al. 2009. Metabolic Pathways Involved in Cold Acclimation Identified by Integrated Analysis of Metabolites and Transcripts Regulated by DREB1A and DREB2A. *Plant Physiol* **150**: 1972-1980.

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* **47**: 1435-1442.

Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J. 2007. The Diurnal Project: Diurnal and Circadian Expression

Profiling, Model-based Pattern Matching, and Promoter Analysis. *Cold Spring Harbor Symposia on Quantitative Biology* **72**: 353-363.
Owen NA, Griffiths H. 2013. A system dynamics model integrating physiology and biochemical regulation predicts extent of crassulacean acid metabolism (CAM) phases. *New Phytol* **200**: 1116-1131.
Salvucci M, Crafts-Brandner SJ. 2004. Mechanism for deactivation of Rubisco under moderate heat stress. *Physiol Plant* **122**: 513-519.
Silvera K, Neubig K, Whitten W, Williams N, Winter K, Cushman JC. 2010. Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology* **37**: 995-1010.
Stern D, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? *Evolution* **62**: 2155-2177.
West-Eberhard MJ, Smith JA, Winter K. 2011. Plant science. Photosynthesis, reorganized. *Science* **332**: 311-312.
Yamori W, Hikosaka K. 2014. Temperature response of photosynthesis in C3, C4, and CAM plants: temperature acclimation and temperature adaptation. *Photosynthesis Research* **119**: 101-117.
Yang X, Cushman JC, Borland AM, Edwards EJ, Wullschleger SD, Tuskan GA, Owen NA, Griffiths H, Smith JA, de Paoli HC et al. 2015. A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist* **207**: 491-504.

Appendix

Table 1. Time-shifts in diel gene expression profiles between orthologs in *Kalanchoë fedtschenkoi* (Kafe), *Ananas comosus* (Anco), and *Arabidopsis thaliana* (Arth).

Only the gene pairs containing *K. fedtschenkoi* and *A. comosus* genes that showed convergent changes in diel expression profile were shown here. "Xcorr" is the cross correlation. "Time" is time-shift in hours.

Name	*Kalanchoë*	Pineapple	*Arabidopsis*	Kafe vs. Arth		Anco vs. Arth		Kafe vs. Anco	
				Time	Xcorr	Time	Xcorr	Time	Xcorr
PHOT2	Kaladp0033s0113	Aco014242	AT5G58140	11	0.58	9	0.63	1	0.87
PPR	Kaladp0058s0382	Aco019522	AT1G02060	12	0.67	11	0.72	0	0.91
DAG	Kaladp0130s0003	Aco002485	AT3G06790	11	0.82	11	0.9	0	0.92
TPX2	Kaladp0062s0086	Aco009655	AT3G23090	9	0.88	9	0.85	1	0.9
MRI1	Kaladp0053s0416	Aco007956	AT2G05830	11	0.87	11	0.72	0	0.78
NUTF2	Kaladp0040s0332	Aco015511	AT2G46100	7	0.84	6	0.79	1	0.88
bcrC	Kaladp0840s0040	Aco007440	AT2G34670	8	0.69	8	0.7	0	0.91
	Kaladp0045s0260	Aco004226	AT2G38820	10	0.9	11	0.94	1	0.85
PLT1	Kaladp0040s0745	Aco000080	AT2G16120	7	0.77	8	0.5	1	0.84
EXS	Kaladp0008s0539	Aco017367	AT1G55610	9	0.79	9	0.73	1	0.88
HSP70	Kaladp0060s0296	Aco031458	AT5G02500	10	0.64	9	0.71	1	0.8
CSLC12	Kaladp0037s0421	Aco011603	AT4G07960	10	0.85	9	0.89	1	0.9
PLC2	Kaladp0059s0034	Aco001556	AT3G08510	9	0.93	8	0.81	0	0.85
Lycopene cyclase protein	Kaladp0056s0132	Aco024665	AT3G10230	9	0.83	10	0.8	0	0.85
CDKI7	Kaladp0028s0063	Aco011817	AT1G49620	10	0.83	8	0.68	0	0.86
NPF	Kaladp0033s0087	Aco013262	AT2G26690	11	0.93	8	0.74	2	0.83
TIM23	Kaladp1244s0001	Aco007546	AT1G72750	6	0.92	7	0.86	1	0.83
IAA29	Kaladp0048s0752	Aco030721	AT4G32280	7	0.86	7	0.69	0	0.87
PRN12	Kaladp0026s0118	Aco015046	AT5G64760	9	0.84	12	0.66	1	0.68
PDK	Kaladp0068s0282	Aco017179	AT3G06483	10	0.91	11	0.88	1	0.76
OMP	Kaladp0040s0029	Aco018518	AT3G52420	9	0.76	12	0.86	2	0.82
TMEM245	Kaladp0048s0390	Aco003087	AT5G55960	12	0.81	12	0.73	1	0.81
TMEM184C	Kaladp1262s0005	Aco014396	AT1G11200	9	0.64	7	0.76	0	0.66
ACLB2	Kaladp0045s0074	Aco001770	AT3G06650	8	0.77	6	0.75	1	0.85
NPH3 family	Kaladp0040s0264	Aco016939	AT5G67440	10	0.95	9	0.77	0	0.78
SS	Kaladp0043s0207	Aco001151	AT2G21940	12	0.83	11	0.87	1	0.9
Ash2l	Kaladp0045s0419	Aco000634	AT1G51450	11	0.78	11	0.95	1	0.88
WRC	Kaladp0011s0273	Aco020634	AT2G42040	8	0.58	8	0.74	1	0.9
SNRPC1	Kaladp0349s0001	Aco016127	AT4G03120	8	0.87	11	0.88	2	0.89
COL2	Kaladp0039s0496	Aco002998	AT1G68520	11	0.73	11	0.94	1	0.85
GK	Kaladp0037s0359	Aco011530	AT1G80460	9	0.81	9	0.87	0	0.92
FH	Kaladp0040s0015	Aco002502	AT5G50950	8	0.93	9	0.87	1	0.87
PIP1-2	Kaladp0059s0048	Aco010251	AT4G23400	9	0.8	8	0.74	0	0.85
AIM1	Kaladp0022s0177	Aco010785	AT4G29010	8	0.86	7	0.82	1	0.92
HAD	Kaladp0071s0244	Aco001521	AT3G45740	11	0.75	12	0.76	0	0.82
SURE	Kaladp0427s0023	Aco023553	AT4G14930	7	0.52	8	0.74	0	0.86
EDR2L	Kaladp0098s0203	Aco027425	AT5G10750	9	0.87	12	0.74	2	0.83
CYB2	Kaladp0035s0007	Aco006777	AT5G38630	10	0.55	12	0.66	1	0.93
SEC14	Kaladp0067s0305	Aco003265	AT1G72160	11	0.73	11	0.64	1	0.88
Gdap2	Kaladp0098s0113	Aco002356	AT3G10210	9	0.56	9	0.62	1	0.84
FAH	Kaladp0060s0413	Aco015620	AT1G12050	9	0.6	10	0.47	0	0.82
	Kaladp0081s0088	Aco014627	AT1G04650	10	0.52	9	0.79	0	0.71
UFC1	Kaladp0095s0421	Aco014085	AT1G27530	9	0.89	9	0.85	1	0.77
ABCG15	Kaladp0045s0418	Aco007271	AT1G17840	11	0.8	10	0.6	1	0.81
	Kaladp0034s0049	Aco005409	AT1G23760	8	0.95	7	0.88	1	0.9
SAG	Kaladp0808s0026	Aco022311	AT2G44670	11	0.89	10	0.85	1	0.97
PPCK1	Kaladp0037s0517	Aco013938	AT1G08650	11	0.84	12	0.94	2	0.81
Starch synthase	Kaladp0055s0317	Aco010848	AT5G24300	9	0.76	7	0.73	3	0.85
SNRPC	Kaladp0018s0298	Aco007284	AT1G09070	10	0.88	10	0.79	0	0.89
Leucine rich repeat	Kaladp1251s0003	Aco008844	AT3G47090	7	0.9	7	0.74	2	0.8
CHR5	Kaladp0011s0810	Aco001980	AT2G13370	6	0.78	6	0.53	0	0.86
Leucine rich repeat	Kaladp0054s0031	Aco008313	AT2G07040	8	0.78	9	0.51	0	0.72
CINV1	Kaladp0550s0020	Aco007782	AT5G22510	11	0.83	11	0.88	1	0.79
TPS	Kaladp0011s0363	Aco012107	AT1G06410	6	0.7	6	0.56	0	0.84

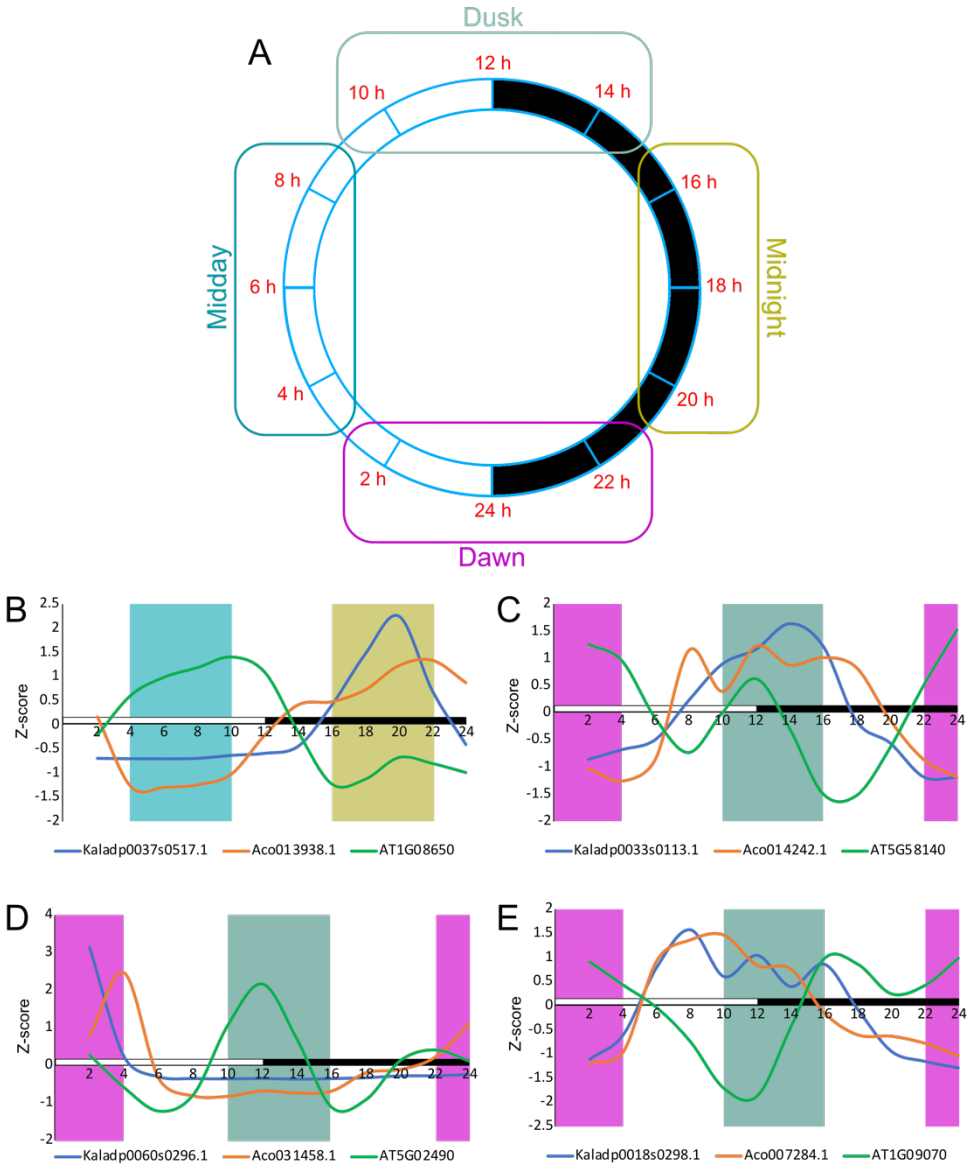


Figure 3. Examples of convergent change in diel expression profile in CAM species.

A) The four time-windows for comparative analysis of temporal changes in gene expression, which were represented by 12 time-points, 2, 4, …, 24 hours after the start of the light period. **B-E)** Comparison of diel gene expression profiles of phosphoenolpyruvate carboxylase kinase 1 (PPCK1), phototropin 2 (PHOT2), heat shock protein 70 (HSP70), and soybean gene regulated by cold 2 (SRC2), respectively, between CAM species (*Kalanchoë fedtschenkoi* and *Ananas comosus*) and C_3 species (*Arabidopsis thaliana*). White and black bars indicate daytime (12-h) and nighttime (12-h), respectively. X-axis represents the time after the start of the light period.

CHAPTER II
CONSERVATION AND DIVERSIFICATION OF CIRCADIAN RHYTHMICITY BETWEEN CAM AND C_3 PHOTOSYNTHESIS PLANTS

This chapter is derived from a manuscript currently in preparation: **Moseley R**, Mewalal R, Motta F, Haase S, Tuskan G, Yang X. Conservation and diversification of circadian rhythmicity between CAM and C_3 photosynthesis plants**.**

All work was conducted by Robert C. Moseley as well the writing of the manuscript. The remaining authors provided guidance and feedback on data analysis and manuscript preparation.

Abstract

Crassulacean acid metabolism (CAM) improves water-use and photosynthetic efficiency. It has been hypothesized that CAM is under the control of the internal circadian clock. However, the role that the circadian clock plays in the evolution of CAM is not well understood. The goal of this research is to investigate the extent of circadian control over the CAM physiology and its impact on CAM evolution. To achieve this goal, rhythmic gene sets were identified in a CAM model plant species (*Kalanchoë fedtschenkoi*) and a C_3 model plant species (*Arabidopsis thaliana*) through analysis of diel time-course gene expression data using multiple periodicity detection algorithms. From the protein sequences, ortholog groups were constructed containing genes from each of these two species. The ortholog groups were categorized into five gene sets based on conservation and diversification of rhythmic gene expression between the two plant species. Minimal functional overlap was observed when comparing the rhythmic gene sets of each species. Metabolic processes were found to be the dominant processes under circadian control in *K. fedtschenkoi*, and numerous genes were found to have retained or gained rhythmic expression in *K. fedtsechenkoi*. Additonally, several rhythmic orthologs, including CAM-related orthologs, displayed phase shifts between species. Results of this analysis point to several mechanisms by which the circadian clock could have played a role in the evolution of CAM and provide testable hypotheses for future experiments.

Introduction

All organisms confront environmental fluctuations on a daily basis, including changes in light, temperature, predation risk and water availability. Many of these changes exhibit predictable diurnal oscillations, which allows organisms to maintain an internal timekeeping system in anticipation of environmental changes. This timekeeping system is referred to as the circadian clock and enables organisms to synchronize their biological activity with the external environment (Bell-Pederson et al. 2005). The circadian clock provides for time-of-day specific biology and is very important for plant survival, due to their sessile life. For instance, plants are unable to retreat from herbivores that pose a threat. Instead, defense pathways have become subordinate to the circadian clock, allowing the plant to activate defense pathways to the time of day when the threat posed by herbivores is maximal (Greenham and McClung 2015). Numerous processes have been documented to be under the control of the circadian clock; however, the extent to which the clock controls plant physiology is still not well understood (Piechulla 1988; Merida et al. 1999; Thain et al. 2000; Hotta et al. 2007; Mallona et al. 2011; McClung 2013; Ming et al. 2015).

Crassulacean acid metabolism (CAM) represents an interesting area of study on circadian control over physiology. CAM is an adaptation to water-limited environments based on a temporal separation of CO_2 fixation and inverted day/night stomatal movement pattern, relative to C_3 plants. These adaptations result in improved photosynthetic and water-use efficiency. It is hypothesized that the strict temporal control of CAM processes is maintained by the circadian clock in CAM plants (Wyka et al. 2004; Boxall et al. 2005; Hartwell 2005; Hartwell 2006; von Caemmerer and Griffiths 2009; Ming et al. 2015). However, a thorough genomic investigation of circadian control in CAM plants has not been performed.

Identification and characterization of circadian-clock regulated genes requires measuring gene expression activity over a 24-hour period (Bar-Joseph et al. 2012; Li et al. 2015; Hughes et al. 2017). Time-course data, of at least 24 hours with 2- to 4-hour sampling, has made the investigation of the control of the circadian clock over an organism's physiology more tractable (Mockler et al. 2007; Michael et al. 2008; Filichkin et al. 2011; Higashi et al. 2016; Koda et al. 2017; Sharma et al. 2017; Wai et al. 2017). Numerous algorithms exist for detecting rhythmic expression within such data sets; however, different algorithms can give inconsistent results when used on the same data set (Keegan et al. 2007; Doherty and Kay 2010; Deckard et al. 2013). Recently, a new approach for detection of rhythmic expression in time-course data developed for integrating multiple rhythmic detection algorithms to produce a robust set of rhythmic genes (Kelliher et al. 2016). Applying a similar method to produce robust, comprehensive rhythmic gene lists in C_3 and CAM plants would provide insights into how the circadian clock propagates environmental signals across the two different photosynthetic types as well as provide clues to how CAM evolved from C_3.

The aim of this study is to investigate the conservation and diversification of rhythmic gene expression between a model C_3 plant *Arabidopsis thaliana* and a model CAM plant *Kalanchoë fedtschenkoi*. We applied community-accepted rhythmic detection algorithms to plant time-course data and identified commonalities and differences in rhythmic genes between *A. thaliana* and *K. fedtschenkoi*, providing a better understanding of how the circadian clock controls gene expression along with the biological processes subordinate to it, in CAM plant species.

Results

Identifying rhythmic genes in A. thaliana and K. fedtschenkoi

To compare putative clock-controlled genes between *A. thaliana* and *K. fedtschenkoi*, genes that display significant rhythmic expression were identified in each species. The periodicity detection algorithms 1) de Lichtenberg (de Lichtenberg et al. 2005), 2) JTK-CYCLE (JTK) (Hughes et al. 2010), and 3) Lomb-Scargle (Scargle 1982) were applied to *A. thaliana* and *K. fedtschenkoi* 24-h time-course datasets to investigate rhythmicity of gene expression in each of these two species. Due to the fact that each algorithm can rank genes differently based on alternate preferences for periodic curve features (Deckard et al. 2013), the rankings given by each algorithm were summed and all genes were ranked by cumulative scores (Tables S1 and S2). There was no clear break

in the distribution of periodicity scores, indicating that no clear "threshold" exists between rhythmic and arrhythmic genes. By visual inspection we estimated that 10,000 and 10,500 genes appeared to have rhythmic expression in *A. thaliana* and *K. fedtschenkoi*, respectively (Figure S1). Comparing the *A. thaliana* rhythmic gene set with previously published *A. thaliana* rhythmic gene sets revealed a 48% overlap (Mockler et al. 2007). This overlap is exceptional, previous studies comparing several rhythmic gene datasets in *Drosophila melanogaster* found an average overlap of 11%, due to differences in periodicity detection algorithms and underlying variation in datasets (Keegan et al. 2007). Next, a score cutoff was applied to each list of rhythmic genes using the JTK algorithm (Hughes et al. 2010). Based on this cut-off, 9,338 genes in *A. thaliana* and 8,769 genes in *K. fedtschenkoi* were estimated to be rhythmically expressed with a 24-h period (Figure 4), representing 34% of the genes in the *A. thaliana* genome and 28% of the genes in the *K. fedtschenkoi* genome. The expression patterns in Figure 4 for rhythmic genes reveal a continuum of phase-specific gene expression, which is a typical behavior of periodic genes controlled by the circadian clock in *A. thaliana* (Covington et al. 2008), tomato (Higashi et al. 2016), and *Brachypodium distachyon* (Koda et al. 2017). Multiple criteria for evaluating expression patterns for individual genes in *A. thaliana* and *K. fedtschenkoi* in our study are provided in Tables S1 and S2.

Functional comparison of rhythmic genes between A. thaliana and K. fedtschenkoi

Determining the dominant processes displaying significant rhythmic expression in their associated genes can give insight into the physiological relevance of the circadian clock. Therefore, a functional enrichment of both rhythmic gene sets was performed. Top functional groups enriched in the *A. thaliana* rhythmic gene set include pigment biosynthetic processes, ribonucleoprotein complex biogenesis, and response to high light intensity, red or far red light, light stimulus and cold (Figure 5A). The *K. fedtschenkoi* rhythmic gene set was enriched for organic acid metabolic processes, carboxylic acid metabolic processes, protein targeting to chloroplast, starch metabolic processes, oxylipin metabolic processes, phosphate-containing compound metabolic processes, photosynthesis (light harvesting), glyceraldehyde-3-phosphate metabolic process, isoprenoid biosynthetic process, and plastid organization (Figure 5B). It was striking that no overlap in enriched functional groups was observed when comparing the dominate rhythmic processes between these two species.

To investigate the physiological relevance of the circadian clock in a temporal context, a chronological gene co-expression network (GCN) was constructed from each rhythmic dataset, with subnetworks representing each phase of the circadian system (Figure 6A-B). Spearman's rank correlation coefficient was calculated for all pair-wise combinations of the rhythmic genes in each dataset. The *A. thaliana* and *K. fedtschenkoi* GCN contained 9,338 rhythmic genes linked by 4,174,924 edges and 8,769 rhythmic genes linked by 4,590,438 edges, respectively. with a $r \geq 0.8$ as a cutoff threshold of co-expression. Visualizing phase calls of every gene with a color scheme, rhythmic genes can be seen chronologically connected in a circular network, consistent with the diagonal stripe observed in the rhythmic gene expression heat maps for each species (Figure 4), as well as in a previous study in *B. distachyon* (Koda et al. 2017). The rhythmic expression

of genes is typically controlled by transcription factors (TFs) that are also rhythmically expressed at the transcript level (Koda et al. 2017). In the rhythmic gene sets for *A thaliana* and *K. fedtschenkoi*, 644 and 923 genes were found to encode putative TFs, respectively. Rhythmic TFs in both species were found to be in all phases of the day and night (Figure 6C-D). Comparing overlapping biological functions between species, only photosynthesis and starch metabolism were common to both species (Figure 6E).

To dissect the conservation of rhythmicity and phase of expression between *A. thaliana* and *K. fedtschenkoi*, ortholog groups were constructed using the protein sequences from these two species. Between the rhythmic gene sets, five types of orthologs and genes were identified: orthologs that are rhythmic in both species (Type 1), orthologs that are rhythmic only in *A. thaliana* (Type 2), orthologs that are rhythmic only in *K. fedtschenkoi* (Type 3), genes found to be rhythmic only in *A. thaliana* (Type 4), and genes that are rhythmic only in *K. fedtschenkoi* (Type 5) (Figure 7). Gene list for each species in each ortholog group type can be found in Table S3. Gene numbers and functional investigations in each type are described below.

Rhythmic ortholog groups shared between A. thaliana and K. fedtschenkoi (Type 1)

2,641 ortholog groups were identified to contain at least one rhythmic gene in each of the two plant species (*A. thaliana* and one *K. fedtschenkoi*), which contained 3,770 and 3,755 genes, respectively. Gene ontology (GO) enrichment of each subset of rhythmic genes displayed a 37% overlap in enriched biological processes (Figure 7), such as protein phosphorylation, organic acid metabolic process, transport, photosynthesis, and generation of precursor metabolites and energy (Table S4). Re-scheduling of gene expression is thought to play a role in the evolution of CAM, however, the extent to which the circadian clock was a factor is not known. Therefore, phase shifts between rhythmic orthologs in these two species were examined. Phase shifts are defined here as the difference between two rhythmic genes' phase calls, which is the time point where maximum transcript abundance occurs for a gene. For simplicity, we will define two orthologs to be *synchronized* if their phase calls are within four hours of each other and *unsynchronized* otherwise.

Several key CAM genes involved in CO_2 fixation in *K. fedtschenkoi* (Yang et al. 2017) were found rhythmic, as well as several of their *A. thaliana* orthologs (Table 2 and Table 3). However, examining the synchronization between the orthologs proved difficult as ortholog groups contained many-to-many relationships. Evidence of both synchronization and desynchronization were found in one-to-one ortholog groups, such as seen with the NADP-malic enzyme (Table 3).

As mentioned earlier, cross-species identification of rhythmic orthologs can be challenging due to many-to-many orthologous relationships and when a set of genes for an ortholog group show differential rhythmicity. Therefore, further analysis of conserved rhythmicity within this group was limited to only those ortholog groups displaying a single expression profile within a species. This restriction resulted in 768 ortholog groups between the two species where both species' genes were rhythmic. Rhythmic genes were observed in all phases of the day and night that were sampled in each species (Figure 8A).

Two sub-types of ortholog groups were identified based on synchronization of gene expression between orthologous gene pairs. Of the 768 ortholog groups, 386 (50%) ortholog groups contained ortholog pairs with synchronized gene expression (Figure 8B). A large number of synchronized ortholog pairs were found phased between morning and evening (Figure 8C). The second sub-type, unsynchronized orthologous gene pairs, contained 382 (50%) ortholog groups (Figure 8). Two groups of orthologous gene pairs were found to have either their *A. thaliana* gene phased to the morning and the *K. fedtschenkoi* ortholog phased to the evening or their *A. thaliana* gene phased to early night and the *K. fedtschenkoi* ortholog phased to midday (Figure 8C), suggesting differential regulation of at least two groups of rhythmic orthologous gene pairs between *A. thaliana* and *K. fedtschenkoi*.

To investigate the potential mechanism underlying the variation in phase calls between orthologs of *A. thaliana* and *K. fedtschenkoi*, TFs were identified in the 768 ortholog groups. Within these ortholog groups, 61 contained genes in *A. thaliana* and *K. fedtschenkoi* annotated as TFs. TFs in both species were found in all phases of the day and night that were sampled in each species (Figure 8D). Of the 61 TF ortholog groups, 40 (66%) contained unsynchronized TF ortholog pairs (Figure 8E). The 61 orthologous TFs in *A. thaliana* and *K. fedtschenkoi* are listed in Table 4, along with the respective *A. thaliana* gene annotation. Several orthologous TFs had similar phase shifts as the two groups of orthologs described earlier (Figure 8F), potentially accounting for the difference in gene expression between orthologous gene pairs.

Rhythmic A. thaliana genes with arrhythmic K. fedtschenkoi orthologs (Type 2)

2,797 ortholog groups were identified that had at least one *A. thaliana* gene that was rhythmic and all *K. fedtschenkoi* genes were arrhythmic. A total of 3,275 *A. thaliana* genes found to be rhythmic that had no rhythmic *K. fedtschenkoi* orthologs (Figure 7). Only two biological processes were enriched in the *A. thaliana* gene set, cellular nitrogen compound metabolic process and response to endoplasmic reticulum stress (Table S4). To further examine differential rhythmicity within Type 2 ortholog groups, ortholog groups that also contained arrhythmic *A. thaliana* gene were filtered out, resulting in 2,258 ortholog groups that only contained rhythmic *A. thaliana* and arrhythmic *K. fedtschenkoi* genes. Clustering of the arrhythmic *K. fedtschenkoi* genes did not display a circular network typical of rhythmic genes, further validating the rhythmic gene expression detection method used in this study (Figure 9A). A slightly higher number of rhythmic *A. thaliana* genes were found with phase calls during the night (Figure 9B). Within these 2,258 ortholog groups, 178 rhythmic *A. thaliana* TFs were identified and were phased to all phases of the day and night (Figure 9B).

Rhythmic K. fedtschenkoi genes with arrhythmic A. thaliana orthologs (Type 3)

2,140 ortholog groups were identified that had at least one rhythmic *K. fedtschenkoi* gene whereas all *A. thaliana* genes were arrhythmic. A total of 2,500 *K. fedtschenkoi* rhythmic genes were found to have no rhythmic *A. thaliana* orthologs in their respective ortholog groups (Figure 7). Numerous processes were enriched in the *K. fedtschenkoi* gene set, e.g., several metabolic processes and transcription from plastids

(Table S4). Differential rhythmicity was further examined and resulted in 1,759 ortholog groups that only contained rhythmic *K. fedtschenkoi* and arrhythmic *A. thaliana* genes. Arrhythmic *A. thaliana* genes did not cluster into a circular pattern as their rhythmic *K. fedtschenkoi* orthologs did (Figure 9C). Plotting the distribution of phase calls of the rhythmic *K. fedtschenkoi* genes revealed a concentration of genes around midday (Figure 9D). Within this set of ortholog groups, 233 *K. fedtschenkoi* TFs were identified and were phased to all phases of the day and night and displayed a bimodal distribution of phase calls (Figure 9D).

Rhythmic genes only in A. thaliana (Type 4)

788 ortholog groups were identified that had at least one rhythmic *A. thaliana* gene, totaling 2,293 rhythmic genes in *A. thaliana*, with no orthologs in *K. fedtschenkoi* (Figure 7). Only toxin metabolic process was enriched in this gene set (Table S4).

Rhythmic genes only in K. fedtschenkoi (Type 5)

1,406 ortholog groups were identified that had at least one rhythmic *K. fedtschenkoi* gene, totaling 2,514 rhythmic genes in *K. fedtschenkoi* with no orthologs in *A. thaliana* (Figure 7). Only monoterpenoid biosynthetic process was enriched in the gene set, however, several *K. fedtschenkoi* had no GO annotations (Table S4). To determine if any genes in this gene set could be CAM-specific or *K. fedtschenkoi*-specific, protein sequences for each gene was BLAST against a collection of plant proteomes spanning monocots and dicots and including all photosynthetic types (i.e., C_3, C_4, and CAM) (Table S5). A relaxed (1e-5) and a strict (1e-10) E-value cutoff were used and results for both are presented in Table S6. Further analysis was conducted on BLAST results using the E-value cutoff of 1e-10. 1,486 *K. fedtschenkoi* genes in this gene set were found in both non-CAM and CAM species. Four *K. fedtschenkoi* genes were found to have homologs in only other CAM species and 443 *K. fedtschenkoi* genes were found to have homologs in only the non-CAM species. Of the four genes found in only CAM species, none were found in all the six CAM species tested. Finally, 581 genes were found to be *K. fedtschenkoi*-specific. A majority of the rhythmic *K. fedtschenkoi*-specific genes had phase calls during midday, around eight hours after light (Table S7), which coincided with the four CAM-specific *K. fedtschenkoi* genes (Table S8).

Core clock genes in A. thaliana and K. fedtschenkoi

The core clock network is generally the driving force behind rhythmic gene expression. Commonly, the *A. thaliana* core clock model (Nohales and Kay 2016) is used to infer conservation of clock genes in other plant species, however, variations in the core clock between species has been observed due to fractionation and/or loss of genes during speciation events (Lou et al. 2012). *K. fedtschenkoi* is no exception to this as variation in the core clock network in terms of gene copy numbers was observed (Table 5). Further variation in the *K. fedtschenkoi* core clock network was observed when assaying each component's expression dynamics. Core clock genes are typically characterized by cardinal circadian parameters, such as high amplitudes and fold-changes and being highly statistically rhythmic (Doherty and Kay 2010; Zhang et al. 2014; Hughes et al. 2017).

Defining high amplitudes as absolute amplitudes >10, fold-changes >2, and statistically rhythmic as JTK p-values <0.05, only 10 of the 23 *K. fedtschenkoi* core clock orthologs met these criteria (Table 5). There were paralogs of RVE6 and ELF4 among the 10 core clock genes, which totals to 8 unique core clock genes. The expression patterns of these 8 core clock genes in *K. fedtschenkoi* were synchronized with their *A. thaliana* orthologs (Figure 10). Of the remaining 13 core clock orthologs, 4 did not pass the gene expression threshold (max FPKM >1) used for before the analysis of gene expression rhythmicity (Table 5). Constructing the core clock network in *K. fedtschenkoi* with this information presents a different architecture of the clock (Figure 11).

Discussion

To date, the majority of studies investigating the control of the circadian clock in plants have been directed towards plant species that conduct C_3 photosynthesis. Despite the reduced focus, CAM features one of the best characterized physiological rhythms in plants and exemplifies an interesting case of circadian clock control (Boxall et al. 2005; Cushman et al. 2008). The aim of this study was to investigate differences and similarities of rhythmicity in gene expression between the model C_3 species *A. thaliana* and the model obligate CAM species *K. fedtschenkoi*. A new approach for detecting rhythmic gene expression was applied to time-course gene expression datasets from both species. Our analyses revealed both conservation and diversification of rhythmic gene expression between species. Additionally, most CAM-related genes were found to be rhythmically expressed in *K. fedtschenkoi* and displayed phase shifts and differential entrainment compared with their *A. thaliana* orthologs, providing another line of evidence that CAM is under circadian control. Moreover, differences were found between the *K. fedtschenkoi* core clock network and the *A. thaliana* model core clock network (Table 5).

Around 30% of all genes were found to be rhythmically expressed in both species, which is consistent with previous circadian studies in other plants species grown under similar conditions and using whole-leaf samples (Filichkin et al. 2011). Interestingly, there was no overlap in top functional groups enriched when comparing the entire rhythmic gene sets. Phase-specific enrichment of functional groups between the two species showed similar exclusiveness, except for starch metabolism and photosynthesis. These two functional groups were enriched around midday in both species, consistent with findings in other studies examining rhythmic processes in other plant species (Filichkin et al. 2011; Koda et al. 2017). The circadian clock is known to play a crucial role in photosynthesis and in regulation of carbohydrate metabolism (Adams and Carre 2011) and these results confirm conservation of these pathways across C_3 and CAM species.

Limited overlap in enriched functional groups between the two species was observed, which could be explained by the circadian clock in *K. fedtschenkoi* controlling different biological processes relative to the clock in *A. thaliana*. Metabolic processes such as glycolysis and gluconeogenesis are known to have day-night rhythms that are unique and essential to CAM photosynthesis for recycling metabolites to conserve energy while maximizing photosynthetic efficiency (Wai et al. 2017). Indeed, processes related

to gluconeogenesis were found enriched in *K. fedtschenkoi*'s rhythmic transcriptome, with many phased to midday. Glyceraldehyde-3-phosphate is a prime end-product of photosynthesis and needed for glucose and sucrose synthesis, essential sugars in plants. Furthermore, organic acid synthesis was also enriched in the rhythmic transcriptome, in line with the strict schedule of CAM-related metabolism to prevent futile cycling between dark-period CO_2 fixation to malate and light-period malate decarboxylation (Borland et al. 2014). The results presented here support the idea that maintaining metabolic homeostasis via strict scheduling of the associated genes is important in CAM plants.

Several studies have compared rhythmic gene sets between different species as well as between different tissues within an organism and found conserved core clock networks but divergent clock output networks (Keegan et al. 2007; Michael et al. 2008; Filichkin et al. 2011; Eckel-Mahan et al. 2013; Boyle et al. 2017). Consistent with these findings, our comparative analysis of rhythmic gene sets between *A. thaliana* and *K. fedtschenkoi* also revealed divergence in clock output networks, e.g., phase shifts between rhythmic orthologs in these two plant species. By categorizing genes together in both species by orthology and rhythmicity, three clock-related phenotypes which were observed in Type 1, 2, and 3 ortholog groups could explain the divergence. The three clock-related phenotypes were observed in Type 1, 2, and 3 ortholog groups. The first clock-related phenotype is of synchronization of rhythmic orthologs. These synchronized orthologs were enriched with functional groups related to photosynthetic processes and a large portion of these orthologs were found phased during the day, coinciding with when photosynthesis takes place in both species. Photosynthesis in CAM species occurs normally as in C_3 species (Brautigam et al. 2017), so it is no surprise that orthologs involved in photosynthesis in both species are synchronized. These results are comparable with the results of phase-specific enrichment of functional groups across the entire rhythmic gene sets of both species, providing further support that the circadian clock plays an important role in photosynthesis in both C_3 and CAM species.

The second clock-related phenotype is conserved rhythmicity between orthologs but divergence in timing of expression. Two scenarios could explain these phase shifts: 1) the rhythmic gene is wired to a different core clock gene or 2) the controlling TF is wired to a different core clock gene thus shifting the expression of the gene. Evidence for the latter scenario is supported by our data where several orthologous TFs were found with similar phase shifts as non-TF orthologs. Several key CAM-related CO_2 fixation genes were found in the Type 1 ortholog groups that displayed this clock phenotype, although, none were found in the group of phase-shifted orthologs and TFs. However, this does not rule out the remaining TFs displaying a different phase shift potentially being the mechanism behind the phase shift of these key CAM-related CO_2 fixation genes. The TFs in Table 4 represent potential mechanisms for how the clock could have altered rhythmic gene expression in the context of CAM evolution beyond just CO_2 fixation and warrant further investigation into how they are integrated into the output of the clock between species and what downstream processes they regulate.

The third clock-related phenotype observed were genes rhythmic in one species and its ortholog arrhythmic in the other species. This raises the question: Are these genes found in Type 2 and 3 ortholog groups uniquely rhythmic to their respective species or

just not rhythmic in this study? It is possible that this phenotype could be partially caused by various technical factors, such as differences in sampling frequency and duration and transcriptomic platform. Additionally, using a different threshold to separate rhythmic genes from arrhythmic genes would produce different gene sets and thus produce different results. These issues are shared among circadian studies as there is currently no "gold" standard to determine rhythmic profiles, however, recent guidelines have been proposed to introduce consistency among studies (Hughes et al. 2017). In a biological sense, one scenario this phenotype could arise from is an adaptive gating system controlling the output network. This system would involve competition between core clock genes and other rhythmic TFs in binding on shared *cis*-elements in a condition dependent manner. The genes found arrhythmic in one species could be rhythmic under a different condition. Indeed, evidence for this scenario was found when comparing rhythmic gene sets between Zebrafish larva and livers and mouse livers (Boyle et al. 2017). The study found that different bHLH factors oscillate between the tissue types and species, suggesting that these bHLH factors could produce different rhythmic gene clusters by promoting or competing with the core clock TFs in a condition dependent manner. Additionally, it was recently reported that rhythmic genes expressed in the mesophyll tissue were not synchronized with their expression in vasculature tissue of *A. thaliana* (Endo et al. 2014). CAM-related processes are under strict temporal control, so it would be interesting to determine if rhythmicity of CAM-related genes would be condition-specific or constitutive in obligate CAM species. Circadian studies in the model CAM plant *K. fedtschenkoi* following the new guidelines in (Hughes et al. 2017) and involving various conditions are needed to address these questions.

Interestingly, in all the ortholog group types that contained rhythmic *K. fedtschenkoi* genes, most rhythmic genes were phased to midday. Several of these genes were either shifted to midday, with respect to their *A. thaliana* ortholog, or entrained to peak during midday. The core clock network is the main cause of rhythmic gene expression so this phase call distribution of rhythmic *K. fedtschenkoi* genes is likely caused by the core clock. Interestingly, the core clock network of *K. fedtschenkoi* differed from the *A. thaliana* core clock network, with several components having multiple copies and others having dampened expression. This raises multiple questions, such as is this the typical core clock network in *K. fedtschenkoi*, or is the network uncovered here the result of unknown internal or external factor feeding back into the clock? While diversity in topology exist between core clock networks, components can be defined as having high amplitudes of expression and being highly statistically rhythmic (Lou et al. 2012). The dampening of some core clock components within this *K. fedtschenkoi* dataset in conjunction with the study in which the *K. fedtschenkoi* data was generated in reporting no abnormalities in sampling method or data processing (Yang et al. 2017), suggest that the core clock network of *K. fedtschenkoi*, in terms of expression dynamics of individual components, is not representative of the core clock network of *K. fedtschenkoi*. Rather, the data here represents the output of the clock as a result of circadian gating and/or compensation by the clock to an unknown factor(s). The high concentration of rhythmic genes to midday and these genes being enriched for processes related to photosynthate production could be the response of the clock to the unknown factor(s). It is not

uncommon to interpret each circadian study as a snapshot of the circadian clock network under the influence of the environment in which the specimen was maintained (Boyle et al. 2017). To better investigate the circadian clock, a more robust circadian study in *K. fedtschenkoi* over a 48-h period, with 2-h sampling, and under diurnal and circadian conditions is needed. To further improve comparative studies, the same experimental conditions and sampling schedules should be carried out in *A. thaliana*.

Conclusion

In summary, several genes in *K. fedtschenkoi*, including CAM-related genes, displayed clock-related phenotypes suggesting that the clock could have played a role via a combination of the three mechanisms or all three. Additionally, the core clock network of *K. fedtschenkoi* was found to differ from the *A. thaliana* core clock model. Several scenarios are presented here that represent hypotheses to be tested in future experiments. Additionally, to improve the investigation of the circadian clock in CAM and how it influenced its evolution from C_3, a new circadian study involving both CAM (*K. fedtschenkoi)* and C_3 (*A. thaliana*) model species needs to be performed that follows the guidelines suggested in (Hughes et al. 2017). Not only will this produce quality resources for future investigations of circadian rhythm in plants, but also facilitate establishing parallels beyond the core clock network between CAM and C_3 output networks, aiding research into understanding the control mechanism behind the temporal dynamics of CAM.

Methods

K. fedtschenkoi and A. thaliana time-course data

The diurnal expression data for *K. fedtschenkoi* and *A. thaliana* were obtained from (Yang et al. 2017) and (Mockler et al. 2007), respectively. The *K. fedtschenkoi* expression data were collected at 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, and 24 h whereas the *A. thaliana* data were collected at 4, 8, 12, 16, 20, and 24 h after the initiation of the light period. Since the *A. thaliana* gene expression data were measured at 4-h intervals and the *K. fedtschenkoi* data were measured at 2-h intervals, the *A. thaliana* data were adjusted to arrive at expression profiles for all *A. thaliana* and *K. fedtschenkoi* genes on the same time scale. Specifically, the piecewise cubic Hermite interpolating polynomial (pchip) interpolation function in the pandas Python library was used to interpolate the sampled *A. thaliana* data to simulate gene expression levels at additional time points so that both time-course data sets consisted of the same time intervals: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, and 24 h. Since measurement of oscillations within the data sets was the goal, pchip was preferred over the more common method of cubic spline interpolation due to cubic spline's tendency to overshoot which introduces oscillations. Pchip maintains the shape of the data as best as possible and has been used on microarray time course data sets (Dong et al. 2011; Luo et al. 2014). *K. fedtschenkoi* genes with a max FPKM<1 were considered noise and removed.

Ranking rhythmic genes in K. fedtschenkoi and A. thaliana

The rhythmic genes in *K. fedtschenkoi* and *A. thaliana* were ranked by using a methodology developed by (Kelliher et al. 2016), which applied four periodicity detection algorithms to time-course gene expression data sets from two yeast species: de Lichtenberg (de Lichtenberg et al. 2005), JTK-CYCLE (JTK) (Hughes et al. 2010), Lomb-Scargle (Scargle 1982), and persistent homology (Lomb 1975). Each algorithm favors slightly different periodic shapes (Deckard et al. 2013) so the periodicity rankings from each algorithm was summed and all yeast genes by cumulative scores for each species were ranked. In this study, all algorithms, as described previously in (Deckard et al. 2013), were applied to *A. thaliana* and *K. fedtschenkoi* 24-h time-course data except persistent homology as it has been shown to give poor results when used on higher organisms such as plants (Deckard et al. 2013). For both time course datasets, a period length of 24 hours and a range of 20-24 hours were applied to search and rank rhythmic genes. The period and length of rhythmic signals were chosen since circadian rhythms were investigated, which is roughly a 24-h cycle.

Tables S1 and S2 summarize the rhythmicity-ranking outputs and gene expression statistics for all *A. thaliana* and *K. fedtschenkoi* genes, respectively. Every 500 genes for each dataset was plotted to visually inspect where rhythmicity in expression begins to fade. A cutoff of 10,000 and 10,500 was determined to identify top ranked genes for *A. thaliana* and *K. fedtschenkoi*, respectively. To validate this method on time course plant data with the length of one period, the top 10,000 ranked genes in *A. thaliana* were compared to a previously published rhythmic gene set using the same data and found a 48% overlap (Mockler et al. 2007). A 100% overlap was not expected as variables such as data normalization, periodicity detection and/or algorithm score cutoff will result in varying sets of genes even from the same data set (Keegan et al. 2007; Kelliher et al. 2016).

To better quantify the datasets for rhythmic genes, a quantitative score cutoff was applied to the 10,000 and 10,500 rhythmic gene sets for *A. thaliana* and *K. fedtschenkoi*, respectively. The JTK algorithm was adapted from statistics for biological data and correlates pairs of points and then computes the significance of the correlation to that of a reference curve. JTK has shown to give preference to cosine and peaked gene expression profiles (Deckard et al. 2013) and a previous study examining rhythmicity in transcription factors and transcription co-regulators in the CAM plant pineapple found cosine and peaked profiles were most successful in identifying cycling transcription factors and transcription co-regulators (Sharma et al. 2017). Graphing the p-values generated by the JTK algorithm, an inflection point in the distribution was observed around a p-value of 0.06. Therefore, a filter was chosen on the JTK algorithm at a p-value cutoff of 0.06. The intersection of genes in the two rhythmic datasets ranked by all algorithms and genes that passed the JTK cutoff was taken, identifying 8,657 rhythmic genes in *A. thaliana* and 7,549 rhythmic genes in *K. fedtschenkoi* (Figure 4).

Functional analysis of rhythmic genes K. fedtschenkoi and A. thaliana

Gene Ontology (GO) terms for the *K. fedtschenkoi* and *A. thaliana* genes were used from the gene annotation information downloaded from Phytozome v12.1

(Goodstein et al. 2012) (Kfedtschenkoi_382_v1.1.annotation_info.txt and Athaliana_167_TAIR10.annotation_info.txt, respectively). *K. fedtschenkoi* genes encoding putative transcription factors were retrieved from (Yang et al. 2017), while *A. thaliana* transcription factors were retrieved from PlanTFDB v4.0 (Jin et al. 2017).

Using ClueGO (Bindea et al. 2009), observed GO biological process were subjected to the right-sided hypergeometric enrichment test at medium network specificity selection and p-value correction was performed using the Holm-Bonferroni step-down method (Holm 1979). There was a minimum of 3 and a maximum of 8 selected GO-tree levels, while each cluster was set to include a minimum of between 3% and 4% of genes associated with each term. GO-term fusion and grouping settings were selected to minimize GO-term redundancy and the term enriched at the highest level of significance was used as the representative term for each functional cluster. The GO terms with p-values less than or equal to 0.05 were considered significantly enriched.

Identification of orthologous genes

Orthologous groups constructed from the protein sequences of 26 plant species including *Amborella trichopoda* (PLAZA 3.0 (Proost, 2015 #126)); available at http://bioinformatics.psb.ugent.be/plaza/), *Ananas comosus* (Ming et al. 2015), *Aquilegia coerulea* (v1.1; available at https://phytozome.jgi.doe.gov), *Arabidopsis thaliana* (PLAZA 3.0), *Beta vulgaris* (PLAZA 3.0), *Brachypodium distachyon* (PLAZA 3.0), *Carica papaya* (PLAZA 3.0), *Citrus sinensis* (PLAZA 3.0), *Eucalyptus grandis* (PLAZA 3.0), *Fragaria vesca* (PLAZA 3.0), *Kalanchoë fedtschenkoi* (v1.1; available at https://phytozome.jgi.doe.gov), *Medicago truncatula* (PLAZA 3.0), *Mimulus guttatus* (PLAZA 3.0), *Musa acuminata* (PLAZA 3.0), *Oryza sativa* (PLAZA 3.0), *Phalaenopsis equestris* (Cai et al. 2015), *Physcomitrella patens* (PLAZA 3.0), *Populus trichocarpa* (PLAZA 3.0), *Prunus persica* (PLAZA 3.0), *Setaria italica* (PLAZA 3.0), *Solanum lycopersicum* (PLAZA 3.0), *Solanum tuberosum* (PLAZA 3.0), *Sorghum bicolor* (PLAZA 3.0), *Theobroma cacao* (PLAZA 3.0), *Vitis vinifera* (PLAZA 3.0), and *Zea may* (PLAZA 3.0) were taken from (Yang et al. 2017). The orthologous groups were constructed using FastOrtho (http://enews.patricbrc.org/fastortho/), a reimplementation of OrthoMCL program (Li et al. 2003), using default parameters, except for a BLASTp E-value cutoff of 1e-5 and an inflation value of 1.3. The longest protein sequence was selected in case of multiple transcripts annotated for one gene locus.

To determine if *K. fedtschenkoi* genes have homologs in other plant species, *K. fedtschenkoi* protein sequences were queried in a BLASTP search of a local plant proteome database. The proteomes of 83 plant species (Table S5), including 70 species from Plaza 4.0 (Van Bel et al. 2017) and 10 species from Phytozome 12.0 (Goodstein et al. 2012), and three *Agave* species with transcriptome sequenced(Gross et al. 2013; Abraham et al. 2016) . National Center for Biotechnology Information (NCBI) BLAST+ (Camacho et al. 2009) executables were used to construct the plant proteome BLAST database (makeblastdb) and to then search the database using BLASTp with E-value cutoffs of 1e-5 and 1e-10.

References

Abraham PE, Yin H, Borland AM, Weighill D, Lim SD, De Paoli HC, Engle N, Jones PC, Agh R, Weston DJ et al. 2016. Transcript, protein and metabolite temporal dynamics in the CAM plant Agave. *Nat Plants* **2**: 16178.

Adams S, Carre IA. 2011. Downstream of the plant circadian clock: output pathways for the control of physiology and development. *Essays in Biochemistry* **1**: 53-69.

Bar-Joseph Z, Gitter A, Simon I. 2012. Studying and modelling dynamic biological processes using time-series gene expression data. *Nat Rev Genet* **13**: 552-564.

Bell-Pederson D, Cassone VM, Earnest DJ, Golden SS, Hardin PE, Thomas TL, Zoran MJ. 2005. CIRCADIAN RHYTHMS FROM MULTIPLE OSCILLATORS: LESSONS FROM DIVERSE ORGANISMS. *Nature Reviews: Genetics*: 1-13.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**: 1091-1093.

Borland AM, Hartwell J, Weston DJ, Schlauch KA, Tschaplinski TJ, Tuskan GA, Yang X, Cushman JC. 2014. Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci* **19**: 327-338.

Boxall SF, Foster JM, Bohnert HJ, Cushman JC, Nimmo HG, Hartwell J. 2005. Conservation and divergence of circadian clock operation in a stress-inducible Crassulacean acid metabolism species reveals clock compensation against stress. *Plant Physiol* **137**: 969-982.

Boyle G, Richter K, Priest HD, Traver D, Mockler TC, Chang JT, Kay SA, Breton G. 2017. Comparative Analysis of Vertebrate Diurnal/Circadian Transcriptomes. *PLoS One* **12**: e0169923.

Brautigam A, Schluter U, Eisenhut M, Gowik U. 2017. On the Evolutionary Origin of CAM Photosynthesis. *Plant Physiol* **174**: 473-477.

Cai J, Liu X, Vanneste K, Proost S, Tsai WC, Liu KW, Chen LJ, He Y, Xu Q, Bian C et al. 2015. The genome sequence of the orchid Phalaenopsis equestris. *Nat Genet* **47**: 65-72.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421.

Covington MF, Maloof JN, Straume M, Kay SA, Harmer SL. 2008. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol* **9**: R130.

Cushman JC, Tillett RL, Wood JA, Branco JM, Schlauch KA. 2008. Large-scale mRNA expression profiling in the common ice plant, Mesembryanthemum crystallinum, performing C3 photosynthesis and Crassulacean acid metabolism (CAM). *J Exp Bot* **59**: 1875-1894.

de Lichtenberg U, Jensen LJ, Fausboll A, Jensen TS, Bork P, Brunak S. 2005. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* **21**: 1164-1171.

Deckard A, Anafi RC, Hogenesch JB, Haase SB, Harer J. 2013. Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data. *Bioinformatics* **29**: 3174-3180.

Doherty CJ, Kay SA. 2010. Circadian control of global gene expression patterns. *Annu Rev Genet* **44**: 419-444.

Dong B, Zhang P, Chen X, Liu L, Wang Y, He S, Chen R. 2011. Predicting Housekeeping Genes Based on Fourier Analysis. *PloS ONE* **6**: 1-11.

Eckel-Mahan KL, Patel VR, de Mateo S, Orozco-Solis R, Ceglia NJ, Sahar S, Dilag-Penilla SA, Dyar KA, Baldi P, Sassone-Corsi P. 2013. Reprogramming of the circadian clock by nutritional challenge. *Cell* **155**: 1464-1478.

Endo M, Shimizu H, Nohales MA, Araki T, Kay SA. 2014. Tissue-specific clocks in Arabidopsis show asymmetric coupling. *Nature* **515**: 419-422.

Filichkin SA, Breton G, Priest HD, Dharmawardhana P, Jaiswal P, Fox SE, Michael TP, Chory J, Kay SA, Mockler TC. 2011. Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. *PLoS One* **6**: e16907.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178-1186.

Greenham K, McClung CR. 2015. Integrating circadian dynamics with physiological processes in plants. *Nat Rev Genet* **16**: 598-610.

Gross S, Martin J, Simpson J, Abraham-Juarez M, Wang Z, Visel A. 2013. De Novo transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. *BMC Genomics* **14**: 1-14.

Hartwell J. 2005. The co-ordination of central plant metabolism by the circadian clock. *Mechanistic and Functional Studies of Proteins* **33**: 945-948.

Hartwell J. 2006. The circadian clock in CAM plants. *Endogenous Plant Rhythms*: 211-236.

Higashi T, Tanigaki Y, Takayama K, Nagano AJ, Honjo MN, Fukuda H. 2016. Detection of Diurnal Variation of Tomato Transcriptome through the Molecular Timetable Method in a Sunlight-Type Plant Factory. *Front Plant Sci* **7**: 87.

Holm S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**: 65-70.

Hotta CT, Gardner MJ, Hubbard KE, Baek SJ, Dalchau N, Suhita D, Dodd AN, Webb AA. 2007. Modulation of environmental responses of plants by circadian clocks. *Plant Cell Environ* **30**: 333-349.

Hughes ME, Abruzzi K, Allada R, Anafi RC, Arpat A, Asher G, Baldi P, Bekker C, Bell-Pederson D, Blau J et al. 2017. Guidelines for Genome-Scale Analysis of Biological Rhythms. *Journal of Biological Rhythms* **32**: 380-393.

Hughes ME, Hogenesch JB, Kornacker K. 2010. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms* **25**: 372-380.

Jin J, Tian F, Yang D, Meng Y, Kong L, Luo J, Gao G. 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* **45**: D1040–D1045.

Keegan KP, Pradhan S, Wang JP, Allada R. 2007. Meta-analysis of Drosophila circadian microarray studies identifies a novel set of rhythmically expressed genes. *PLoS Comput Biol* **3**: e208.

Kelliher CM, Leman AR, Sierra CS, Haase SB. 2016. Investigating Conservation of the Cell-Cycle-Regulated Transcriptional Program in the Fungal Pathogen, Cryptococcus neoformans. *PLoS Genet* **12**: e1006453.

Koda S, Onda Y, Matsui H, Takahagi K, Yamaguchi-Uehara Y, Shimizu M, Inoue K, Yoshida T, Sakurai T, Honda H et al. 2017. Diurnal Transcriptome and Gene Network Represented through Sparse Modeling in Brachypodium distachyon. *Front Plant Sci* **8**: 2055.

Li J, Grant GR, Hogenesch JB, Hughes ME. 2015. Considerations for RNA-seq analysis of circadian rhythms. *Methods Enzymol* **551**: 349-367.

Li L, Stoeckert C, Roos D. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: 2178-2189.

Lomb NR. 1975. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science* **39**: 447-462.

Lou P, Wu J, Cheng F, Cressman LG, Wang X, McClung CR. 2012. Preferential retention of circadian clock genes during diploidization following whole genome triplication in Brassica rapa. *Plant Cell* **24**: 2415-2426.

Luo Z, Azencott R, Zhao Y. 2014. Modeling miRNA-mRNA interactions: fitting chemical kinetics equations to microarray data. *BMC Systems Biology* **9**: 1-15.

Mallona I, Egea-Cortines M, Weiss J. 2011. Conserved and divergent rhythms of crassulacean acid metabolism-related and core clock gene expression in the cactus Opuntia ficus-indica. *Plant Physiol* **156**: 1978-1989.

McClung CR. 2013. Beyond Arabidopsis: the circadian clock in non-model plant species. *Semin Cell Dev Biol* **24**: 430-436.

Merida A, Rodriquez-Galan JM, Vincent C, Romero JM. 1999. Expression of the Granule-Bound Starch Synthase I (Waxy) Gene from Snapdragon Is Developmentally and Circadian Clock Regulated. *Plant Physiology* **120**: 401-409.

Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM et al. 2008. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet* **4**: e14.

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* **47**: 1435-1442.

Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J. 2007. The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis. *Cold Spring Harbor Symposia on Quantitative Biology* **72**: 353-363.

Nohales MA, Kay SA. 2016. Molecular mechanisms at the core of the plant circadian oscillator. *Nat Struct Mol Biol* **23**: 1061-1069.

Piechulla B. 1988. Plastid and nuclear mRNA fluctuations in tomato leaves - diurnal and circadian rhythms during extended dark and light periods. *Plant Molecular Biology* **11**: 345-353.

Scargle JD. 1982. STUDIES IN ASTRONOMICAL TIME SERIES ANALYSIS. II. STATISTICAL ASPECTS OF SPECTRAL ANALYSIS OF UNEVENLY SPACED DATA. *The Astrophysical Journal* **263**: 835-853.

Sharma A, Wai CM, Ming R, Yu Q. 2017. Diurnal Cycling Transcription Factors of Pineapple Revealed by Genome-Wide Annotation and Global Transcriptomic Analysis. *Genome Biol Evol* **9**: 2170-2190.

Thain S, Hall A, Millar AJ. 2000. Functional independence of circadian clocks that regulate plant gene expression. *Current Biology* **10**: 951-956.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Candepoele K. 2017. PLAZA 4.0: an integrative resource for functional evolutionary and comparative plant genomices. *Nucleic Acids Research* **46**: D1190-D1196.

von Caemmerer S, Griffiths H. 2009. Stomatal responses to CO2 during a diel Crassulacean acid metabolism cycle in Kalanchoe daigremontiana and Kalanchoe pinnata. *Plant Cell Environ* **32**: 567-576.

Wai CM, VanBuren R, Zhang J, Huang L, Miao W, Edger PP, Yim WC, Priest HD, Meyers BC, Mockler T et al. 2017. Temporal and spatial transcriptomic and microRNA dynamics of CAM photosynthesis in pineapple. *Plant J* **92**: 19-30.

Wyka TP, Bohn A, Duarte HM, Kaiser F, Luttge UE. 2004. Perturbations of malate accumulation and the endogenous rhythms of gas exchange in the Crassulacean acid metabolism plant Kalanchoe daigremontiana: testing the tonoplast-as-oscillator model. *Planta* **219**: 705-713.

Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, Liu D, Weigel D, Cheol Yim W, Ha J et al. 2017. The Kalanchoë genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications* **8**: 1-15.

Zhang R, Lahens N, Ballance H, Hughes ME, Hogenesch JB. 2014. A circadian gene expression atlas in mammals: Implications for biology and medicine. *PNAS* **111**: 16219-16224.

Appendix

Table 2. Carboxylation genes in *Kalanchoë fedtschenkoi* and their orthologs in *Arabidopsis thaliana*.

The Ortholog group type in Figure 8.

Gene Symbol	*K. fedtschenkoi* ID	Gene Description	Ortholog Group	Max Expression (FPKM)	*K. fedtschenkoi* Rhythmicity	*K. fedtschenkoi* Phase Call	*A. thaliana* ID	*A. thaliana* Rhythmicity	*A. thaliana* Phase Call	Phase Shift (hours)
β-CA	Kaladp0018s0287	Beta carbonic anhydrase	Type 5	219.37	rhythmic	20				
β-CA	Kaladp0018s0289	Beta carbonic anhydrase	Type 5	352.68	rhythmic	20				
β-CA	Kaladp0034s0051	Beta carbonic anhydrase	Type 5	267.513	rhythmic	4				
PPCK1	Kaladp0037s0517	Phosphoenolpyruvate carboxylase kinase	Type 1	1187.84	rhythmic	20	AT3G04530			
							AT1G08650	rhythmic	8	12
PPCK2	Kaladp0604s0001	Phosphoenolpyruvate carboxylase kinase	Type 1	376.3	rhythmic	16	AT3G04530			
							AT1G08650	rhythmic	8	8
PEPC2	Kaladp0048s0578	Phosphoenolpyruvate carboxylase 2	Type 1	280.38	rhythmic	18	AT2G42600			
							AT3G14940	rhythmic	20	2
							AT1G68750			
							AT1G53310	rhythmic	12	6
PEPC1	Kaladp0095s0055	Phosphoenolpyruvate carboxylase 1	Type 1	8338.11	rhythmic	12	AT2G42600			
							AT3G14940	rhythmic	20	8
							AT1G68750			
							AT1G53310	rhythmic	12	0
MDH	Kaladp0001s0257	Malate dehydrogenase	Type 1	383.29	rhythmic	12	AT2G22780	rhythmic	12	0
MDH	Kaladp0022s0111	Malate dehydrogenase	Type 1	43.82			AT2G22780	rhythmic	12	0
MDH	Kaladp0082s0194	Malate dehydrogenase	Type 3	1925.18	rhythmic	8	AT5G56720			
							AT5G43330			
							AT1G04410			

Table 2. Continued.

Gene Symbol	*K. fedtschenkoi* ID	Gene Description	Ortholog Group	Max Expression (FPKM)	*K. fedtschenkoi* Rhythmicity	*K. fedtschenkoi* Phase Call	*A. thaliana* ID	*A. thaliana* Rhythmicity	*A. thaliana* Phase Call	Phase Shift (hours)
ALMT	Kaladp0024s0194	Tonoplast aluminum-activated malate transporter	Type 2	147.01			AT1G25480 AT3G18440 AT2G17470 AT1G18420 AT1G68600	rhythmic	8	
ALMT 6	Kaladp0073s0021	Tonoplast aluminum-activated malate transporter 6	Type 1	440.84	rhythmic	2	AT1G25480 AT3G18440 AT2G17470 AT1G18420 AT1G68600	rhythmic	8	6
ALMT 6	Kaladp0062s0038	Tonoplast aluminum-activated malate transporter	Type 1	66.01	rhythmic	16	AT1G25480 AT3G18440 AT2G17470 AT1G18420 AT1G68600	rhythmic	8	8

Table 3. Decarboxylation genes in *Kalanchoë fedtschenkoi* and their orthologs in *Arabidopsis thaliana*.

The Ortholog group type in Figure 8.

Gene Symbol	*K. fedtschenkoi* ID	Gene Description	Ortholog Group	Max Expression (FPKM)	*K. fedtschenkoi* Rhythmicity	*K. fedtschenkoi* Phase Call	*A. thaliana* ID	*A. thaliana* Rhythmicity	*A. thaliana* Phase Call	Phase Shift (hours)
TDT	Kaladp0042s0251	Tonoplast dicarboxylate transporter	Type 1	1617.63	rhythmic	2	AT5G47560	rhythmic	24	2
PPDK	Kaladp0039s0092	Pyruvate, orthophosphate dikinase	Type 5	4744.54	rhythmic	24				
PPDK	Kaladp0076s0229	Pyruvate, orthophosphate dikinase	Type 2	9054.27			AT4G15530	rhythmic	16	8
PPDK-RP	Kaladp0010s0106	PPDK regulatory protein	Type 3	121.13	rhythmic	8	AT4G21210			
PPDK-RP	Kaladp0060s0363	PPDK regulatory protein	Type 3	81.78	rhythmic	10	AT4G21210			
NAD-ME	Kaladp0033s0124	NAD-dependent malic enzyme	Type 1	269.31	rhythmic	2	AT4G00570	rhythmic	16	10
							AT2G13560			
NADP-ME	Kaladp0092s0166	NADP-malic enzyme	Type 1	2948.96	rhythmic	24	AT1G79750	rhythmic	20	4
							AT5G11670	rhythmic	12	12
							AT5G25880	rhythmic	16	8
							AT2G19900	rhythmic	24	0

Table 4. Type 1 rhythmic orthologous transcription factors in *Kalanchoë fedtschenkoi* and *Arabidopsis thaliana*.

K. fedtschenkoi ID	*K. fedtschenkoi* Phase Call	*A. thaliana* ID	*A. thaliana* Phase Call	Phase shift (hours)	*A. thaliana* Symbol	*A. thaliana* Description
Kaladp0057s0097	2	AT1G09530	20	6	PIF3	phytochrome interacting factor 3
Kaladp0015s0032	4	AT4G38890	16	12		FMN-linked oxidoreductases superfamily protein
Kaladp0003s0081	6	AT5G42200	12	6		RING/U-box superfamily protein
Kaladp0011s0211	6	AT3G45880	16	10		2-oxoglutarate (2OG) and Fe (II)-dependent oxygenase superfamily protein
Kaladp0098s0059	6	AT1G76710	20	10	ASHH1	SET domain group 26
Kaladp0087s0172	6	AT4G00090	20	10		Transducin/WD40 repeat-like superfamily protein
Kaladp0053s0598	6	AT5G64730	20	10		Transducin/WD40 repeat-like superfamily protein
Kaladp0024s0982	6	AT2G47450	24	6	CAO	chloroplast signal recognition particle component
Kaladp0058s0485	6	AT5G51110	24	6		Transcriptional coactivator/pterin dehydratase
Kaladp0039s0495	8	AT1G68550	16	8	CRF10	Integrase-type DNA-binding superfamily protein
Kaladp0095s0706	8	AT1G75430	16	8	BLH11	BEL1-like homeodomain 11
Kaladp0037s0533	8	AT5G28300	16	8		Duplicated homeodomain-like superfamily protein
Kaladp1129s0043	8	AT4G17060	20	12	FIP2	FRIGIDA interacting protein 2
Kaladp0095s0494	8	AT4G32570	20	12	TIFY8	TIFY domain protein 8
Kaladp0050s0018	8	AT5G28640	24	8	AN3	SSXT family protein
Kaladp0034s0058	10	AT1G10610	4	6		basic helix-loop-helix (bHLH) DNA-binding superfamily protein
Kaladp0068s0095	10	AT4G12240	16	6		zinc finger (C2H2 type) family protein
Kaladp0809s0115	10	AT4G29940	16	6	PRHA	pathogenesis related homeodomain protein A
Kaladp0042s0203	10	AT5G14370	16	6		CCT motif family protein
Kaladp0081s0357	10	AT5G63420	16	6	emb2746	RNA-metabolizing metallo-beta-lactamase family protein
Kaladp0039s0570	10	AT3G52190	20	10	PHF1	phosphate transporter traffic facilitator1
Kaladp0044s0029	10	AT2G32000	24	10		DNA topoisomerase, type IA, core
Kaladp0808s0018	12	AT1G22860	4	8		Vacuolar sorting protein 39
Kaladp0033s0181	12	AT3G18640	20	8		Zinc finger C-x8-C-x5-C-x3-H type family protein
Kaladp0015s0185	12	AT1G49040	24	12	SCD1	stomatal cytokinesis defective / SCD1 protein
Kaladp0099s0116	12	AT5G50970	24	12		transducin family protein / WD-40 repeat family protein
Kaladp0085s0129	12	AT5G56780	24	12	ET2	effector of transcription2
Kaladp0004s0037	14	AT2G04240	8	6	XERICO	RING/U-box superfamily protein
Kaladp0746s0006	14	AT2G47270	8	6	UPB1	transcription factor UPBEAT protein
Kaladp0011s0744	14	AT1G08810	24	10	MYB60	myb domain protein 60
Kaladp0008s0534	14	AT2G39810	24	10	HOS1	ubiquitin-protein ligase
Kaladp0055s0201	14	AT4G37650	24	10	SHR	GRAS family transcription factor
Kaladp0093s0080	16	AT4G00150	24	8	HAM3	GRAS family transcription factor
Kaladp0048s0281	16	AT5G01160	24	8	HAKAI	RING/U-box superfamily protein
Kaladp0060s0155	18	AT3G50590	8	10		Transducin/WD40 repeat-like superfamily protein
Kaladp0011s0085	18	AT1G58025	12	6		DNA-binding bromodomain-containing protein
Kaladp0050s0333	22	AT1G68190	8	10	BBX27	B-box zinc finger family protein
Kaladp0048s0675	22	AT3G23490	8	10	CYN	cyanase
Kaladp0011s0272	22	AT5G03720	12	10	HSFA3	heat shock transcription factor A3
Kaladp0036s0299	24	AT5G58410	16	8		HEAT/U-box domain-containing protein

Table 5. Circadian parameters of *Kalanchoë fedtschenkoi* core clock orthologs.

Clock Gene	*K. fedtschenkoi* Id	Normalized Rank	Mean FPKM	Absolute Amplitude	Fold-change	JTK-CYCLE p-value
LUX/NOX	Kaladp0033s0047	34	19.45	42.12	10.95	1.06E-05
ELF3	Kaladp0039s0732	177	30.61	64.15	11.13	5.39E-04
ELF4a	Kaladp0037s0163	207	11.75	35.32	47.72	4.84E-05
ELF4b	Kaladp0045s0206	737	36.91	122.38	622.21	1.46E-03
PRR9/5	Kaladp0032s0115	818	36.87	127.70	327.39	4.84E-05
RVE8	Kaladp0577s0020	1180	40.88	143.53	992.09	4.84E-05
RVE6a	Kaladp0019s0045	2589	52.55	76.76	3.73	7.92E-03
RVE6b	Kaladp0022s0168	2672	154.49	700.15	476.50	4.84E-05
LNK1a	Kaladp0607s0046	4139	13.85	45.15	33.56	1.65E-02
LNK2a	Kaladp0099s0129	4214	41.61	99.10	15.51	3.22E-02
RVE6c	Kaladp0055s0349	5756	13.42	8.30	1.97	1.65E-02
ELF4c	Kaladp0059s0037	6111	6.34	7.77	3.67	3.22E-02
LNK2b	Kaladp0060s0264	6361	60.40	124.40	11.48	1.05E-01
GI	Kaladp0040s0489	6902	4.18	9.37	6.60E+07	2.89E-01
LNK1b	Kaladp0047s0123	12058	3.17	14.92	6.81E+05	1.00E+00
PRR7	Kaladp0001s0237	12078	0.74	3.23	NA	6.92E-01
CCA/LHYa	Kaladp0066s0115	13093	1.74	8.30	3.97E+07	4.54E-01
CHE	Kaladp0032s0054	14327	23.19	22.66	2.54	2.89E-01
LWD1/2	Kaladp0048s0797	15008	13.89	4.12	1.34	4.54E-01
TOC1	Kaladp0040s0446	NA	0.18	0.67	26665.50	NA
PRR7	Kaladp0101s0041	NA	0.22	0.63	2.19E+43	NA
CCA/LHYb	Kaladp0496s0018	NA	0.10	0.58	2.85E+23	NA

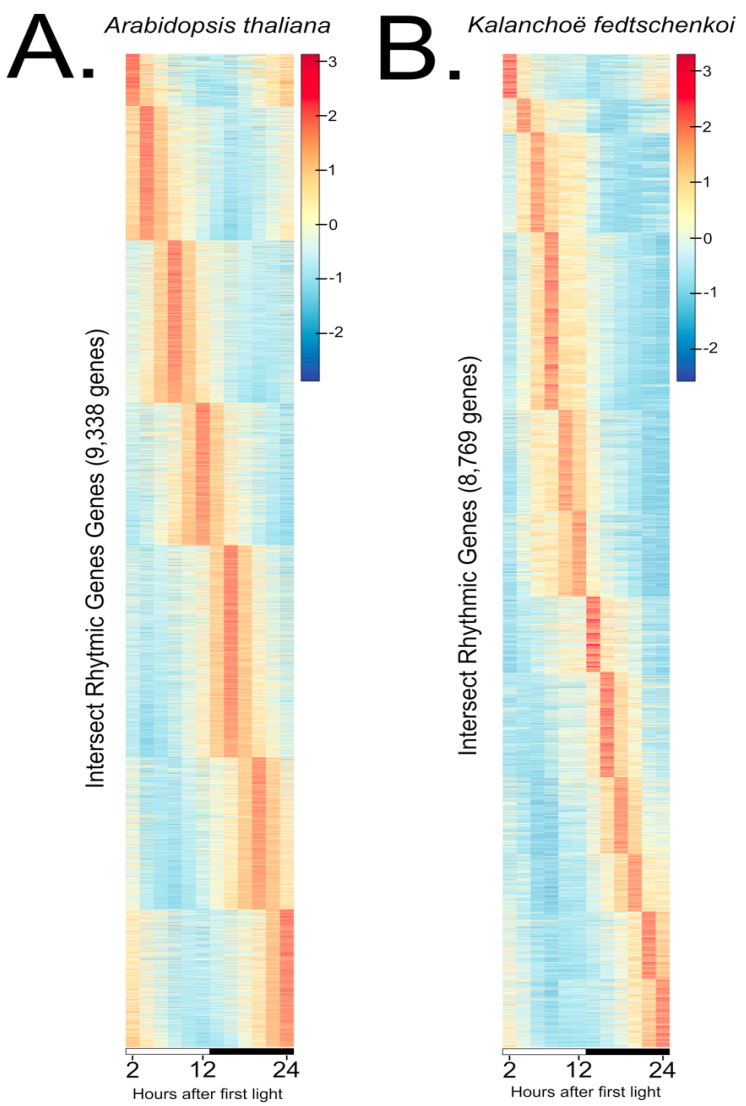


Figure 4. Rhythmically expressed genes in *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi*.

A JTK cutoff of 0.06 was applied to each set of rhythmic genes identified by visual inspective to apply a qualitative selection for highly rhythmic genes in each species. **A)** 9,338 genes were identified as highly rhythmic in *A. thaliana*. **B)** 8,769 genes were identified as highly rhythmic in *K. fedtschenkoi*. X-axis corresponds to number of hours after first light. Time points 12 and 24 correspond to subjective dusk and dawn, respectively. Transcript levels are depicted as a z-score change relative to mean expression for each gene, where values represent the number of standard deviations away from the mean.

Figure 5. Gene ontology categories overrepresented among *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi* rhythmic genes.

The radius of each circle denotes the number of genes in each category. Gene ontology overrepresentation bar chart and graph were generated using the network visualization tool Cytoscape with the add-on application ClueGO. **A)** Biological functions enriched in *A. thaliana* rhythmic gene set. **B)** Biological functions enriched in *K. fedtschenkoi* rhythmic gene set.

Figure 5. Continued.

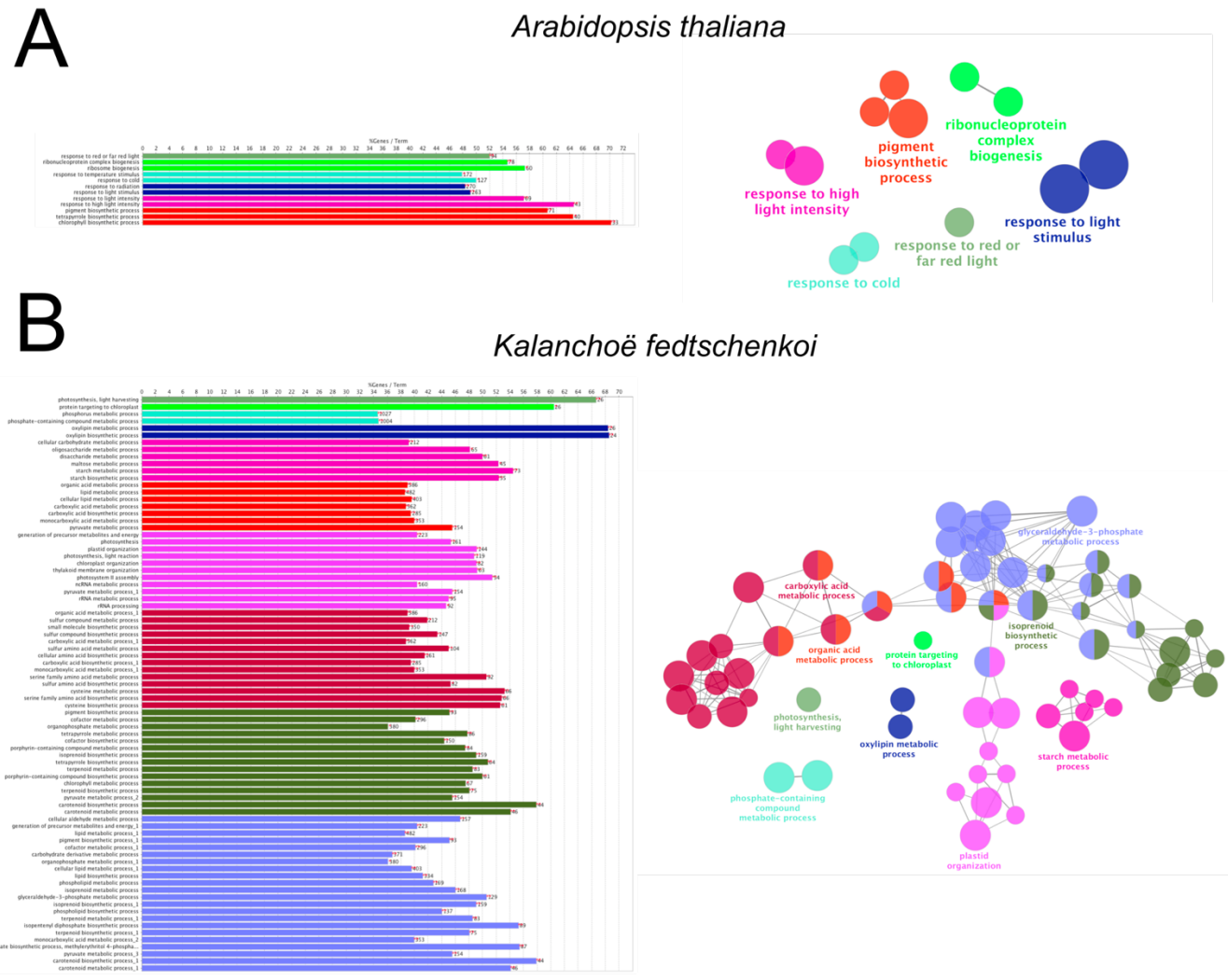


Figure 6. Gene co-expression networks of the rhythmic gene sets in *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi* rhythmic gene sets.

Gene co-expression networks for each species were constructed by calculating the Spearman ranked correlation coefficient for all pair-wise combinations of gene expression in each species' rhythmic gene set. Gene ontology overrepresentations were determined using the Cytoscape add-on ClueGO. **A)** A circular gene co-expression network for *A. thaliana* rhythmic genes with nodes colored to their phase call in their diurnal expressions. **B)** A circular gene co-expression network for *K. fedtschenkoi* rhythmic genes with nodes colored to their phase call in their diurnal expressions. **C)** Transcription factors encoded by the rhythmic gene in *A. thaliana*. **D)** Transcription factors encoded by the rhythmic gene in *K. fedtschenkoi*. **E)** Enriched biological functions in each phase of the day and night for the rhythmic genes in *A. thaliana* (left) and *K. fedtschenkoi* (right).

Figure 6. Continued.

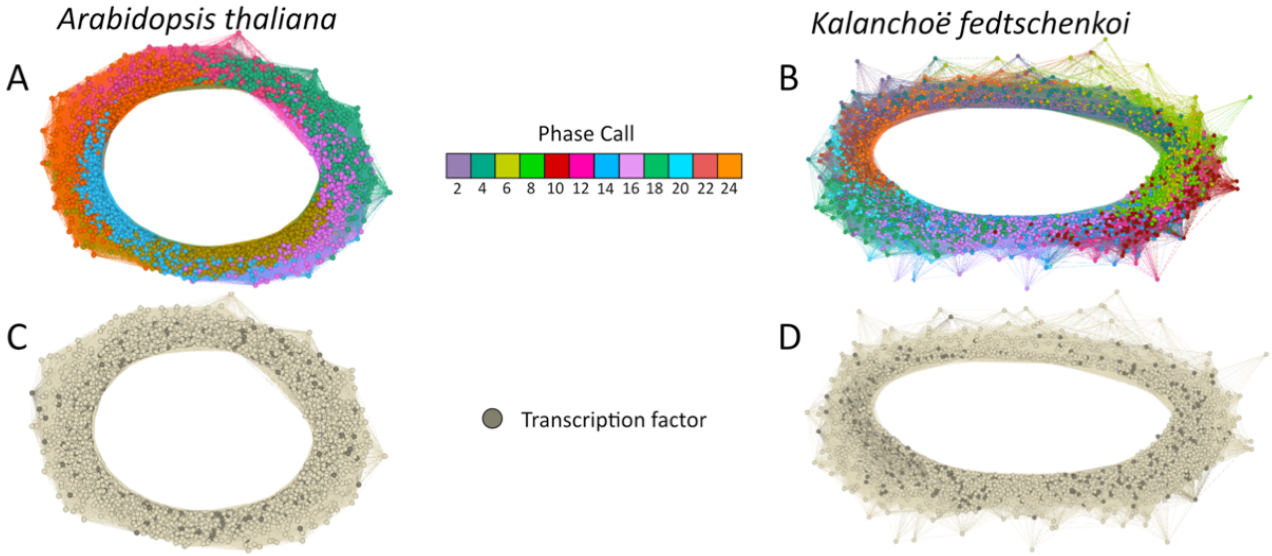


Figure 6. Continued.

E

Arabidopsis thaliana												GOID	Description	*Kalanchoë fedtschenkoi*										
24	22	20	18	16	14	12	10	8	6	4	2			2	4	6	8	10	12	14	16	18	20	22	24
												GO:0009765	photosynthesis, light harvesting	1.34E-20											
												GO:0008610	lipid biosynthetic process	1.28E-07											
												GO:0009069	serine family amino acid metabolic process	9.17E-05											
												GO:0016104	triterpenoid biosynthetic process	1.64E-04											
												GO:0009644	response to high light intensity	0.0067											
												GO:0042548	regulation of photosynthesis, light reaction	0.0199											
												GO:0072521	purine-containing compound metabolic process	0.0254											
												GO:0006885	regulation of pH	0.039											
												GO:0010025	wax biosynthetic process	0.0497											
										1.80E-02		GO:0006139	nucleobase-containing compound metabolic process												
												GO:0010200	response to chitin		0.0123										
												GO:0009697	salicylic acid biosynthetic process		0.0259										
												GO:0019761	glucosinolate biosynthetic process		0.0312										
												GO:0032269	negative regulation of cellular protein metabolic process		0.0397										
												GO:0019682	glyceraldehyde-3-phosphate metabolic process			2.02E-07			3.00E-05						
												GO:0009657	plastid organization			4.96E-09			2.30E-05						
												GO:0034470	ncRNA processing			1.80E-04									
												GO:0019684	photosynthesis, light reaction			2.80E-04									
						5.02E-03						GO:0005982	starch metabolic process			0.0061		4.78E-04							
												GO:0045036	protein targeting to chloroplast			0.0132			0.046						
												GO:0010016	shoot system morphogenesis			0.0396									
								1.70E-03				GO:0006914	autophagy												
										1.80E-05		GO:0015979	photosynthesis				2.27E-04								
												GO:0008654	phospholipid biosynthetic process				0.0037								
								4.16E-02				GO:0034976	response to endoplasmic reticulum stress												
												GO:0016554	cytidine to uridine editing					2.24E-04							
												GO:0009451	RNA modification					4.13E-04							
												GO:0006468	protein phosphorylation					5.61E-05							
												GO:0016556	mRNA modification						1.21E-05						
												GO:0031425	chloroplast RNA processing						0.001						
												GO:1902531	regulation of intracellular signal transduction							0.0046					
												GO:0048583	regulation of response to stimulus							0.0129					
												GO:0006566	threonine metabolic process								0.0012				
												GO:0006913	nucleocytoplasmic transport								0.0082	0.0026			
												GO:0006563	L-serine metabolic process								0.0358				
												GO:0015031	protein transport									1.66E-08			
												GO:0016192	vesicle-mediated transport									7.18E-07			
												GO:0031408	oxylipin biosynthetic process									3.60E-05			
												GO:0006101	citrate metabolic process									3.49E-04			
												GO:1901565	organonitrogen compound catabolic process										4.06E-06		
												GO:0006412	translation										6.47E-04		
												GO:0010035	response to inorganic substance										0.005		
		1.85E-12										GO:0022613	ribonucleoprotein complex biogenesis												
		3.22E-05										GO:0006396	RNA processing												
		3.90E-03										GO:0006413	translational initiation												
												GO:0009446	putrescine biosynthetic process											0.0047	
												GO:0006021	inositol biosynthetic process											0.009	
												GO:0005985	sucrose metabolic process											0.0447	
4.05E-03												GO:0050794	regulation of cellular process												

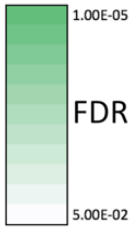


Figure 7. Categories of rhythmic orthologous genes in *Arabidopsis thaliana* and *Kalanchoë fedtschenkoi.*

Group 1 refers to ortholog groups containing at least one rhythmic genes in each of the two plant species (*A. thaliana* and *K. fedtschenkoi*). Group 2 refers to ortholog groups containing at least one rhythmic genes in *A. thaliana* ortholog whereas all the *K. fedtschenkoi* orthologs are arrhythmic. Group 3 refers to ortholog groups containing at least one rhythmic genes in *K. fedtschenkoi* ortholog whereas all the *A. thaliana* orthologs are arrhythmic. Group 4 refers to ortholog groups containg at least one rhythmic genes in *A. thaliana* ortholog but no orthologs in *K. fedtschenkoi*. Group 5 refers to ortholog groups containing at least on rhythmic genes in *K. fedtschenkoi* ortholog but no *A. thaliana* orthologs in *A. thaliana*. Groups are designated by their group names and separated by red dotted lines. Vertical numbers are the number of genes in each group from each plant species.

Figure 7. Continued.

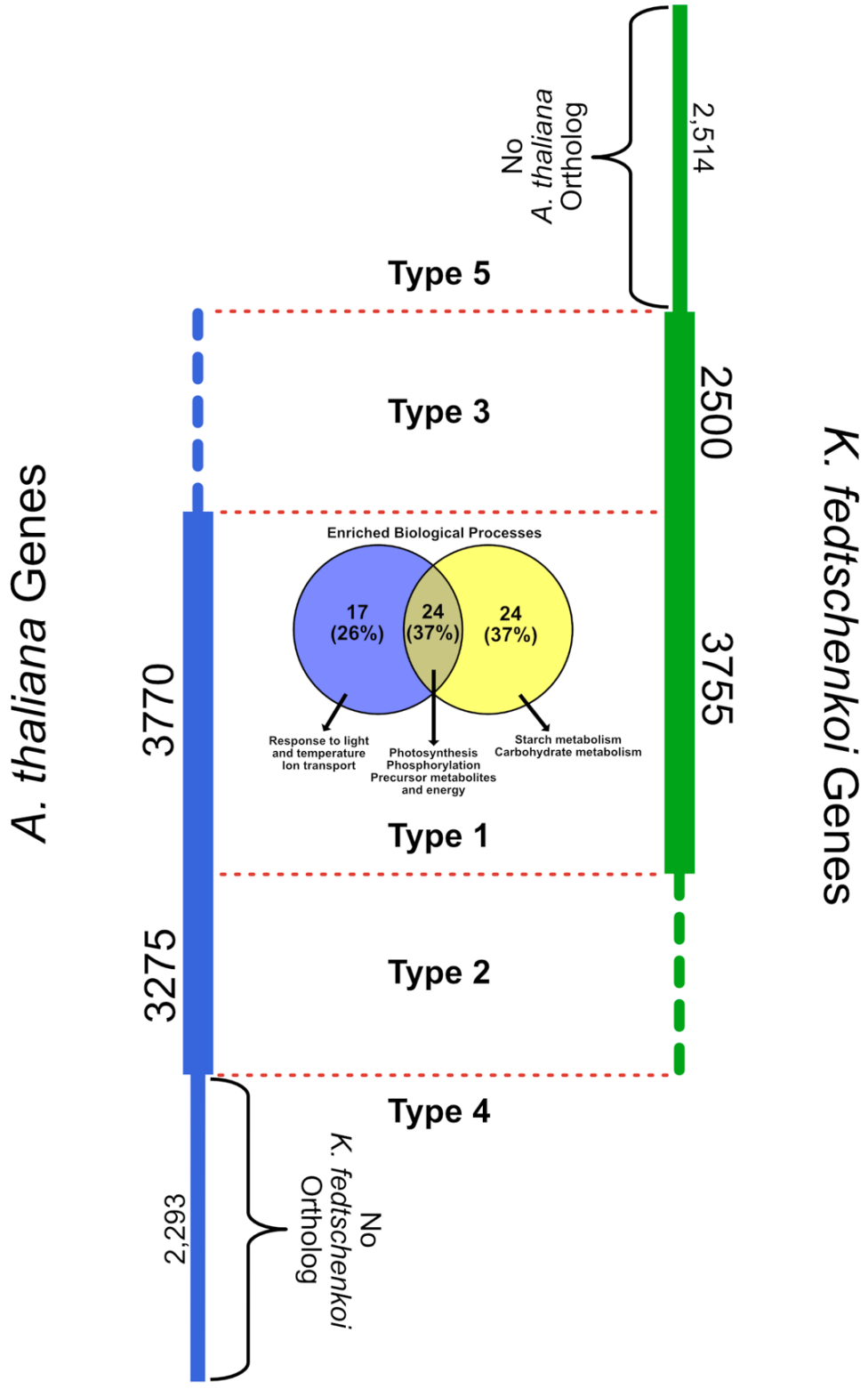
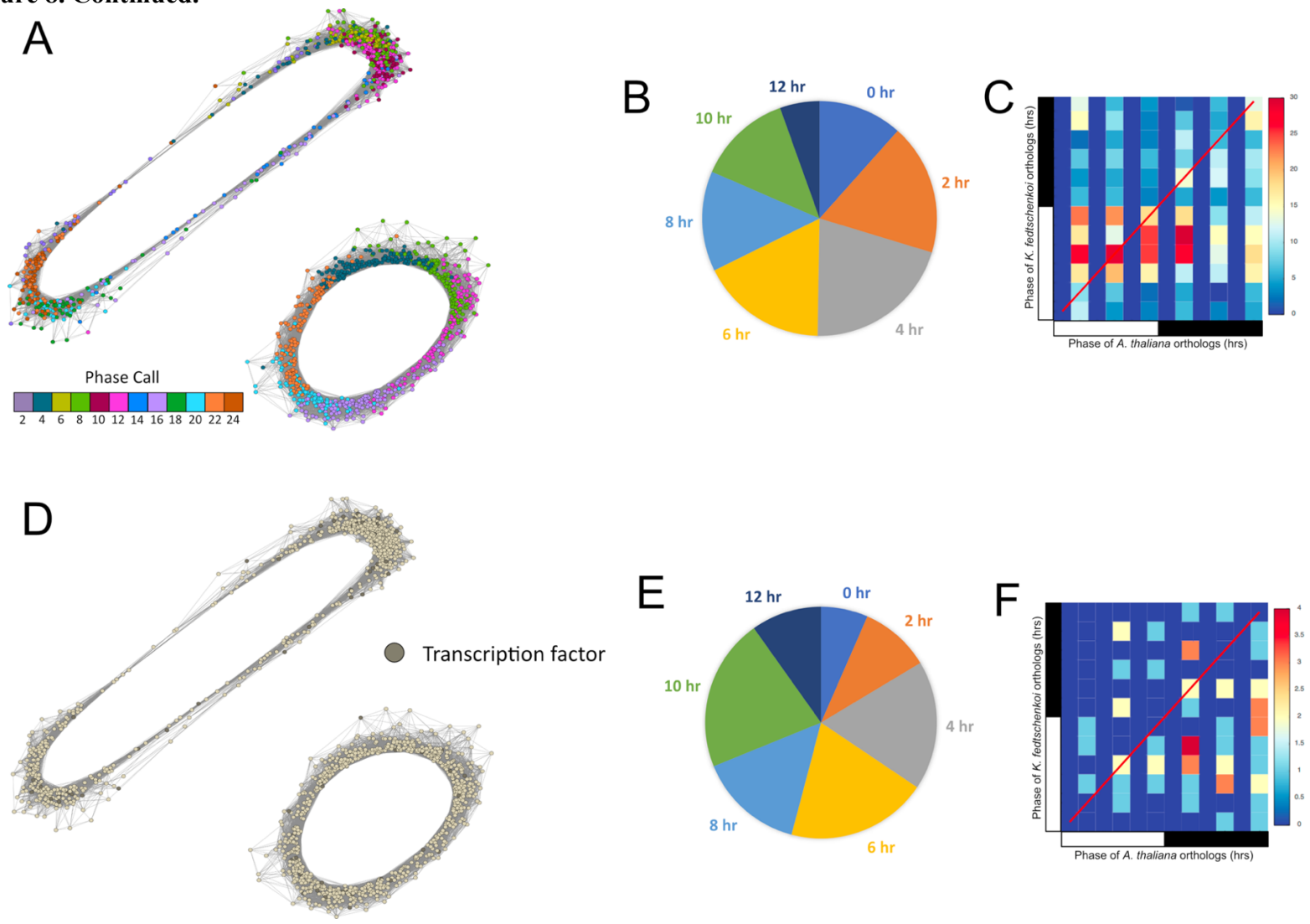


Figure 8. Gene co-expression networks of Type 1 ortholog groups containing only one rhythmic gene in each of the two plant species (*Arabidopsis thaliana* and *Kalanchoë fedtschenkoi)*.

A) Circular gene co-expression networks of rhythmic *A. thaliana* (top) and *K. fedtschenkoi* (bottom) orthologs. Genes are colored according to their phase call. **B)** Pie chart categorizing the different phase shifts between *A. thaliana* and *K. fedtschenkoi* orthologs. **C)** Heat map depicting the phase calls of orthologs between *A. thaliana* (x-axis) and *K. fedtschenkoi* (y-axis). The red line represents positions of identical phases between both species. The black and white bars indicated nighttime (12 h) and daytime (12 h), respectively. **D)** Genes annotated as transcription factors in *A. thaliana* (top) and *K. fedtschenkoi* (bottom). **E)** Pie chart categorizing the different phase shifts between the transcription factor orthologs in *A. thaliana* and *K. fedtschenkoi*. **F)** Heat map depicting the phase calls of transcription factor orthologs between *A. thaliana* (x-axis) and *K. fedtschenkoi* (y-axis). The red line represents positions of identical phases between both species. The black and white bars indicated nighttime (12 h) and daytime (12 h), respectively.

Figure 8. Continued.



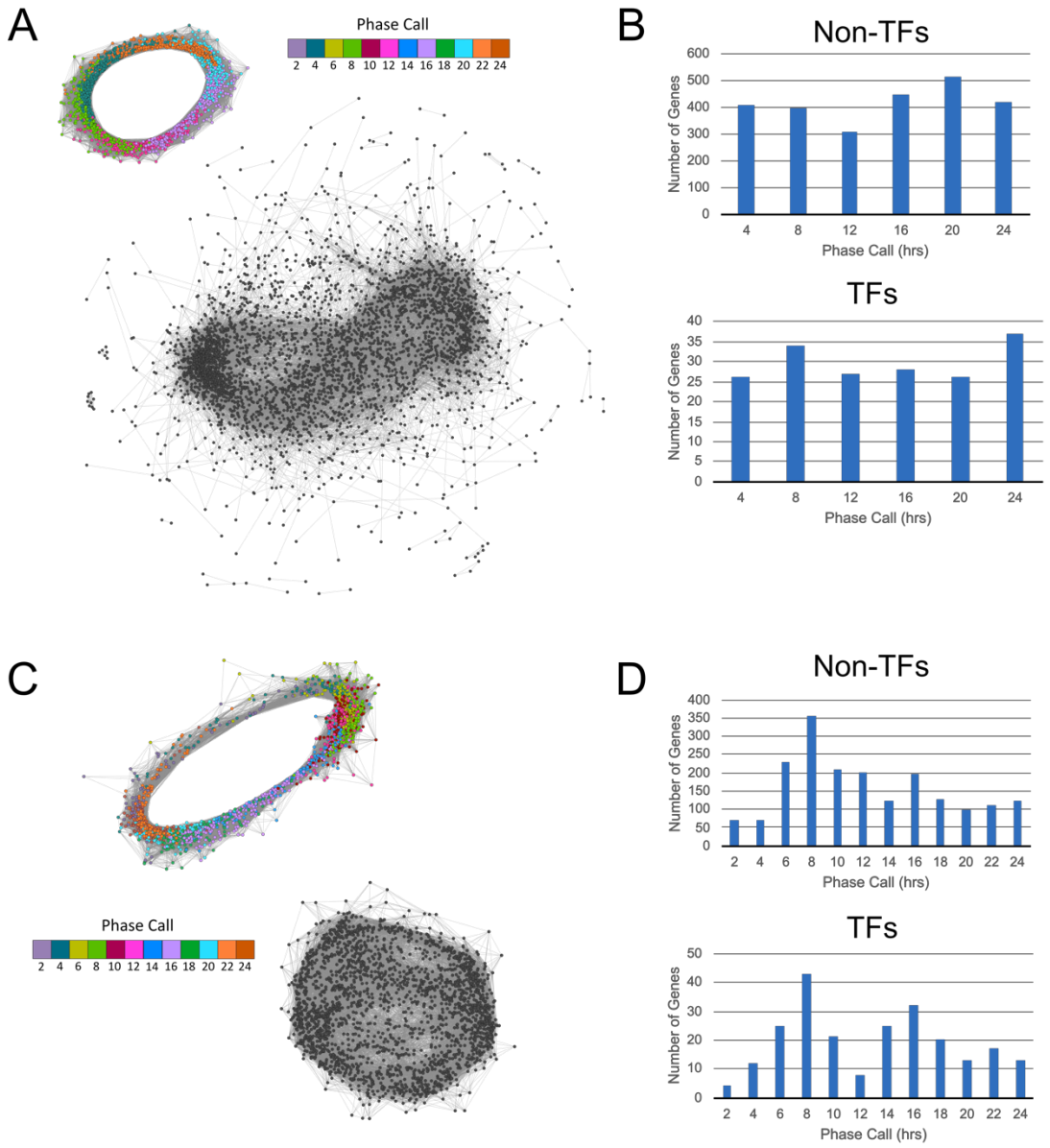


Figure 9. Gene co-expression networks and phase call distributions of Type 2 and Type 3 ortholog groups described in Figure 8.

A) Circular gene co-expression networks of rhythmic genes in *A. thaliana* (top) and arrhythmic genes in *K. fedtschenkoi* (bottom). Genes are colored according to their phase call. B) Phase call distributions of rhythmic *A. thaliana* non-TF genes (top) and rhythmic *A. thaliana* TF genes (bottom). C) Circular gene co-expression networks of rhythmic genes in *K. fedtschenkoi* (top) and arrhythmic genes in *A. thaliana* (bottom) orthologs. Genes are colored according to their phase call. D) Phase call distributions of rhythmic *K. fedtschenkoi* non-TF genes (top) and rhythmic *K. fedtschenkoi* TF genes (bottom)**.**

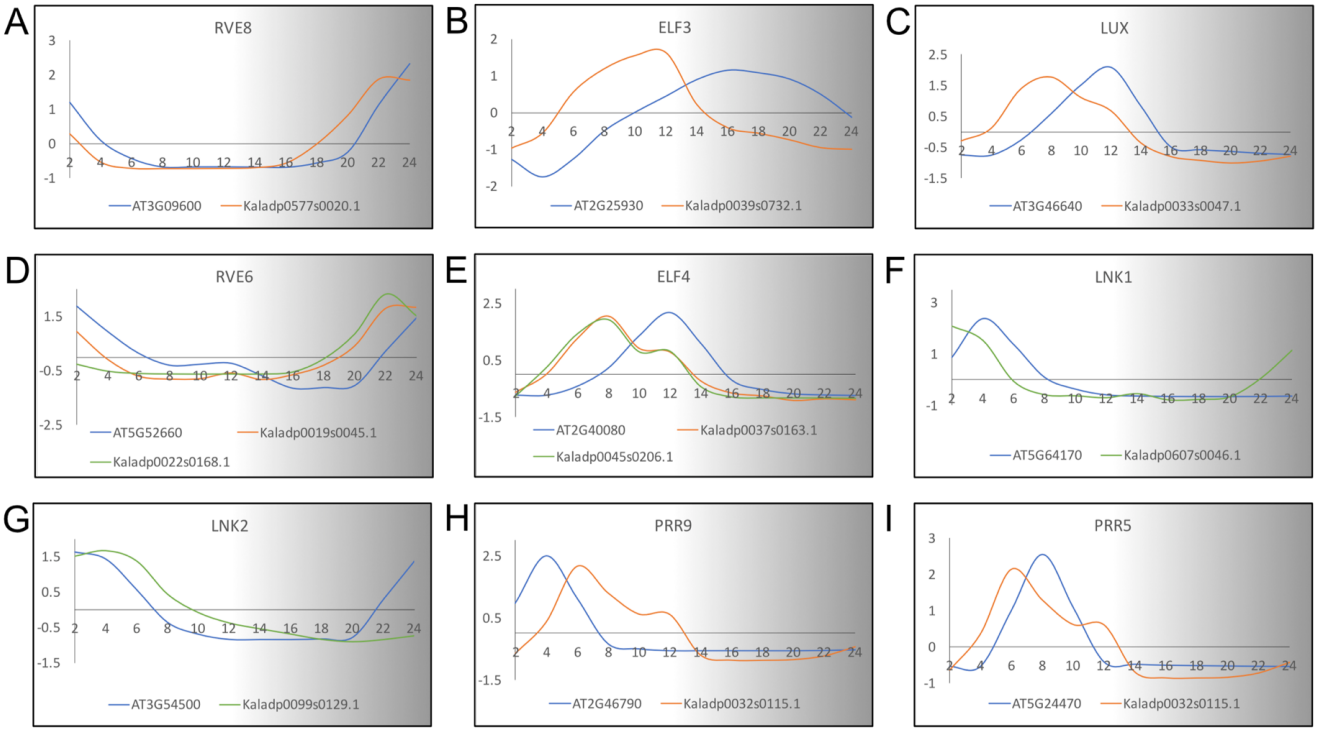


Figure 10. Diurnal expression profiles of *Arabidopsis thaliana* core clock genes and their *Kalanchoë fedtschenkoi* orthologs.

A) Expression of *RVE8* orthologs. **B)** Expression of *ELF3* orthologs. **C)** Expression of LUX orthologs. **D)** Expression of *RVE6* orthologs. **E)** Expression of *ELF4* orthologs. **F)** Expression of *LNK1* orthologs. **G)** Expression of *LNK2* orthologs. The Arabidopsis gene names starts with "AT" and gene names starts with "Ka". X-axis corresponds to number of hours after first light. Time points 2 and 14 correspond to subjective dawn and dusk, respectively. Transcript levels are depicted as a z-score change relative to mean expression for each gene, where values represent the number of standard deviations away from the mean.

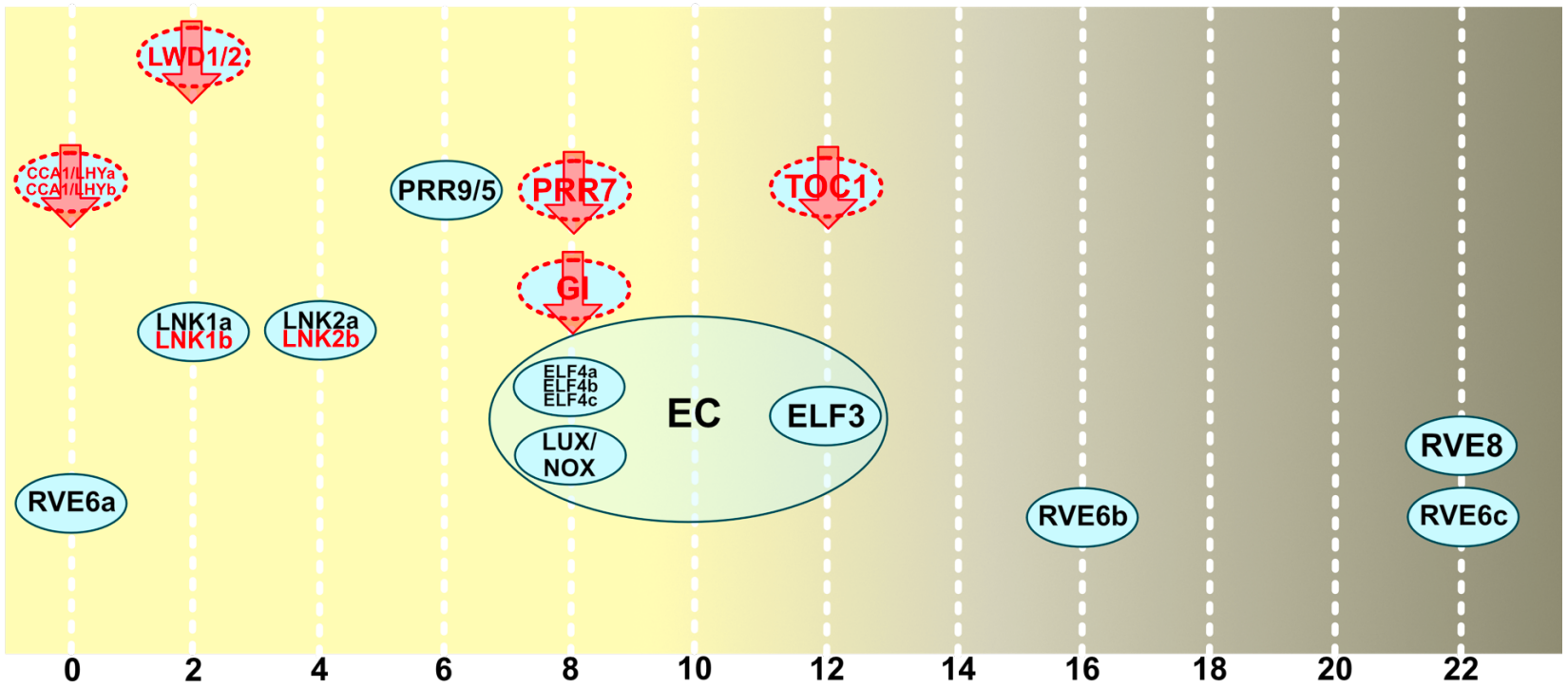


Figure 11. Phase of day for core clock genes in *Kalanchoë fedtschenkoi.*

Ovals represent core clock genes. Red text and arrows indicate the gene does have either an absolute amplitude ≥10, fold-change ≥2, and/or a JTK-CYCLE p-value ≤0.06. Genes with a red dotted outline indicates that the timing of their expression, as indicated here, is based on the phase call of their *A. thaliana* ortholog. The numbers arrayed at the bottom of the image indicate the number of hours passed after first light. EC: Evening complex**.**

CHAPTER III
COMPARATIVE GENOMICS ANALYSIS PROVIDES NEW INSIGHT INTO MOLECULAR BASIS OF STOMATAL DEVELOPMENT AND MOVEMENT IN *KALANCHOË FEDTSCHENKOI*

This chapter is derived from a manuscript currently in preparation: **Moseley R**, Yang X. Comparative genomic analysis provides new insight into molecular basis of stomatal development and movement in *Kalanchoë fedtschenkoi.*

All work was conducted by Robert C. Moseley as well the writing of the manuscript. Dr. Xiaohan Yang provided guidance and feedback on data analysis and manuscript preparation.

Abstract

CO_2 uptake and water loss in plants are regulated by microscopic pores on surface of leaves, called stomata. This enablement of gas exchange by the opening and closing of stomata is one of the most essential processes in plant photosynthesis and transpiration, affecting water-use efficiency (WUE) and thus drought susceptibility. In plant species with crassulacean acid metabolism (CAM) photosynthesis, diel stomatal movement pattern is inverted relative to C_3 and C_4 photosynthesis species, resulting in much higher WUE and drought tolerance. However, little is known about the molecular components involved in stomatal movement and the mechanisms that confer this inversion in CAM species. The goal of this study is to identify candidate genes that could play a role in inversion of day/night stomatal movement pattern in an obligate CAM species, *Kalanchoë fedtschenkoi.* By way of a text-mining approach, proteins were identified in various plant species, spanning C_3, C_4, and CAM photosynthetic types, which are orthologous to proteins known to be involved in stomatal movement or stomatal development. A comparative analysis of diel gene expression time-course data was performed between a *Kalanchoë fedtschenkoi* and two C_3 species (*Arabidopsis thaliana, Solanum lycopersicum*) to identify differential gene expression between the dusk and dawn phases of the 24-hour cycle. A rescheduled catalase gene known to be involved in stomatal movement was identified, suggesting a role for H_2O_2 in CAM-like stomatal movement. Overall, these results provide new insights into the molecular regulation of stomatal development and movement in CAM plants, facilitating genetic improvement of drought tolerance in agricultural crops through manipulation of stomata-related genes.

Introduction

Crassulacean acid metabolism (CAM) is a plant adaptation that involves a carbon concentrating mechanism that is based on a temporal separation of CO_2 fixation (Ehleringer and Monson 1993). A key mechanism facilitating this is the diurnal inversion of stomatal opening and closing relative to the stomatal movement pattern typically seen in C_3 species (Males and Griffiths 2017). Specifically, CAM species open their stomata during the night allowing for uptake of atmospheric CO_2. Combined with internal CO_2, the atmospheric CO_2 is converted into malate, which is subsequently stored in the vacuole. During the onset of day, malate is released from the vacuole and re-fixed into CO_2. In combination with this, stomata begin to close and subsequently trap the released CO_2 within the leaf, allowing for CO_2 to be concentrated around Rubisco, for normal photosynthetic processes (Ehleringer and Monson 1993). Additionally, the inversion of stomatal movement also decreases evapotranspiration in the plant, which gives CAM

plants a drought avoidance/tolerance mechanism. Engineering of these traits into non-CAM species has brought increased attention to understanding the molecular mechanisms behind CAM, and more specifically, behind stomatal movement in CAM (Borland et al. 2014; Yang et al. 2015) (Figure 12A).

Aside from stomata research in CAM, research on stomata in general has recently increased (Figure 12B). In all plants, stomata play essential roles in controlling water losses caused by transpiration and CO_2 uptake for photosynthesis, modulating the transpiration-driven water flow through the soil-plant-atmosphere continuum, and plant adaptation to changing environmental conditions and to stress (Damour et al. 2010). Fully understanding stomatal processes, such as development, movement, and patterning, across various species, can facilitate engineering efforts to improve these traits. However, two issues today hinder the pace of advancement in stomatal research: 1) most work on these stomatal processes has been conducted in the model C_3 plant *Arabidopsis thaliana*. This can present an issue when wanting to transfer the knowledge gained from these studies to another species as the genes involved in these processes can vary in regulation and/or function, as seen in CAM species (Abraham et al. 2016; Males and Griffiths 2017; Yang et al. 2017) and in the grass species *Brachypodium distachyon* (Raissig et al. 2016; Raissig et al. 2017). And 2) large-scale genome sequencing has identified millions of protein-coding genes; however, the rate of discovery has surpassed the rate at which the protein information can be curated (Chang et al. 2016). Homology between two protein sequences can help infer protein function as proteins (Clark and Radivojac 2011). The issue is many proteins' functional information is missing from annotation databases or is hidden in the scientific literature. For instance, a protein of interest could be homologous to a protein characterized to function in a stomata-related process, but this information is not known because the characterized protein lacks the gene ontology annotation or the study it was involved in is difficult to find or the information is buried in the large volume of publications.

Various tools today present alternatives to homology-based searches, such as text-mining tools, e.g., (Hoffmann and Valencia 2005; Mandloi and Chakrabarti 2015), that could aid in stomatal research. These tools enable a researcher to find articles about a protein of interest and determine its function by "reading" the articles themselves. However, many of these tools focus on model organisms and do not solve the first issue of furthering stomatal research. There are text-mining tools that when given a protein of interest, search for information about similar proteins by combining BLAST searches with links to articles from certain databases (Gilchrist et al. 2008; Li et al. 2013; Jaroszewski et al. 2014). However, these tools do not search the literature, therefore, limiting their analysis. A new web-based tool called PaperBLAST (Price and Arkin 2017), that combines BLAST searches and text-mining of linked articles, was recently published that can help alleviate the two issues facing stomatal research described earlier, and therefore aid in discovering molecular mechanisms behind inverted stomatal movement seen in CAM species.

The aim of this study is to first aid the broader stomatal research community by identifying stomata-related genes that contain no stomata-related annotations or are not known as key stomata genes in 13 plant species that either conduct C_3, C_4, or CAM

photosynthesis through utilization of PaperBLAST and each species' respective proteins. Next, this data will be used to identify any genes that could be involved in the inverted stomatal movement seen in CAM plants through a comparative analysis of time-course gene expression data between a CAM species (*Kalanchoe fedtschenkoi*) and two C_3 species (*A. thaliana* and *Solanum lycopersicum*). The gene catalogues generated for each plant species will provide a resource for future experimental work in stomatal research of non-model species, and more specifically, provide gene targets for engineering CAM-like traits into non-CAM species.

Methods

Ortholog groups and phylogenetic analysis

Construction of ortholog groups between 26 plant species was taken from (Yang et al. 2017). Ortholog group construction used the proteomes of 26 plant species and the methodology and species used are described in (Yang et al. 2017). Phylogenetic trees were constructed from protein sequence alignments using the web-based MUSCLE program (https://www.ebi.ac.uk/Tools/msa/muscle/) with default parameters (Katoh and Standley 2013; McWilliam et al. 2013; Li et al. 2015). Trees were generated using the web-based IQ-Tree program (http://www.iqtree.org/) with default parameters (Trifinopoulos et al. 2016).

Gene ontology analysis of protein sequences

Gene ontology terms for *A. thaliana* and *K. fedtschenkoi* genes were taken from the gene annotation information downloaded from Phytozome v12.1 (Goodstein et al. 2012). Using ClueGO (Bindea et al. 2009), the GO biological processes were subjected to the right-sided hypergeometric enrichment test at medium network specificity selection and p-value correction was performed using the Holm-Bonferroni step-down method (Holm 1979). There was a minimum of 3 and a maximum of 8 selected GO-tree levels, while each cluster was set to include a minimum of between 3% and 4% of genes associated with each term. GO-term fusion and grouping settings were selected to minimize GO-term redundancy and the term enriched at the highest level of significance was used as the representative term for each functional cluster. The GO terms with p-values less than or equal to 0.05 were considered significantly enriched.

Data-mining using paperBLAST

The proteomes of 13 plant species including *Amborella trichopoda, Ananas comosus*, *A. thaliana*, *B. distachyon*, *K. fedtschenkoi*, *Mimulus guttatus*, *Musa acuminate*, *Oryza sativa*, *Phalaenopsis equestris*, *Setaria italica*, *Solanum lycopersicum*, *Sorghum bicolor*, and *Vitis vinifera* were taken from PLAZA 4.0 (Van Bel et al. 2017). Each protein sequence was used as a search query in PaperBLAST (Price and Arkin 2017) using an E value threshold of <1e-3. The subsequent HTML files were collected and parsed through searching for any presence of keywords (i.e., 'stomata' and 'guard cell') in each PaperBLAST hit's affiliated data using in-house python scripts. The relevant extracted data for each species are summarized in Table S9.

Analysis of time-course gene expression data

The diurnal expression data for *K. fedtschenkoi* were taken from (Yang et al. 2017). Whole leaf tissue samples of *K. fedtschenkoi* were collected from plants grown in a growth chamber under a 12-h light/12-h dark cycle and collected every 2 hours. The diurnal expression data for *A. thaliana* were taken from (Mockler et al. 2007). Whole-leaf samples of *A. thaliana* were collected from plants grown in a growth chamber under a 12-h light/12-h dark cycle and collected every 4 hours. Raw sequencing reads for *S. lycopersicum* were taken from the DDBJ Sequence Read Archive (http://trace.ddbj.nig.ac.jp/DRASearch) under the accession numbers DRA003529 and DRA0035530. Whole-leaf samples of *S. lycopersicum* were taken every two hours in duplicates from plants grown in a sunlight-type factory under ambient light conditions, resulting in a 10-h light/14-h dark cycle (Higashi et al. 2016). *S. lycopersicum* raw sequencing reads were quality checked by FastQC (Andrew 2014), trimmed using Trimmomatic v0.36 (Bogler et al. 2014) with default parameters, and quality checked again by FastQC. The *S. lycopersicum* RNA-Seq reads were mapped to the *S. lycopersicum* iTAG v2.3 reference genome using TopHat2 v1.0.1 (Kim et al. 2013) using default parameters. Read counts were computed using Cufflink v2.2.1 (Trapnell et al. 2012) and fragments per kilobase of transcript per million mapped reads (FPKM) were calculated. Processing of all raw sequencing data was performed using Kbase (Arkin et al. 2016). FPKM data for *S. lycopersicum* were averaged. Genes with an average FPKM less than 0.01 and with less than half of their sampling points having an FPKM equal to zero were filtered out from the *K. fedtschenkoi* and *S. lycopersicum* expression data sets.

Comparative analysis of gene expression

Both *K. fedtschenkoi* and *S. lycopersicum* gene expression data were measured at two-hour intervals; however, *A. thaliana* gene expression data was measured at four-hour intervals. Therefore, the *A. thaliana* data was adjusted to arrive at expression profiles on the same time scale as *K. fedtschenkoi* and *S. lycopersicum* gene expression data. Here, the cubic interpolation algorithm in the pandas Python library (https://pandas.pydata.org/) was used to simulate the *A. thaliana* gene expression data at additional time points so that all three data sets consisted of sampling points spaced every two hours over a 24-h period.

To compare transcript expression between the time-windows dusk and dawn, the same procedure described in (Yang et al. 2017) was followed. The dusk and dawn phases for *K. fedtschenkoi* and *A. thaliana* were 10, 12, and 14 and 22, 24, and 2 h from the starting of the light period, respectively (Figure 13A). For *S. lycopersicum*, which was grown under a slightly different light regime than *K. fedtschenkoi* and *A. thaliana,* the dusk and dawn phases were 8, 10, and 12 and 22, 24, and 2 h, respectively (Figure 13A). For each species, any genes with negative gene expression levels were removed and expression data for each gene was normalized as in (Yang et al. 2017) and a two-column matrix was created, where rows represented genes, one column represented the sum of all transformed dusk time points, and the other column represented the sum of all transformed dawn time points. For each gene, the right-tailed Fisher Exact Test was used to determine if that gene's expression was enriched in dusk, dawn, or neither according to

the contingency tables in (Yang et al. 2017) (Figure 13B). The Fisher Exact Test was performed using fisher_exact function from the scipy Python library (https://www.scipy.org/). The False Discovery Rate (Benjamini and Hochberg 1995) was controlled per species and time-window at a p-value of <0.05. This comparative analysis was perform to identify ortholog groups containing re-scheduled gene expression in CAM species in comparison with C_3 species, in whichthe *S. lycopersicum* and *A. thaliana* gene were enriched in the same time-window whereas the *K. fedtschenkoi* gene was enriched in the opposite time-window.

Results

Identification of stomata-related genes in 13 plant species

The proteins of 13 plant species were used to search the PaperBLAST database (Table 6). PaperBLAST is a tool that given a protein sequence, finds homologous proteins using protein-protein BLAST and then uses the IDs (e.g., UniProt accessions and RefSeq protein identifiers) of the homologous proteins to perform text searches of various databases (Price and Arkin 2017). The output of each PaperBLAST search contains various information on the homologous proteins, such as functional information, article name, and text snippets from articles containing the homologous protein's IDs. For each protein of the 13 species used to search the PaperBLAST database an output file was created, and text was searched for the key words "stomata" and "guard cell". Any homologous proteins that did not contain the keywords in their related data were filtered out. For each species, this resulted in an average of 18,546 times per species' protein matched with a protein in the PaperBLAST database (Table 7). Redundancy is to be expected as one protein can match to more than one protein in terms of sequence similarity. Unique proteins for each species that matched with a protein in the PaperBLAST database that contained at least one of the keywords was filtered out resulting in an average of 5,196 proteins that could be involved in stomata-related processes for each species. On average, these genes account for 16.4% of the respective plant genome.

A total of 321 proteins in the PaperBLAST database had data containing the keywords 'stomata' and 'guard cell' (Table 8). Most of these proteins (304) were *A. thaliana* proteins, with the second most (9) belonging to *S. lycopersicum*. Only *A. thaliana* and *S. lycopersicum* had proteins that matched to all 321 proteins in the PaperBLAST database (Table 7). Each protein in PaperBLAST is linked to a publication(s) that contains information on the respective protein. 272 publications were catalogued containing at least one of the 321 proteins (Table 7). Since a majority of the 321 proteins were *A. thaliana* proteins and that *A. thaliana* is a well-annotated organism, annotations of the 304 *A. thaliana* proteins were examined. Using the current annotation of *A. thaliana* proteins, 178 proteins were annotated with stomata- or guard cell-related terms. Additional genes known to play key roles in stomatal movement and development were added as well (Ohashi-Ito and Bergmann 2006; MacAllister et al. 2007; Pillitteri et al. 2007; Kanaoka et al. 2008; Kollist et al. 2014), totaling 198 proteins (Table 9). 97 of the 198 proteins were found within the 304 *A. thaliana* proteins found within

PaperBLAST, resulting in 207 stomata-related proteins in *A. thaliana* (Table 10 and Figure 14). For clarity, these 207 stomata-related genes will be referred to as new stomata-related genes/proteins from here on. Gene ontology terms overrepresented in these 207 new stomata-related proteins were signal transduction, phosphorylation-related processes, and several response processes (Figure 15).

PaperBLAST uses an E value threshold <1e-3 to infer homology between two proteins, and generally this E value cutoff can be used reliably for this purpose (Pearson 2013). However, to produce a more confident set of stomata-related genes for each species, the ortholog groups (OGs) of the 321 proteins linked to stomata-related processes in PaperBLAST were extracted. In the construction of these OGs, (Yang et al. 2017) used a more stringent E value threshold of <1e-5 to infer homology and included proteins from 26 plant species. 257 OGs were extracted. Individually, the number of proteins from each species that shared the same OG as the PaperBLAST protein ranged from just over 500 proteins in *A. trichopoda* to over 1,200 proteins in both *A. thaliana* and *M. acuminate* (Table 11). Further categorization of these OGs on whether the respective species' proteins were orthologous to an *A. thaliana* protein in any of the three stomata gene categories (i.e., annotated as stomata- or guard cell-related, known key stomatal genes, or new stomata-related genes) was performed next with results shown in Table 12.

Differential enrichment of gene expression in dusk and dawn

It is hypothesized that the rewiring of gene expression played a role in the evolution of CAM plants from C_3 plants, where genes that are expressed highly during one phase of the day in C_3 species are expressed in a different phase of the day in CAM species (Yang et al. 2017). To determine if any of the *K. fedtschenkoi* genes orthologous to an *A. thaliana* gene in any of the three stomata gene categories could play a role in the inversion of stomatal movement, their expression profiles were compared to their ortholog's expression profiles in *S. lycopersicum* and *A. thaliana*. Specifically, the gene expression profiles of 188 OGs of genes from all three species with an FPKM >0.01 at six or more time-points were compared. These 188 OGs were selected as they contained stomata-related genes identified from the previous section in *K. fedtschenkoi*, *S. lycopersicum*, and *A. thaliana* (Figure 16).

Two criteria were used to identify stomata-related genes that display differential gene expression between dawn and dusk phases of the day between a CAM species (*K. fedtschenkoi*) and two C_3 species (*S. lycopersicum* and *A. thaliana*): 1) The *K. fedtschenkoi* gene must be significantly enriched in either dawn or dusk, whereas the *S. lycopersicum* and *A. thaliana* orthologs must be significantly enriched in the opposite phase of the day, and 2) the *K. fedtschenkoi* gene must have a spearman rank correlation coefficient <-0.6 with both the *S. lycopersicum* and *A. thaliana* orthologs, whereas the *S. lycopersicum* and *A. thaliana* orthologs must have a Spearman rank correlation coefficient >0.6 among themselves. 16 OGs were identified containing orthologs with gene expression adhering to these criteria (Table 13 and Figure 17A). One OG contained *A. thaliana* genes annotated as stomata-related, CIPK23 (AT1G30270) and 1 OG contained an *A. thaliana* gene known to play a role in a stomata-related process, RCAR3

(AT5G53160) (Table 13). The remaining 14 OGs contained *A. thaliana* genes that are not annotated or known as stomata-related but have been reported to be involved in a stomata-related process, or in other words, new stomata-related genes.

Comparison of studies

Two previous studies investigated re-scheduling of genes expression across CAM and C_3 species (Abraham et al. 2016; Yang et al. 2017); however, at the whole-genome level, used slightly different metrics and criteria for defining gene expression differences between orthologs, and either included other plant species or did not use some of the plant species used here. However, the *A. thaliana* gene expression data used here were used in both studies, and the *K. fedtschenkoi* gene expression data used here were used in (Yang et al. 2017). To determine if any overlap in genes identified here overlapped with the two previous studies, results of all three studies were compared. In (Yang et al. 2017), 54 ortholog pairs in two CAM species (*K. fedtschenkoi* and *Ananas comosus*) were found to have convergent gene expression shifts relative to their ortholog in a C_3 species (*A. thaliana*). Of these 54 gene groups, two gene groups were found to have the *A. thaliana* gene annotated or known as a stomata-related gene and 5 gene groups were found to be new stomata-related genes (Table 14). No overlap of genes found to be shifted in gene expression between CAM and C_3 was found between this study and (Yang et al. 2017). In (Abraham et al. 2016), 641 reciprocal best BLAST hit (RBBH) gene pairs between a CAM species (*Agave americana*) and a C_3 species (*A. thaliana*) were found to have reciprocal gene expression, which was inferred as orthologs having undergone a re-scheduling in gene expression relative to each other. Of these 641 RBBHs, 18 had the *A. thaliana* gene annotated or known as a stomata-related gene, and 9 RBBHs were found to be new stomata-related genes (Table 14). An overlap of 2 genes was found between this study and (Abraham et al. 2016), with one of them being known as a stomata-related gene. Comparing the results of both studies to themselves revealed no overlap in genes that displayed re-rescheduling of gene expression between orthologs in a CAM species and a C_3 species (Table 14).

Phylogenetic investigation

The 3 species studied here have undergone multiple whole-genome duplications which can cause for multiple copies of one gene to occur within a gene family, otherwise known as paralogs. The occurrence of paralogs within gene families can introduce ambiguity when trying to perform cross-species analysis of ortholog gene expression. Therefore, phylogenetic trees were constructed for each of the 16 ortholog groups using the protein sequences from only the three species to further refine homology and to better identify a change in expression pattern between closely related orthologs (Figure S2). Using the phylogenetic trees, the 16 OGs were further filtered based on if the genes listed in Table 13 for each OG were at most two speciation events from each other. This resulted in only one OG being identified which contained *A. thaliana* genes annotated as catalases (Figure 17B).

Three genes from both *A. thaliana* and *S. lycopersicum* were located in the catalase OG which is consistent with literature (Mhamdi et al. 2012), however, 6 *K.*

fedtschenkoi genes were present. The *A. thaliana* gene listed in Table 13 is annotated as catalase 2 (CAT2). CAT2 is a part of the photorespiratory pathway and aids in the detoxification of H_2O_2 (Queval et al. 2007; Mhamdi et al. 2010; Mhamdi et al. 2012). Additionally, CAT2 has been linked to abscisic acid (ABA) signaling in stomatal closure as a negative regulator as ABA induces stomatal closure through H_2O_2 production associated with cytosolic free Ca^{2+}concentration oscillation (Jannat et al. 2011). Furthermore, in line with the timing of stomatal opening, CAT2 gene expression peaks in the morning. Lastly, CAT2 gene expression is controlled by the circadian clock (Zhong et al. 1994; Zhong and McClung 1996). The CAT2 clade in the catalase OG phylogenetic tree (Figure 17B) includes all the genes listed in Table 13 for this OG, but also includes another *K. fedtschenkoi* (Kaladp0052s0025), suggesting there are two CAT2 genes in *K. fedtschenkoi*. The *A. thaliana* and *S. lycopersicum* genes have gene expression enriched and peaking in the morning, as well as one of the *K. fedtschenkoi* CAT2 (CAT2.1) genes (Figure 17C). The second *K. fedtschenkoi* gene (CAT2.2) has gene expression enriched and peaking during dusk (Figure 17C). This presents an interesting scenario as the proteins each catalase gene encodes for have relatively specific roles in determining accumulation of H_2O_2 produced through various metabolic pathways (Mhamdi et al. 2012).

Discussion

Research into stomata's role in controlling water loss and gas exchange as well as in facilitating protection against changing environmental conditions and stresses has increased recently, with the purpose of understanding the molecular mechanisms behind these features. However, the functional information generated from these studies is either missing from annotation databases or is hidden in the scientific literature (Price and Arkin 2017), therefore, hindering progress. To help alleviate this problem via a text-mining approach and a comparative genomics approach, this study identified genes that were neither annotated or known as stomata genes across 13 plant species that were cited as being involved in a stomata-related process. To demonstrate the utility of this new resource on stomata-related genes, the molecular mechanism behind the inversion of stomatal movement in CAM species, relative to C_3 species, was investigated using the gene sets generated for one CAM species (*K. fedtschenkoi*) and two C_3 species (*A. thaliana* and *S. lycopersicum*). Several genes were identified as candidates for further investigation into inverted stomatal movement in CAM.

PaperBLAST was used to generate data for each protein of each of the 13 plant species for subsequent filtering based on the key words "stomata" and "guard cell". False positives are expected when using PaperBLAST as articles about genome-wide gene expression often include tables of upregulated and downregulated genes, resulting in the mentioning of many genes in the text snippet presented in PaperBLAST results. In this study, false positives have mostly been remedied due to searching the text snippet given by PaperBLAST for keywords as the listing of genes generally encompasses the whole-text snippet. Manual-curation was used to improve the confidence of identifying new stomata-related proteins.

A little over 300 genes were found to have been published on and implicated in stomata-related processes, with a majority of them belonging to *A. thaliana*. Stomata-related genes were identified in the 13 plant species used in this study by identifying evolutionary orthologs in each species of the over 300 genes found with PaperBLAST. The subset of genes for each species were further categorized based on sequence similarity with *A. thaliana* genes found with PaperBLAST that were either annotated or known as stomata-related genes or are new stomata-related genes. The new stomata-related genes reported in this study serve as a resource for future investigations into the molecular mechanisms behind stomata-related processes across all 13 species; however, it must be noted that these genes and their associated functional predictions are based on homology with the protein sequences of *A. thaliana* genes identified as new stomata-related genes. The genes identified in the other 12 plant species (i.e., not *A. thaliana*) could very well serve other functional roles, even though they belong in the same ortholog group. For instance, it has been recently reported that orthologous genes involved in stomatal development and patterning in *Brachypodium distachyon* and *A. thaliana* display divergence in stomata-related function and regulation, even though they are orthologous genes (Raissig et al. 2016; Raissig et al. 2017). Additionally, there could be genes in each of the other 12 plant species (i.e., not *A. thaliana*) not identified here that could be involved in stomatal development, movement, or patterning that do not exist in *A. thaliana*.

Results, from previous studies examining gene expression rescheduling between CAM and C_3 species (Abraham et al. 2016; Yang et al. 2017), were re-analyzed to determine if any new stomata-related genes were identified in each study and to determine if there was any overlap among them and this study. Overlaps existed pairwise between some studies yet no overlap of orthologs displaying rescheduling of gene expression existed between all 3 studies. One reason for this is likely due to each study using slightly different methods and criteria for identifying orthologs and comparing gene expression. Additionally, there is evidence for non-uniform regulation of CAM-like stomatal movement across different CAM species, for instance in malate signaling (Pantoja and Smith 2002; Santelia and Lawson 2016) and blue light signaling (Ceusters et al. 2014; Abraham et al. 2016; Yang et al. 2017). Each study, including this one, used different combinations of CAM and C_3 species and since current literature points to CAM species evolving different mechanistic routes for stomatal movement, an overlap among all 3 studies is not surprising.

Identifying orthologs across multiple species to infer protein function is difficult, particularly in plant species due to the multiple genome-wide duplication events common among plants. Comparative analyses of gene expression dynamics between orthologs is made even more complicated due to this feature of plant evolution. The two previous studies investigated expression changes between orthologs to identify the molecular components involved in the evolution of CAM from C_3 photosynthesis (Abraham et al. 2016; Yang et al. 2017) and used different approaches to reduce false positives, such as only looking at ortholog groups containing one gene from each species (Abraham et al. 2016). Significant insights behind the molecular mechanisms of CAM stomatal movement were produced from both studies, however, neither study investigated the

phylogenetic relationships within OGs. This study examined the phylogenetic relationships within OGs that displayed differential gene expression in dusk and dawn between orthologs of different species. One OG was identified which contained the catalase genes from all three species, with the *K. fedtschenkoi* CAT2 genes representing high-confidence genes that could have played a role in the inversion of stomatal movement seen in *K. fedtschenkoi*.

There are generally three catalases genes in a plant's genome, and each gene is considered functionally conserved between species (Mhamdi et al. 2012). Specific to the CAT2 gene, CAT2 is expressed in photosynthetic mesophyll cells, is involved in photorespiration, and shows day-night rhythms in transcript abundance, with peaks in the morning (Zhong et al. 1994; Zhong and McClung 1996; Queval et al. 2007; Mhamdi et al. 2010; Mhamdi et al. 2012). CAT2's primary role is in H_2O_2 detoxification, specifically H_2O_2 produced as a result of photosynthesis (Mhamdi et al. 2010). Rubisco is a bifunctional enzyme that also catalyzes oxygenation of RuBP, which produces 2-phosphogylycolate as one of its products. This small molecule is not metabolized via the Calvin-Benson cycle, but is dephosphorylated to produce glycolate, which is transported to the peroxisomes. Within the peroxisomes, glycolate is oxidized to glyoxylate using oxygen as a co-factor, which results in H_2O_2 being produced. This production is negatively controlled by CAT2 (Figure 18A).

CAM is a mechanism to reduce photorespiration by reducing the CO_2:O_2 ratio in cells (Figure 18B). In other organisms that have evolved a method to reduce glycolate production the main fate of glycolate is excretion or oxidation to glyoxylate via a mitochondrial dehydrogenase using NAD^+ as the final electron acceptor (Stabenau and Winkler 2005). Whether this is the case in *K. fedtschenkoi* is yet to be determined, however, our results suggest a reduced role for CAT2.1 in *K. fedtschenkoi*. Interestingly, CAT2 seems to be duplicated in *K. fedtschenkoi*, with one copy having conserved gene expression and the other copy being expressed at dawn (Figure 17C). In a study in *A. thaliana*, CAT2 was found to be involved in ABA signaling for stomatal closure as *cat2* mutants had significantly enhanced ABA-induced stomatal closure (Jannat et al. 2011). These authors presented a model where ABA activates respiratory burst oxidases (RBOHs) in the membranes of guard cells which results in rapid production of H_2O_2 in the guard cells (Kwak et al. 2003). The H_2O_2 then activates Ca^{2+} channels causing stomatal closure during the night (Pei et al. 2000; Murata et al. 2001) (Figure 18C). CAT2 plays a negative role in ABA signaling in stomatal closure by detoxifying the guard cells. This results in Ca^{2+} channels not being activated which leaves the stomata open (Figure 18C). Due to the role of CAT2 in stomatal opening, we propose that CAT2.2, expressed during dusk in *K. fedtschenkoi*, could play a role in stomatal opening, by inhibiting ABA-induced stomatal closure by reducing H_2O_2 concentrations (Figure 18D). Several other authors have suggested that H_2O_2 activation of Ca^{2+} channels represents a possible convergent point for multiple stress signaling pathways (Gechev and Hille 2005; Sewelam et al. 2014; Niu and Liao 2016). Functional characterization of both CAT2 genes in *K. fedtschenkoi* is needed to fully understand both of their roles in stomatal opening, as well as if CAT2.2 is circadian regulated. Further determination of the role of H_2O_2 signaling in *K. fedtschenkoi* stomatal movement is needed.

Conclusion

Several new stomata-related genes were identified across multiple plant species that will enable quicker identification of gene targets for stomata-related investigations. Additionally, these new stomata-related genes were used to investigate the molecular mechanism behind the inversion of stomatal movement seen between CAM and C_3 species. Compilation of the results in this study with two other similar studies, identified numerous new stomata-related genes that displayed re-scheduled gene expression between orthologs in various CAM and C_3 species, and provide a starting point for further investigation into the inversion of stomatal movement. Additionally, two CAT2 genes were identified in *K. fedtschenkoi* as high confident molecular components in the inversion and warrant further investigation into their functional roles. Such experiments could involve phenotyping *cat2* mutants in *K. fedtschenkoi* and insertions of the evening-phased *K. fedtschenkoi* CAT2 and its promoter into *A. thaliana*.

References

Abraham PE, Yin H, Borland AM, Weighill D, Lim SD, De Paoli HC, Engle N, Jones PC, Agh R, Weston DJ et al. 2016. Transcript, protein and metabolite temporal dynamics in the CAM plant Agave. *Nat Plants* **2**: 16178.

Andrew S. 2014. FastQC A Quality Control tool for High Throughput Sequence Data.

Arkin AP, Stevens RL, Cottingham RW, Maslov S, Henry CS, Dehal P, Ware D, Perez F, Harris NL, Canon S et al. 2016. The DOE Systems Biology Knowledgebase (KBase). *bioRxiv* doi:10.1101/096354.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B* **57**: 289-300.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**: 1091-1093.

Bogler A, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114-2120.

Borland AM, Hartwell J, Weston DJ, Schlauch KA, Tschaplinski TJ, Tuskan GA, Yang X, Cushman JC. 2014. Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci* **19**: 327-338.

Ceusters J, Borland AM, Taybi T, Frans M, Godts C, De Proft MP. 2014. Light quality modulates metabolic synchronization over the diel phases of crassulacean acid metabolism. *J Exp Bot* **65**: 3705-3714.

Chang YC, Hu Z, Rachlin J, Anton BP, Kasif S, Roberts RJ, Steffen M. 2016. COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res* **44**: D330-335.

Clark W, Radivojac P. 2011. Analysis of protein function and its prediction from amino acid sequence. *Proteins* **79**: 2086-2096.

Damour G, Simonneau T, Cochard H, Urban L. 2010. An overview of models of stomatal conductance at the leaf level. *Plant Cell Environ* **33**: 1419-1438.

Ehleringer JR, Monson RK. 1993. EVOLUTIONARY AND ECOLOGICAL ASPECTS OF PHOTOSYNTHETIC PATHWAY VARIATION. *Annu Rev Ecol Syst* **24**.

Gechev TS, Hille J. 2005. Hydrogen peroxide as a signal controlling plant programmed cell death. *J Cell Biol* **168**: 17-20.

Gilchrist MJ, Christensen MB, Harland R, Pollet N, Smith JC, Ueno N, Papalopulu N. 2008. Evading the annotation bottleneck: using sequence similarity to search non-sequence gene data. *BMC Bioinformatics* **9**: 442.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178-1186.

Higashi T, Tanigaki Y, Takayama K, Nagano AJ, Honjo MN, Fukuda H. 2016. Detection of Diurnal Variation of Tomato Transcriptome through the Molecular Timetable Method in a Sunlight-Type Plant Factory. *Front Plant Sci* **7**: 87.

Hoffmann R, Valencia A. 2005. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics* **21 Suppl 2**: ii252-258.
Holm S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**: 65-70.
Jannat R, Uraji M, Morofuji M, Hossain MA, Islam MM, Nakamura Y, Mori IC, Murata Y. 2011. The roles of CATALASE2 in abscisic acid signaling in Arabidopsis guard cells. *Biosci Biotechnol Biochem* **75**: 2034-2036.
Jaroszewski L, Koska L, Sedova M, Godzik A. 2014. PubServer: literature searches by homology. *Nucleic Acids Res* **42**: W430-435.
Kanaoka MM, Pillitteri LJ, Fujii H, Yoshida Y, Bogenschutz NL, Takabayashi J, Zhu JK, Torii KU. 2008. SCREAM/ICE1 and SCREAM2 specify three cell-state transitional steps leading to arabidopsis stomatal differentiation. *Plant Cell* **20**: 1775-1785.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg S. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**: 1-13.
Kollist H, Nuhkat M, Roelfsema MR. 2014. Closing gaps: linking elements that control stomatal movement. *New Phytol* **203**: 44-62.
Kwak JM, Mori IC, Pei Z, Leonhardt N, Torres M, Dangl J, Bloom R, Bodde S, Jones J, Schroeder JI. 2003. NADPH oxidase AtrbohD and AtrbohF genes function in ROS-dependent ABA signaling in Arabidopsis. *The EMBO Journal* **22**: 2623-2633.
Li W, Cong Q, Kinch L, Grishin N. 2013. Seq2Ref: a web server to facilitate functional interpretation. *BMC Bioinformatics* **14**: 1-7.
Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Res* **43**: W580-584.
MacAllister C, Ohashi-Ito K, Bergmann DC. 2007. Transcription factor control of asymmetric cell divisions that establish the stomatal lineage. *Nature* **445**: 537-540.
Males J, Griffiths H. 2017. Stomatal Biology of CAM Plants. *Plant Physiol* **174**: 550-560.
Mandloi S, Chakrabarti S. 2015. PALM-IST: Pathway Assembly from Literature Mining--an Information Search Tool. *Sci Rep* **5**: 10021.
McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res* **41**: W597-600.
Mhamdi A, Noctor G, Baker A. 2012. Plant catalases: peroxisomal redox guardians. *Arch Biochem Biophys* **525**: 181-194.
Mhamdi A, Queval G, Chaouch S, Vanderauwera S, Van Breusegem F, Noctor G. 2010. Catalase function in plants: a focus on Arabidopsis mutants as stress-mimic models. *J Exp Bot* **61**: 4197-4220.

Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J. 2007. The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis. *Cold Spring Harbor Symposia on Quantitative Biology* **72**: 353-363.

Murata Y, Pei Z, Mori IC, Schroeder JI. 2001. Abscisic Acid Activation of Plasma Membrane Ca2 Channels in Guard Cells Requires Cytosolic NAD(P)H and Is Differentially Disrupted Upstream and Downstream of Reactive Oxygen Species Production in abi1-1 and abi2-1 Protein Phosphatase 2C Mutants. *The Plant Cell* **13**: 2513-2523.

Niu L, Liao W. 2016. Hydrogen Peroxide Signaling in Plant Development and Abiotic Responses: Crosstalk with Nitric Oxide and Calcium. *Front Plant Sci* **7**: 230.

Ohashi-Ito K, Bergmann DC. 2006. Arabidopsis FAMA controls the final proliferation/differentiation switch during stomatal development. *Plant Cell* **18**: 2493-2505.

Pantoja O, Smith JA. 2002. Sensitivity of the plant vacuolar malate channel to pH, Ca2+ and anion-channel blockers. *J Membr Biol* **186**: 31-42.

Pearson W. 2013. An Introduction to Sequence Similarity ("Homology") Searching. In *Current Protocols in Bioinformatics*.

Pei Z, Murata Y, Benning G, Thomine S, Klusener B, Allen G, Grill E, Schroeder JI. 2000. Calcium channels activated by hydrogen peroxide mediate abscisic acid signalling in guard cells. *Nature* **406**: 731-734.

Pillitteri LJ, Sloan D, GBogenschutz N, Torli K. 2007. Termination of asymmetric cell division and differentiation of stomata. *Nature* **445**: 501-505.

Price MN, Arkin AP. 2017. PaperBLAST: Text Mining Papers for Information about Homologs. *mSystems* **2**.

Queval G, Issakidis-Bourguet E, Hoeberichts FA, Vandorpe M, Gakiere B, Vanacker H, Miginiac-Maslow M, Van Breusegem F, Noctor G. 2007. Conditional oxidative stress responses in the Arabidopsis photorespiratory mutant cat2 demonstrate that redox state is a key modulator of daylength-dependent gene expression, and define photoperiod as a crucial factor in the regulation of H2O2-induced cell death. *Plant J* **52**: 640-657.

Raissig MT, Abrash E, Bettadapur A, Vogel JP, Bergmann DC. 2016. Grasses use an alternatively wired bHLH transcription factor network to establish stomatal identity. *Proc Natl Acad Sci U S A* **113**: 8326-8331.

Raissig MT, Matos J, Gil M, Kornfeld A, Bettadapur A, Abrash E, Allison H, Badgley G, Vogel JP, Berry J et al. 2017. Mobile MUTE specifies subsidiary cells to build physiologically improved grass stomata. *Science* **355**: 1215-1218.

Santelia D, Lawson T. 2016. Rethinking Guard Cell Metabolism. *Plant Physiol* **172**: 1371-1392.

Sewelam N, Jaspert N, Van Der Kelen K, Tognetti VB, Schmitz J, Frerigmann H, Stahl E, Zeier J, Van Breusegem F, Maurino VG. 2014. Spatial H2O2 signaling specificity: H2O2 from chloroplasts and peroxisomes modulates the plant transcriptome differentially. *Mol Plant* **7**: 1191-1210.

Stabenau H, Winkler U. 2005. Glycolate metabolism in green algae. *Physiol Plant* **123**: 235-245.

Trapnell C, Robert A, Goff L, Pertea G, Kim D, Kelley D, Pimentel H, Salzberg S, Rinn J, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**: 562-578.

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Res* **44**: W232-235.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Candepoele K. 2017. PLAZA 4.0: an integrative resource for functional evolutionary and comparative plant genomices. *Nucleic Acids Research* **46**: D1190-D1196.

Yang X, Cushman JC, Borland AM, Edwards EJ, Wullschleger SD, Tuskan GA, Owen NA, Griffiths H, Smith JA, de Paoli HC et al. 2015. A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist* **207**: 491-504.

Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, Liu D, Weigel D, Cheol Yim W, Ha J et al. 2017. The Kalanchoë genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications* **8**: 1-15.

Zhong H, McClung C. 1996. The circadian clock gates expression of two Arabidopsis catalase genes to distinct and opposite circadian phases. *Mol Gen Genet* **251**: 196-203.

Zhong H, Young J, Pease E, Hangarter R, McClung C. 1994. Interactions between Light and the Circadian Clock in the Regulation of CAT2 Expression in Arabidopsis. *Plant Physiol* **104**: 889-898.

Appendix

Table 6. Plant species used to in PaperBLAST search.

Species Abbrev.	Species Name	Version	Data Provider	PubMed ID
arth	*Arabidopsis thaliana*	Araport11	Plaza 4.0	27862469
vivi	*Vitis vinifera*	12xMarch2010	Plaza 4.0	17721507
soly	*Solanum lycopersicum*	itag2.4	Plaza 4.0	22660326
migu	*Mimulus guttatus*	v2.0	Phytozome v12	17460045
kafe	*Kalanchoe fedtschenkoi*	v1.1	Phytozome v12	29196618
sobi	*Sorghum bicolor*	v3.1	Plaza 4.0	19189423
seit	*Setaria italica*	v2.2	Plaza 4.0	22580950
brdi	*Brachypodium distachyon*	v3.1	Plaza 4.0	20148030
orsaja	*Oryza sativa ssp. japonica*	v7.0	Plaza 4.0	16100779
anco	*Ananas comosus*	v3.0	Plaza 4.0	26523774
muac	*Musa acuminata*	v1.0	Plaza 4.0	22801500
pheq	*Phalaenopsis equestris*	v1.0	Plaza 4.0	25420146
amtr	*Amborella trichopoda*	JGI v1.0	Plaza 4.0	24357323

Table 7. Results of PaperBLAST search.

The full names of the plants species are listed in Table 6.

Species	Amtr	Anco	Arth	Brdi	Kafe	Migu	Muac	Orsa	Pheq	Seit	Sobi	Soly	Vivi
# of Proteins	26846	27024	27615	34310	30964	28140	36528	42189	29415	34584	34211	34725	26346
Total Filtered Proteins	12205	14724	17537	17464	19563	18914	27326	22094	13800	20780	19081	19798	17812
Total Unique Proteins	3647	4146	5197	4953	5595	5649	6956	5936	3998	5718	5371	5572	4813
% of Genome	0.136	0.153	0.188	0.144	0.181	0.201	0.19	0.141	0.136	0.165	0.157	0.16	0.183
Total Unique Hit Proteins	318	330	318	319	320	319	317	319	318	318	318	321	319
Total Unique Publications	270	268	271	270	271	271	269	271	270	270	270	271	271

Table 8. Species in PaperBLAST database with stomata-related proteins.

Hit Species	Total Count of Hit Genes	Unique Count of Hit Genes
Arabidopsis thaliana	226698	304
Solanum lycopersicum	6531	9
Glycine max	3548	1
Oryza sativa Japonica Group	3283	3
Vitis vinifera	775	2
Zea mays	263	2
Total		321

Table 9. *Arabidopsis thaliana* genes with stomata-related annotations or known as key stomatal genes.

The gene function annotation was obtained from Gene Ontology (GO), and the "known" stomatal genes were obtained from literature.

ID	Description	Symbol	Type.
AT5G46240	1	KAT1	Known
AT1G74710	ADC synthase superfamily protein	EDS16	GO
AT3G57230	AGAMOUS-like 16	AGL16	GO
AT1G25490	ARM repeat superfamily protein	RCN1	GO
AT4G22330	Alkaline phytoceramidase (aPHC)	ATCES1	GO
AT2G46370	Auxin-responsive GH3 family protein	JAR1	GO
AT5G52440	Bacterial sec-independent translocation protein mttA/Hcf106	HCF106	GO
AT1G12480	C4-dicarboxylate transporter/malic acid transport protein	OZS1	GO/Known
AT1G30270	CBL-interacting protein kinase 23	CIPK23	GO
AT5G22920	CHY-type/CTCHY-type/RING-type Zinc finger protein		GO
AT1G15570	CYCLIN A2;3	CYCA2;3	GO
AT5G65420	CYCLIN D4;1	CYCD4;1	GO
AT4G18160	Ca2+ activated outward rectifying K+ channel 6	KCO6	Known
AT1G08860	Calcium-dependent phospholipid-binding Copine family protein	BON3	GO
AT2G17290	Calcium-dependent protein kinase family protein	CPK6	GO/Known
AT2G33380	Caleosin-related family protein	RD20	GO
AT3G11130	Clathrin, heavy chain		GO
AT1G80370	Cyclin A2;4	CYCA2;4	GO
AT2G15430	DNA-directed RNA polymerase family protein	NRPB3	GO
AT5G65590	Dof-type zinc finger DNA-binding family protein		GO
AT1G14350	Duplicated homeodomain-like superfamily protein	FLP	GO
AT1G80133	EPIDERMAL PATTERNING FACTOR-like protein		GO
AT4G37810	EPIDERMAL PATTERNING FACTOR-like protein		GO
AT5G10310	EPIDERMAL PATTERNING FACTOR-like protein		GO
AT4G14723	EPIDERMAL PATTERNING FACTOR-like protein	CLL2	GO
AT5G62230	ERECTA-like 1	ERL1	GO
AT5G07180	ERECTA-like 2	ERL2	GO
AT2G31470	F-box and associated interaction domains-containing protein	DOR	GO
AT1G71010	FORMS APLOID AND BINUCLEATE CELLS 1C	FAB1C	GO
AT2G26300	G protein alpha subunit 1	GP ALPHA 1	GO
AT4G40100	GRAM domain family protein	PRSL1	GO
AT1G33240	GT-2-like 1	GTL1	GO
AT2G18960	H(+)-ATPase 1	HA1	GO
AT3G49260	IQ-domain 21	iqd21	GO
AT2G40220	Integrase-type DNA-binding superfamily protein	ABI4	GO
AT5G43410	Integrase-type DNA-binding superfamily protein		GO
AT2G26650	K+ transporter 1	KT1	GO
AT1G70300	K+ uptake permease 6	KUP6	Known
AT1G75640	Leucine-rich receptor-like protein kinase family protein		GO
AT2G26330	Leucine-rich receptor-like protein kinase family protein	ER	GO
AT4G20940	Leucine-rich receptor-like protein kinase family protein	GHR1	GO/Known
AT1G80080	Leucine-rich repeat (LRR) family protein	TMM	GO
AT2G43790	MAP kinase 6	MPK6	GO
AT3G21220	MAP kinase kinase 5	MKK5	GO
AT4G22790	MATE efflux family protein	RHC1	GO/Known
AT5G05660	NF-X1-type zinc finger protein NFXL2	NFXL2	GO
AT4G24020	NIN like protein 7	NLP7	GO
AT5G03280	NRAMP metal ion transporter family protein	EIN2	GO
AT5G27150	Na+/H+ exchanger 1	NHX1	GO
AT1G10300	Nucleolar GTP-binding protein		GO
AT5G55630	Outward rectifying potassium channel protein	KCO1	GO/Known

Table 9. Continued.

ID	Description	Symbol	Type.
AT1G65480	PEBP (phosphatidylethanolamine-binding protein) family protein	FT	GO
AT4G26690	PLC-like phosphodiesterase family protein	SHV3	GO
AT5G59220	PP2C protein (Clade A protein phosphatases type 2C)	HAI1	GO
AT5G46790	PYR1-like 1	PYL1	Known
AT1G08750	Peptidase C13 family		GO
AT3G58490	Phosphatidic acid phosphatase (PAP2) family protein	SPP1	GO
AT1G09780	Phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent	iPGAM1	GO
AT3G08590	Phosphoglycerate mutase, 2,3-bisphosphoglycerate-independent	iPGAM2	GO
AT2G26670	Plant heme oxygenase (decyclizing) family protein	TED4	GO
AT3G49220	Plant invertase/pectin methylesterase inhibitor superfamily		GO
AT2G16630	Pollen Ole e 1 allergen and extensin family protein		GO
AT2G31870	Poly (ADP-ribose) glycohydrolase (PARG)	TEJ	GO
AT4G17870	Polyketide cyclase/dehydrase and lipid transport superfamily protein	PYR1	Known
AT5G14880	Potassium transporter family protein	KUP8	Known
AT5G40280	Prenyltransferase family protein	ERA1	GO
AT4G14480	Protein kinase superfamily protein	BLUS1	GO/Known
AT1G62400	Protein kinase superfamily protein	HT1	GO/Known
AT4G33950	Protein kinase superfamily protein	OST1	GO/Known
AT5G35410	Protein kinase superfamily protein	SOS2	Known
AT1G63700	Protein kinase superfamily protein	YDA	GO
AT4G26080	Protein phosphatase 2C family protein	ABI1	GO/Key
AT5G57050	Protein phosphatase 2C family protein	ABI2	GO/Key
AT4G28400	Protein phosphatase 2C family protein		Known
AT1G34245	Putative membrane lipoprotein	EPF2	GO
AT1G20090	RHO-related protein from plants 2	ROP2	GO
AT3G09770	RING/U-box superfamily protein	LOG2	GO
AT2G39940	RNI-like superfamily protein	COI1	GO
AT1G17040	SH2 domain protein A	SHA	GO
AT5G55480	SHV3-like 1	SVL1	GO
AT5G24030	SLAC1 homologue 3	SLAH3	Known
AT1G60940	SNF1-related protein kinase 2.10	SNRK2.10	Known
AT1G66340	Signal transduction histidine kinase, hybrid-type, ethylene sensor	ETR1	GO
AT1G04110	Subtilase family protein	SDD1	GO
AT1G20160	Subtilisin-like serine endopeptidase family protein	ATSBT5.2	GO
AT5G46860	Syntaxin/t-SNARE family protein	VAM3	GO
AT2G32950	Transducin/WD40 repeat-like superfamily protein	COP1	GO
AT5G64010	U2 small nuclear ribonucleoprotein auxiliary factor-like protein		GO
AT5G05860	UDP-glucosyl transferase 76C2	UGT76C2	GO
AT5G20090	Uncharacterized protein family	UPF0041	GO
AT2G30370	allergen-like protein	CHAL	GO
AT3G22820	allergen-like protein	CLL1	GO
AT1G25480	aluminum activated malate transporter family protein		GO
AT4G17970	aluminum-activated, malate transporter 12	ALMT12	GO/Known
AT1G05805	basic helix-loop-helix (bHLH) DNA-binding superfamily protein		GO
AT1G51140	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	FBH3	GO
AT3G24140	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	FMA	GO/Known
AT3G26744	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	ICE1	GO/Known
AT3G06120	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	MUTE	GO/Known
AT1G12860	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	SCRM2	Known
AT5G53210	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	SPCH	GO/Known
AT1G70410	beta carbonic anhydrase 4	BCA4	GO/Known
AT1G52400	beta glucosidase 18	BGLU18	GO
AT5G60880	breaking of asymmetry in the stomatal lineage	BASL	GO
AT4G17615	calcineurin B-like protein 1	CBL1	GO
AT5G47100	calcineurin B-like protein 9	CBL9	GO
AT4G37640	calcium ATPase 2	ACA2	Known
AT5G23060	calcium sensing receptor	CaS	GO

Table 9. Continued.

ID	Description	Symbol	Type.
AT4G04720	calcium-dependent protein kinase 21	CPK21	Known
AT4G04740	calcium-dependent protein kinase 23	CPK23	Known
AT4G23650	calcium-dependent protein kinase 6	CDPK6	GO
AT4G33050	calmodulin-binding family protein	EDA39	GO
AT3G01500	carbonic anhydrase 1	CA1	GO/Known
AT3G50360	centrin2	CEN2	GO
AT5G49890	chloride channel C	CLC-C	Known
AT2G21660	cold, circadian rhythm, and RNA binding 2	GRP7	GO
AT5G44790	copper-exporting ATPase	RAN1	GO
AT4G08920	cryptochrome 1	CRY1	GO/Known
AT1G04400	cryptochrome 2	CRY2	GO/Known
AT5G27620	cyclin H;1	CYCH;1	GO
AT5G10440	cyclin d4;2	CYCD4;2	GO
AT3G54180	cyclin-dependent kinase B1;1	CDKB1;1	GO
AT2G38620	cyclin-dependent kinase B1;2	CDKB1;2	GO
AT3G63440	cytokinin oxidase/dehydrogenase 6	CKX6	GO
AT5G23720	dual specificity protein phosphatase family protein	PHS1	GO
AT1G07670	endomembrane-type CA-ATPase 4	ECA4	Known
AT2G20875	epidermal patterning factor 1	EPF1	GO
AT5G58430	exocyst subunit exo70 family protein B1	EXO70B1	GO
AT1G07000	exocyst subunit exo70 family protein B2	EXO70B2	GO
AT1G69530	expansin A1	EXPA1	GO
AT1G62580	flavin containing monooxygenase FMO GS-OX-like protein	NOGC1	GO
AT5G05620	gamma-tubulin complex protein 2	GCP2	GO
AT3G61650	gamma-tubulin	TUBG1	GO
AT5G37500	gated outwardly-rectifying K+ channel	GORK	Known
AT5G25980	glucoside glucohydrolase 2	TGG2	GO
AT2G17260	glutamate receptor 2	GLR2	GO
AT5G02500	heat shock cognate protein 70-1	HSC70-1	GO
AT5G56030	heat shock protein 81-2	HSP81-2	GO
AT4G29130	hexokinase 1	HXK1	GO
AT2G17820	histidine kinase 1	HK1	GO
AT5G10720	histidine kinase 5	HK5	GO
AT2G25930	hydroxyproline-rich glycoprotein family protein	ELF3	GO
AT2G39050	hydroxyproline-rich glycoprotein family protein	EULS3	GO
AT1G80180	hypothetical protein		GO
AT3G59700	lectin-receptor kinase	HLECRK	GO
AT5G54270	light-harvesting chlorophyll B-binding protein 3	LHCB3	GO
AT3G45640	mitogen-activated protein kinase 3	MPK3	GO
AT1G51660	mitogen-activated protein kinase kinase 4	MKK4	GO
AT1G05100	mitogen-activated protein kinase kinase kinase 18	MAPKKK18	GO
AT3G50310	mitogen-activated protein kinase kinase kinase 20	MAPKKK20	GO
AT5G11300	mitotic-like cyclin 3B from Arabidopsis	CYC3B	GO
AT1G16540	molybdenum cofactor sulfurase (LOS5)	ABA3	GO
AT5G08120	movement protein binding protein 2C	MPB2C	GO
AT2G47800	multidrug resistance-associated protein 4	ABCC4	GO
AT1G04120	multidrug resistance-associated protein 5	ABCC5	GO
AT1G57560	myb domain protein 50	MYB50	GO
AT4G01680	myb domain protein 55	MYB55	GO
AT1G08810	myb domain protein 60	MYB60	GO
AT1G09540	myb domain protein 61	MYB61	GO
AT2G02820	myb domain protein 88	MYB88	GO
AT1G69850	nitrate transporter 1:2	NRT1:2	GO
AT5G11270	overexpressor of cationic peroxidase 3	OCP3	GO
AT1G15440	periodic tryptophan protein 2	PWP2	GO
AT2G40180	phosphatase 2C5	PP2C5	GO
AT3G56960	phosphatidyl inositol monophosphate 5 kinase 4	PIP5K4	GO

Table 9. Continued.

ID	Description	Symbol	Type.
AT3G14270	phosphatidylinositol-4-phosphate 5-kinase family protein	FAB1B	GO
AT2G19690	phospholipase A2-beta	PLA2-BETA	GO
AT3G15730	phospholipase D alpha 1	PLDALPHA1	GO
AT4G35790	phospholipase D delta	PLDDELTA	GO
AT2G05070	photosystem II light harvesting complex protein 2.2	LHCB2.2	GO
AT2G20890	photosystem II reaction center PSB29 protein	PSB29	GO
AT3G45780	phototropin 1	PHOT1	GO/Known
AT5G58140	phototropin 2	PHOT2	GO/Known
AT2G18790	phytochrome B	PHYB	GO
AT1G10560	plant U-box 18	PUB18	GO
AT5G06970	plant/protein	DUF810	GO
AT4G20260	plasma-membrane associated cation-binding protein 1	PCAP1	GO
AT2G26770	pectin-like protein	SCAB1	GO
AT4G18290	potassium channel KAT1-like protein	KAT2	GO/Known
AT4G22200	potassium transport 2/3	KT2/3	Known
AT2G28930	protein kinase 1B	PK1B	GO
AT2G42500	protein phosphatase 2A-3	PP2A-3	Known
AT3G11410	protein phosphatase 2CA	PP2CA	GO
AT5G53160	regulatory components of ABA receptor 3	RCAR3	Known
AT1G64060	respiratory burst oxidase protein F	RBOH F	GO
AT4G16110	response regulator 2	RR2	GO
AT3G12280	retinoblastoma-related 1	RBR1	GO
AT3G05030	sodium hydrogen exchanger 2	NHX2	GO
AT4G12970	stomagen	STOMAGEN	GO
AT3G11820	syntaxin of plants 121	SYP121	GO
AT5G26000	thioglucoside glucohydrolase 1	TGG1	GO
AT3G16640	translationally controlled tumor protein	TCTP	GO
AT5G46220	transmembrane protein	DUF616	GO
AT4G03560	two-pore channel 1	TPC1	GO/Known
AT4G32150	vesicle-associated membrane protein 711	VAMP711	GO
AT5G13740	zinc induced facilitator 1	ZIF1	GO
AT5G13750	zinc induced facilitator-like 1	ZIFL1	GO
AT3G43790	zinc induced facilitator-like 2	ZIFL2	GO

Table 10. *Arabidopsis thaliana* genes identified as new stomata-related genes.

The new candidate genes were predicted using PaperBLAST as described in the Methods section.

ID	Description	Symbol
AT4G33240	1-phosphatidylinositol-3-phosphate 5-kinase FAB1A	FAB1A
AT1G76690	12-oxophytodienoate reductase 2	OPR2
AT1G25450	3-ketoacyl-CoA synthase 5	KCS5
AT5G06530	ABC-2 type transporter family protein	ABCG22
AT3G47760	ABC2 homolog 4	ABCA5
AT2G45660	AGAMOUS-like 20	AGL20
AT3G25610	ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein	
AT4G38510	ATPase, V1 complex, subunit B protein	VAB2
AT1G13180	Actin-like ATPase superfamily protein	DIS1
AT4G37610	BTB and TAZ domain protein 5	BT5
AT1G74720	C2 calcium/lipid-binding plant phosphoribosyltransferase family protein	QKY
AT4G25990	CCT motif family protein	CIL
AT2G01830	CHASE domain containing histidine kinase protein	WOL
AT4G29750	CRS1 / YhbY (CRM) domain-containing protein	
AT1G77390	CYCLIN A1;2	CYCA1;2
AT5G67260	CYCLIN D3;2	CYCD3;2
AT3G55470	Calcium-dependent lipid-binding (CaLB domain) family protein	AT3G55470
AT3G14590	Calcium-dependent lipid-binding (CaLB domain) family protein	NTMC2T6.2
AT1G62810	Copper amine oxidase family protein	CuAO1
AT1G49620	Cyclin-dependent kinase inhibitor family protein	ICK5
AT5G08630	DDT domain-containing protein	
AT5G60410	DNA-binding protein with MIZ/SP-RING zinc finger, PHD-finger and SAP domain-containing protein	SIZ1
AT2G35720	DNAJ heat shock N-terminal domain-containing protein	OWL1
AT3G27925	DegP protease 1	DEG1
AT2G46090	Diacylglycerol kinase family protein	LCBK2
AT1G63750	Disease resistance protein (TIR-NBS-LRR class) family	
AT4G29920	Double Clp-N motif-containing P-loop nucleoside triphosphate hydrolases superfamily protein	
AT1G54150	E3 Ubiquitin ligase family protein	SPL2
AT5G42190	E3 ubiquitin ligase SCF complex subunit SKP1/ASK1 family protein	
AT2G25490	EIN3-binding F box protein 1	EBF1
AT5G25350	EIN3-binding F box protein 2	EBF2
AT5G51550	EXORDIUM like 3	EXL3
AT2G41475	Embryo-specific protein 3	ATS3
AT3G52940	Ergosterol biosynthesis ERG4/ERG24 family	FK
AT3G18490	Eukaryotic aspartyl protease family protein	ASPG1
AT3G20015	Eukaryotic aspartyl protease family protein	
AT1G34260	FORMS APLOID AND BINUCLEATE CELLS 1A	FAB1D
AT4G01120	G-box binding factor 2	GBF2
AT5G10450	G-box regulating factor 6	GRF6
AT1G48270	G-protein-coupled receptor 1	GCR1
AT2G20570	GBF's pro-rich region-interacting factor 1	GPRI1
AT1G35160	GF14 protein phi chain	GF14 PHI
AT5G44190	GOLDEN2-like 2	GLK2
AT5G66770	GRAS family transcription factor	
AT1G06390	GSK3/SHAGGY-like protein kinase 1	GSK1
AT1G22690	Gibberellin-regulated family protein	
AT4G21870	HSP20-like chaperones superfamily protein	
AT1G79920	Heat shock protein 70 (Hsp 70) family protein	Hsp70-15
AT3G25690	Hydroxyproline-rich glycoprotein family protein	CHUP1
AT1G31150	K-box region protein	DUF1985
AT5G14760	L-aspartate oxidase	AO
AT4G39400	Leucine-rich receptor-like protein kinase family protein	BRI1
AT5G46330	Leucine-rich receptor-like protein kinase family protein	FLS2

Table 10. Continued.

ID	Description	Symbol
AT3G47570	Leucine-rich repeat protein kinase family protein	
AT1G51800	Leucine-rich repeat protein kinase family protein	IOS1
AT3G18040	MAP kinase 9	MPK9
AT3G10550	Myotubularin-like phosphatases II superfamily	MTM1
AT1G01010	NAC domain containing protein 1	NAC001
AT1G28010	P-glycoprotein 14	ABCB14
AT4G18480	P-loop containing nucleoside triphosphate hydrolases superfamily protein	CHLI1
AT1G64110	P-loop containing nucleoside triphosphate hydrolases superfamily protein	DAA1
AT4G20370	PEBP (phosphatidylethanolamine-binding protein) family protein	TSF
AT3G27400	Pectin lyase-like superfamily protein	
AT3G14205	Phosphoinositide phosphatase family protein	
AT5G20840	Phosphoinositide phosphatase family protein	
AT3G55940	Phosphoinositide-specific phospholipase C family protein	
AT2G06925	Phospholipase A2 family protein	PLA2-ALPHA
AT2G30520	Phototropic-responsive NPH3 family protein	RPT2
AT1G23200	Plant invertase/pectin methylesterase inhibitor superfamily	
AT3G15220	Protein kinase superfamily protein	
AT3G57770	Protein kinase superfamily protein	
AT3G05840	Protein kinase superfamily protein	ATSK12
AT4G18710	Protein kinase superfamily protein	BIN2
AT1G43900	Protein phosphatase 2C family protein	
AT5G66080	Protein phosphatase 2C family protein	
AT5G62880	RAC-like 10	RAC10
AT2G45890	RHO guanyl-nucleotide exchange factor 4	ROPGEF4
AT2G42620	RNI-like superfamily protein	MAX2
AT4G28560	ROP-interactive CRIB motif-containing protein 7	RIC7
AT3G25070	RPM1 interacting protein 4	RIN4
AT5G63860	Regulator of chromosome condensation (RCC1) family protein	UVR8
AT1G75950	S phase kinase-associated protein 1	SKP1
AT1G61370	S-locus lectin protein kinase family protein	
AT4G31120	SHK1 binding protein 1	SKB1
AT2G25700	SKP1-like 3	SK3
AT1G41830	SKU5-similar 6	SKS6
AT1G62280	SLAC1 homologue 1	SLAH1
AT1G69230	SPIRAL1-like2	SP1L2
AT1G48410	Stabilizer of iron transporter SufD / Polynucleotidyl transferase	AGO1
AT1G65370	TRAF-like family protein	
AT1G30290	Tetratricopeptide repeat (TPR)-like superfamily protein	
AT1G19850	Transcriptional factor B3 family protein / auxin-responsive factor AUX/IAA-like protein	MP
AT5G66690	UDP-Glycosyltransferase superfamily protein	UGT72E2
AT3G13550	Ubiquitin-conjugating enzyme family protein	FUS9
AT5G39510	Vesicle transport v-SNARE family protein	SGR4
AT2G46400	WRKY DNA-binding protein 46	WRKY46
AT2G40750	WRKY DNA-binding protein 54	WRKY54
AT3G56400	WRKY DNA-binding protein 70	WRKY70
AT1G29860	WRKY DNA-binding protein 71	WRKY71
AT4G31805	WRKY family transcription factor	
AT1G45249	abscisic acid responsive elements-binding factor 2	ABF2
AT4G34000	abscisic acid responsive elements-binding factor 3	ABF3
AT2G27150	abscisic aldehyde oxidase 3	AAO3
AT3G27000	actin related protein 2	ARP2
AT5G47840	adenosine monophosphate kinase	AMK2
AT5G42650	allene oxide synthase	AOS
AT3G18440	aluminum-activated malate transporter 9	ALMT9
AT3G04080	apyrase 1	APY1

Table 10. Continued.

ID	Description	Symbol
AT5G18280	apyrase 2	APY2
AT4G34710	arginine decarboxylase 2	ADC2
AT5G61440	atypical CYS HIS rich thioredoxin 5	ACHT5
AT5G67110	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	ALC
AT4G00050	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	UNE10
AT3G25710	basic helix-loop-helix 32	BHLH32
AT1G23730	beta carbonic anhydrase 3	BCA3
AT4G33580	beta carbonic anhydrase 5	BCA5
AT1G58180	beta carbonic anhydrase 6	BCA6
AT2G32860	beta glucosidase 33	BGLU33
AT2G39660	botrytis-induced kinase1	BIK1
AT5G55990	calcineurin B-like protein 2	CBL2
AT1G48090	calcium-dependent lipid-binding family protein	
AT1G18890	calcium-dependent protein kinase 1	CDPK1
AT3G51850	calcium-dependent protein kinase 13	CPK13
AT5G19450	calcium-dependent protein kinase 19	CDPK19
AT1G50700	calcium-dependent protein kinase 33	CPK33
AT2G41010	calmodulin (CAM)-binding protein of 25 kDa	CAMBP25
AT5G67380	casein kinase alpha 1	CKA1
AT1G20620	catalase 3	CAT3
AT2G38170	cation exchanger 1	CAX1
AT3G53720	cation/H+ exchanger 20	CHX20
AT3G48750	cell division control 2	CDC2
AT5G40890	chloride channel A	CLC-A
AT5G26570	chloroplastidic phosphoglucan, water dikinase (ATGWD3)	PWD
AT2G34510	choice-of-anchor C domain protein	DUF642
AT1G11210	cotton fiber protein, putative	DUF761
AT5G51020	crumpled leaf	CRL
AT5G57940	cyclic nucleotide gated channel 5	CNGC5
AT2G23980	cyclic nucleotide-gated channel 6	CNGC6
AT5G62430	cycling DOF factor 1	CDF1
AT4G19230	cytochrome P450, family 707, subfamily A, polypeptide 1	CYP707A1
AT5G53290	cytokinin response factor 3	CRF3
AT4G25480	dehydration response element B1A	DREB1A
AT3G55610	delta 1-pyrroline-5-carboxylate synthase 2	P5CS2
AT2G36490	demeter-like 1	DML1
AT4G24570	dicarboxylate carrier 2	DIC2
AT2G36640	embryonic cell protein 63	ECP63
AT4G05110	equilibrative nucleoside transporter 6	ENT6
AT5G49150	gamete expressed 2	GEX2
AT4G09000	general regulatory factor 1	GRF1
AT1G34760	general regulatory factor 11	GRF11
AT1G78220	general regulatory factor 13	GRF13
AT5G38480	general regulatory factor 3	GRF3
AT2G42590	general regulatory factor 9	GRF9
AT1G22770	gigantea protein	GI
AT2G36850	glucan synthase-like 8	GSL8
AT5G61030	glycine-rich RNA-binding protein 3	GR-RBP3
AT5G49720	glycosyl hydrolase 9A1	GH9A1
AT5G20635	guanine nucleotide-binding protein subunit gamma	AGG3
AT1G79930	heat shock protein 91	HSP91
AT1G35515	high response to osmotic stress 10	HOS10
AT5G35750	histidine kinase 2	HK2
AT1G27320	histidine kinase 3	HK3
AT3G29350	histidine-containing phosphotransmitter 2	AHP2
AT1G05230	homeodomain GLABROUS 2	HDG2
AT2G31650	homologue of trithorax	ATX1

Table 10. Continued.

ID	Description	Symbol
AT5G07370	inositol polyphosphate kinase 2 alpha	IPK2a
AT1G54340	isocitrate dehydrogenase	ICDH
AT1G77280	kinase with adenine nucleotide alpha hydrolases-like domain-containing protein	
AT1G15080	lipid phosphate phosphatase 2	LPP2
AT5G13630	magnesium-chelatase subunit chlH, chloroplast, putative / Mg-protoporphyrin IX chelatase, putative (CHLH)	GUN5
AT4G25080	magnesium-protoporphyrin IX methyltransferase	CHLM
AT4G25200	mitochondrion-localized small heat shock protein 23.6	HSP23.6-MITO
AT2G46070	mitogen-activated protein kinase 12	MPK12
AT3G62700	multidrug resistance-associated protein 10	ABCC14
AT5G67300	myb domain protein r1	MYBR1
AT3G03530	non-specific phospholipase C4	NPC4
AT3G23430	phosphate 1	PHO1
AT4G37870	phosphoenolpyruvate carboxykinase 1	PCK1
AT2G46500	phosphoinositide 4-kinase gamma 4	PI4K GAMMA 4
AT1G14280	phytochrome kinase substrate 2	PKS2
AT3G54820	plasma membrane intrinsic protein 2;5	PIP2;5
AT3G53420	plasma membrane intrinsic protein 2A	PIP2A
AT1G15520	pleiotropic drug resistance 12	ABCG40
AT5G40770	prohibitin 3	PHB3
AT5G54190	protochlorophyllide oxidoreductase A	PORA
AT5G02810	pseudo-response regulator 7	PRR7
AT4G05590	pyruvate carrier-like protein	
AT3G25010	receptor like protein 41	RLP41
AT1G01360	regulatory component of ABA receptor 1	RCAR1
AT5G47910	respiratory burst oxidase homologue D	RBOHD
AT3G16857	response regulator 1	RR1
AT4G31920	response regulator 10	RR10
AT2G25180	response regulator 12	RR12
AT4G38430	rho guanyl-nucleotide exchange factor 1	ROPGEF1
AT1G55740	seed imbibition 1	SIP1
AT1G34210	somatic embryogenesis receptor-like kinase 2	SERK2
AT4G21540	sphingosine kinase 1	SPHK1
AT1G49040	stomatal cytokinesis defective / SCD1 protein	SCD1
AT5G61820	stress up-regulated Nod 19 protein	
AT1G53300	tetratricopetide-repeat thioredoxin-like 1	TTL1
AT5G54770	thiazole biosynthetic enzyme, chloroplast	ARA6
AT2G47270	transcription factor UPBEAT protein	UPB1
AT4G17770	trehalose phosphatase/synthase 5	TPS5
AT2G22500	uncoupling protein 5	UCP5
AT2G14740	vaculolar sorting receptor 3	VSR3
AT1G60490	vacuolar protein sorting 34	VPS34
AT2G04880	zinc-dependent activator protein-1	ZAP1

Table 11. Categorization of whether genes in each species were placed in the same ortholog group (OG) as the gene they matched to in the PaperBLAST database.

The full names of the plant species are listed in Table 6.

Species	Amtr	Arth	Brdi	Migu	Muac	Orsa	Pheq	Seit	Sobi	Soly	Vivi	Anco	Kafe
Same OG	541	1253	908	989	1271	993	668	1020	895	986	821	716	1045
Different OG	6256	9341	9965	10609	13496	12596	6531	12177	10977	10566	9277	7732	10714
No OG (query only)	1415	749	878	1009	2184	1510	1607	922	1289	1385	1881	1062	777
No OG (hit only)	85	145	118	115	183	135	98	131	131	127	104	92	158
No OG (both)	13	21	6	17	42	16	26	8	12	15	26	22	18

Table 12. Categorization of genes placed in the same ortholog group as the gene they matched to in the PaperBLAST database based on whether the matched gene was annotated as stomata-related or known as a key stomatal gene.

The full names of the plant species are listed in Table 6.

Species	Amtr	Arth	Brdi	Migu	Muac	Orsa	Pheq	Seit	Sobi	Soly	Vivi	Anco	Kafe
Annotated or Key orthologs	125	157	286	191	181	282	193	149	205	190	180	173	203
New Orthologs	395	530	914	687	768	945	768	493	784	674	755	616	792

Table 13. Orthologous genes *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Kalanchoë fedtschenkoi* that displayed differential enrichment between dusk and dawn in gene expression.

A. thaliana Symbol	*A. thaliana* Description	*A. thaliana* ID	*S. lycopersicum* ID	*K. fedtschenkoi* ID
ACA.l	autoinhibited Ca2+/ATPase II	AT1G13210	Solyc01g011100	Kaladp0043s0103
AT1G17500	ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein	AT1G17500	Solyc01g096930	Kaladp0050s0103
AT1G26130	ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein	AT1G26130		
AT1G72700	ATPase E1-E2 type family protein / haloacid dehalogenase-like hydrolase family protein	AT1G72700		
PDR4	pleiotropic drug resistance 4	AT2G26910	Solyc05g053570	Kaladp0058s0071
PDR6	pleiotropic drug resistance 6	AT2G36380	Solyc05g055330	Kaladp0068s0280
			Solyc06g065670	Kaladp0322s0001
KCS2	3-ketoacyl-CoA synthase 2	AT1G04220	Solyc05g013220	Kaladp0020s0110
KCS8	3-ketoacyl-CoA synthase 8	AT2G15090	Solyc12g006820	Kaladp0029s0057
				Kaladp0050s0301
				Kaladp0062s0076
				Kaladp0095s0482
AT1G72180	Leucine-rich receptor-like protein kinase family protein	AT1G72180	Solyc02g091860	Kaladp0062s0167
AT5G25930	kinase family with leucine-rich repeat domain-containing protein	AT5G25930	Solyc09g064520	Kaladp0090s0003
			Solyc12g098100	
RK1	receptor kinase 1	AT1G65790	Solyc04g077370	Kaladp0095s0362
			Solyc04g077390	Kaladp0266s0009
CIPK23	CBL-interacting protein kinase 23	AT1G30270	Solyc01g008850	Kaladp0053s0051
CIPK3	CBL-interacting protein kinase 3	AT2G26980	Solyc12g009570	
HXK2	hexokinase 2	AT2G19860	Solyc03g121070	Kaladp0037s0285
				Kaladp0064s0085
CPK4	calcium-dependent protein kinase 4	AT4G09570	Solyc10g081740	Kaladp0092s0084
			Solyc11g065660	
NAK	Protein kinase superfamily protein	AT5G02290	Solyc05g053930	Kaladp0058s0603
			Solyc06g005500	
PMR6	Pectin lyase-like superfamily protein	AT3G54920	Solyc03g111690	Kaladp0024s0371
			Solyc05g014000	
RCAR1	regulatory component of ABA receptor 1	AT1G01360	Solyc08g082180	Kaladp0042s0353
RCAR3	regulatory components of ABA receptor 3	AT5G53160		
CNGC5	cyclic nucleotide gated channel 5	AT5G57940	Solyc03g114110	Kaladp0008s0414
CLC-B	chloride channel B	AT3G27170	Solyc02g094060	Kaladp0011s0070
ATML1	Homeobox-leucine zipper family protein / lipid-binding START domain-containing protein	AT4G21750	Solyc10g005330	Kaladp0093s0030
NAP5	non-intrinsic ABC protein 5	AT1G71330	Solyc12g044820	Kaladp0040s0675
CAT2	catalase 2	AT4G35090	Solyc02g082760	Kaladp0001s0016

Table 14. Comparison of results between this study and two other studies investigating gene expression rescheduling.

A. thaliana Symbol	Annotated	Known	New	Present Study			Yang et al., 2017			Abraham et al., 2016	
				A. thaliana	*S. lycopersicum*	*K. fedtschenkoi*	*A. thaliana*	*A. comosus*	*K. fedtschenkoi*	*A. thaliana*	*A. americana*
Shared											
RCAR3	AT5G53160	AT5G53160		AT5G53160	Solyc08g082180	Kaladp0042s0353				AT5G53160	Aam022092
			AT5G25930	AT5G25930	Solyc02g091860	Kaladp0062s0167				AT5G25930	Aam087252
			AT5G25930	AT5G25930	Solyc12g098100	Kaladp0062s0167					
			AT5G25930	AT5G25930	Solyc09g064520	Kaladp0062s0167					
			AT5G25930	AT5G25930	Solyc12g098100	Kaladp0090s0003					
			AT5G25930	AT5G25930	Solyc09g064520	Kaladp0090s0003					
Not Shared											
PHOT2	AT5G58140	AT5G58140					AT5G58140	Aco014242	Kaladp0033s0113		
	AT1G80180									AT1G80180	Aam307072
MKK5	AT3G21220									AT3G21220	Aam003588
EULS3	AT2G39050									AT2G39050	Aam010610
OCP3	AT5G11270									AT5G11270	Aam077009
PWP2	AT1G15440									AT1G15440	Aam079149
iPGAM1	AT1G09780									AT1G09780	Aam051734
NLP7	AT4G24020									AT4G24020	Aam024504
TED4	AT2G26670									AT2G26670	Aam333043
HSC70-1		AT5G02500					AT5G02500	Aco031458	Kaladp0060s0296		
OST1	AT4G33950	AT4G33950								AT4G33950	Aam349853
HT1	AT1G62400	AT1G62400								AT1G62400	Aam018566
SNRK2.10		AT1G60940								AT1G60940	Aam332354
SOS2		AT5G35410								AT5G35410	Aam080324
ECA4		AT1G07670								AT1G07670	Aam088048
CLC-C		AT5G49890								AT5G49890	Aam081659
		AT4G28400								AT4G28400	Aam012848
ACA2		AT4G37640								AT4G37640	Aam003442
KT2/3		AT4G22200								AT4G22200	Aam018832
BRL1			AT1G55610				AT1G55610	Aco017367	Kaladp0008s0539		
PLC2			AT3G08510				AT3G08510	Aco001556	Kaladp0059s0034		
ICK5			AT1G49620				AT1G49620	Aco011817	Kaladp0028s0063		
			AT3G47090				AT3G47090	Aco008844	Kaladp1251s0003		
TPS7			AT1G06410				AT1G06410	Aco012107	Kaladp0011s0363		
GSL8			AT2G36850							AT2G36850	Aam013331
ARP2			AT3G27000							AT3G27000	Aam076219
ABCC2			AT2G34660							AT2G34660	Aam000228
cICDH			AT1G65930							AT1G65930	Aam080817
AGO1			AT1G48410							AT1G48410	Aam056884
			AT5G21105							AT5G21105	Aam023405
AHP1			AT3G21510							AT3G21510	Aam095535
DAA1			AT1G64110							AT1G64110	Aam087482
			AT5G61820							AT5G61820	Aam054195

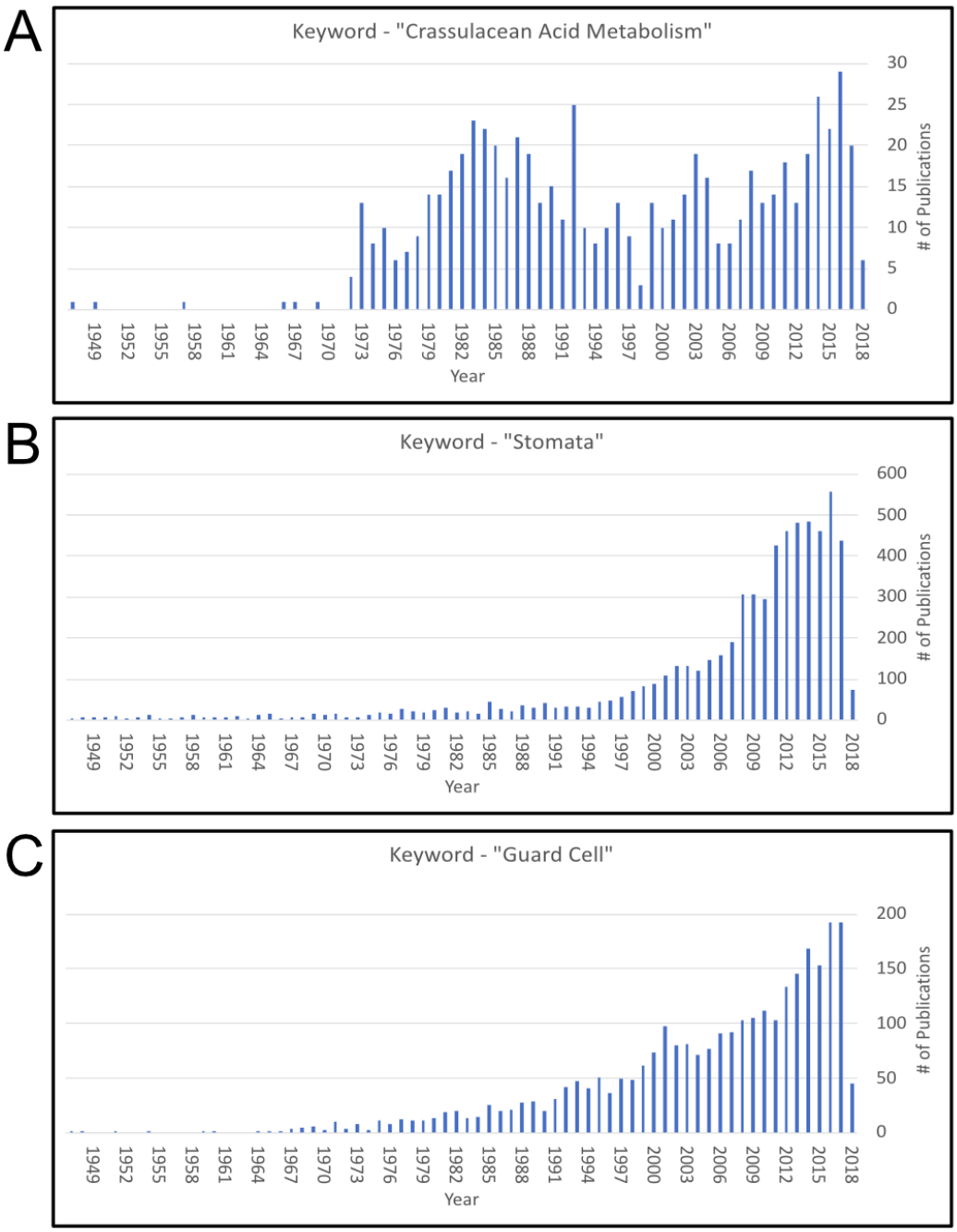


Figure 12. History of number of publications related to crassulacean acid metabolism, stomata, and guard cell research.

Results of searching the PubMed database using the key words "crassulacean acid metabolism" **(A)**, "stomata" **(B)**, and "guard cell" **(C)**.

A

Species	2	4	6	8	10	12	14	16	18	20	22	24
K. fedtschenkoi												
Phase					Dusk						Dawn	

Species	2	4	6	8	10	12	14	16	18	20	22	24
S. lycopersicum												
Phase				Dusk							Dawn	

Species	2	4	6	8	10	12	14	16	18	20	22	24
A. thaliana												
Phase					Dusk						Dawn	

B

	Dawn	Dusk
Gene i	d_{i1}	d_{i2}
Gene i	$\sum_{j \neq i} d_{j1}$	$\sum_{j \neq i} d_{j2}$

	Dawn	Dusk
Gene i	d_{i2}	d_{i1}
Gene i	$\sum_{j \neq i} d_{j2}$	$\sum_{j \neq i} d_{j1}$

Figure 13. Description of dawn and dusk phases for each species and the contingency tables used to gene expression enrichment.

A) The time points from each species' time-course data that were considered as occurring during dusk and dawn. **B)** Contingency tables (adapted from Yang et al., (2017)) for Fisher Exact Test to determine gene expression enrichment between dusk and dawn.

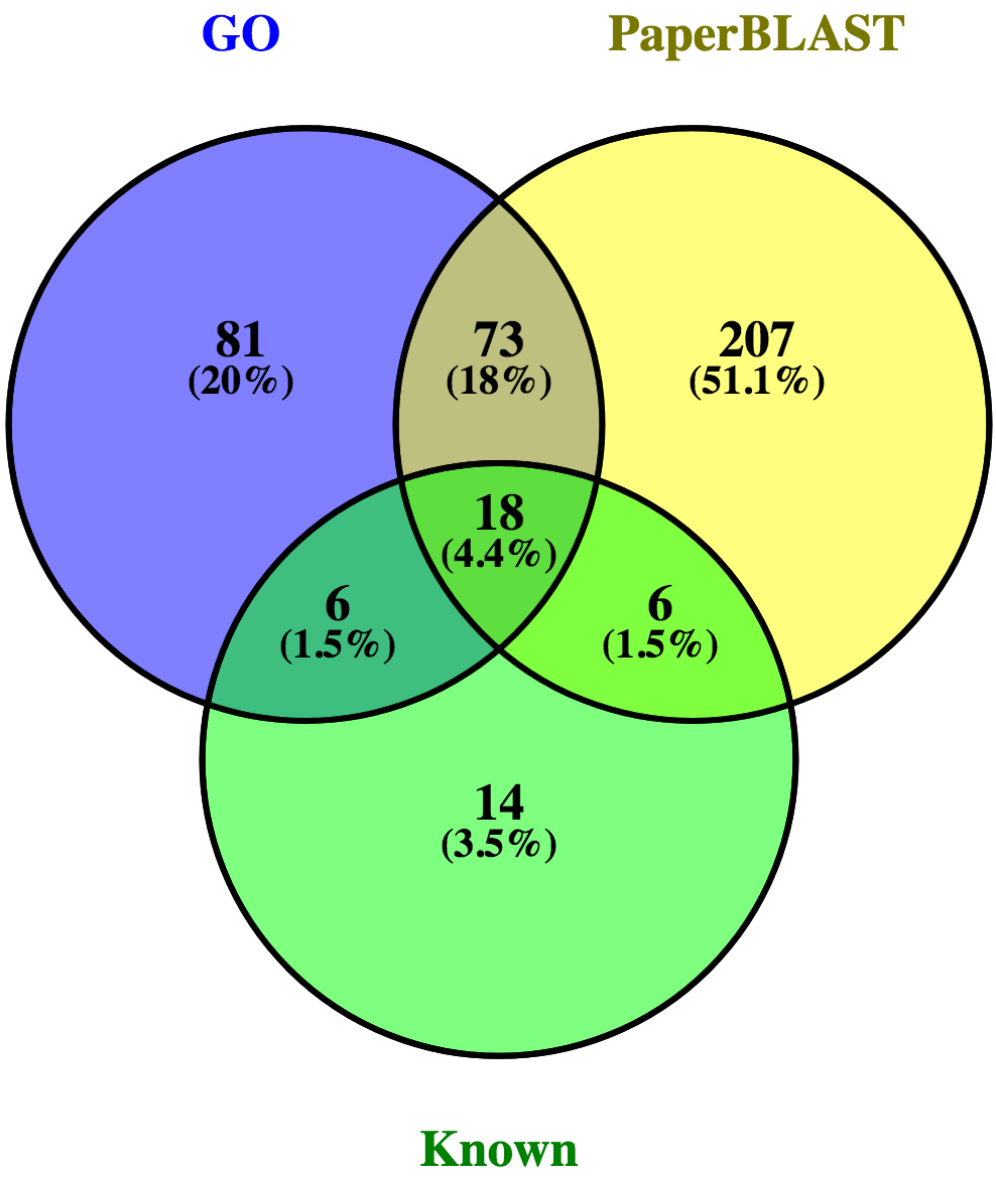


Figure 14. Overlap of *Arabidopsis thaliana* genes with annotations related to stomata, known as key stomatal genes, and identified in PaperBLAST database as stomata-related.

GO: genes annotated as stomata-related from Gene Ontology analysis. Known: genes known as stomata-related from the literature. PaperBLAST: stomata-related genes identified by using PaperBLAST search.

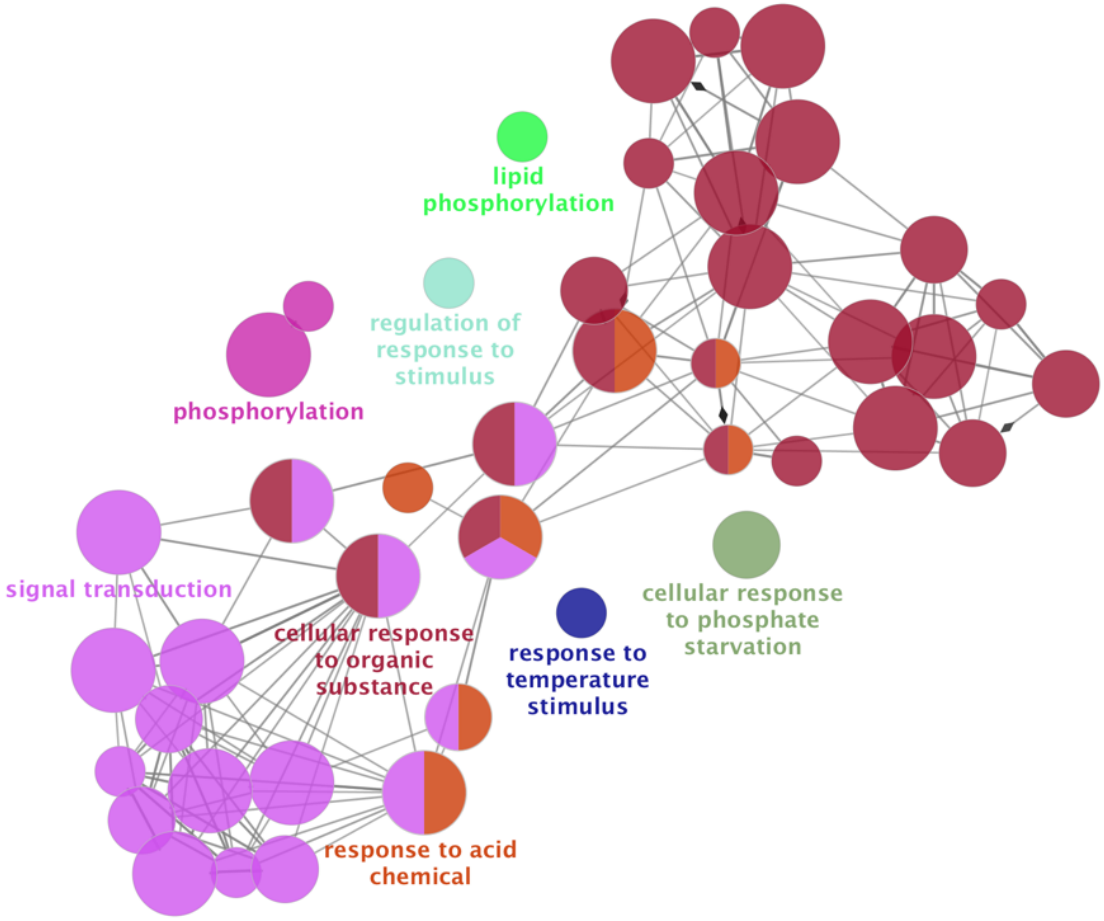


Figure 15. Gene ontology enrichment of new *Arabidopsis thaliana* stomatal genes.

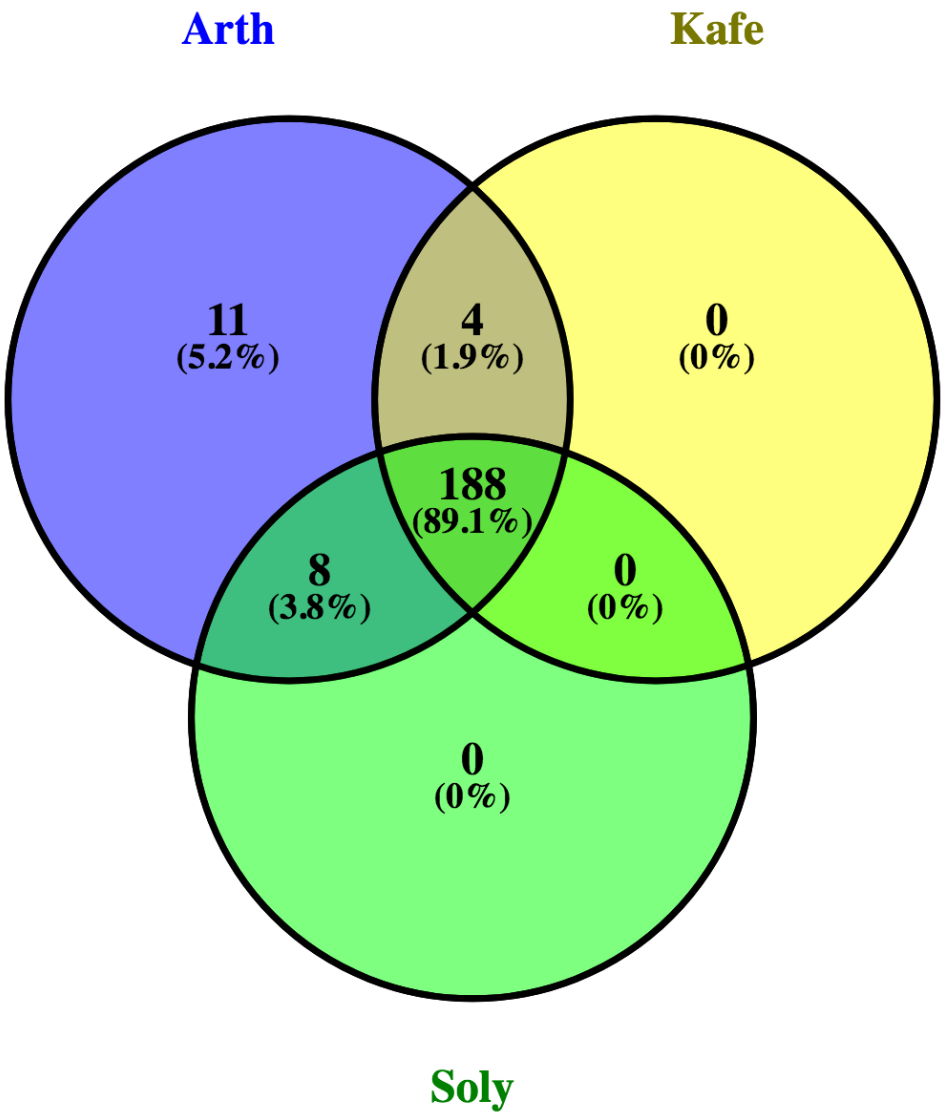
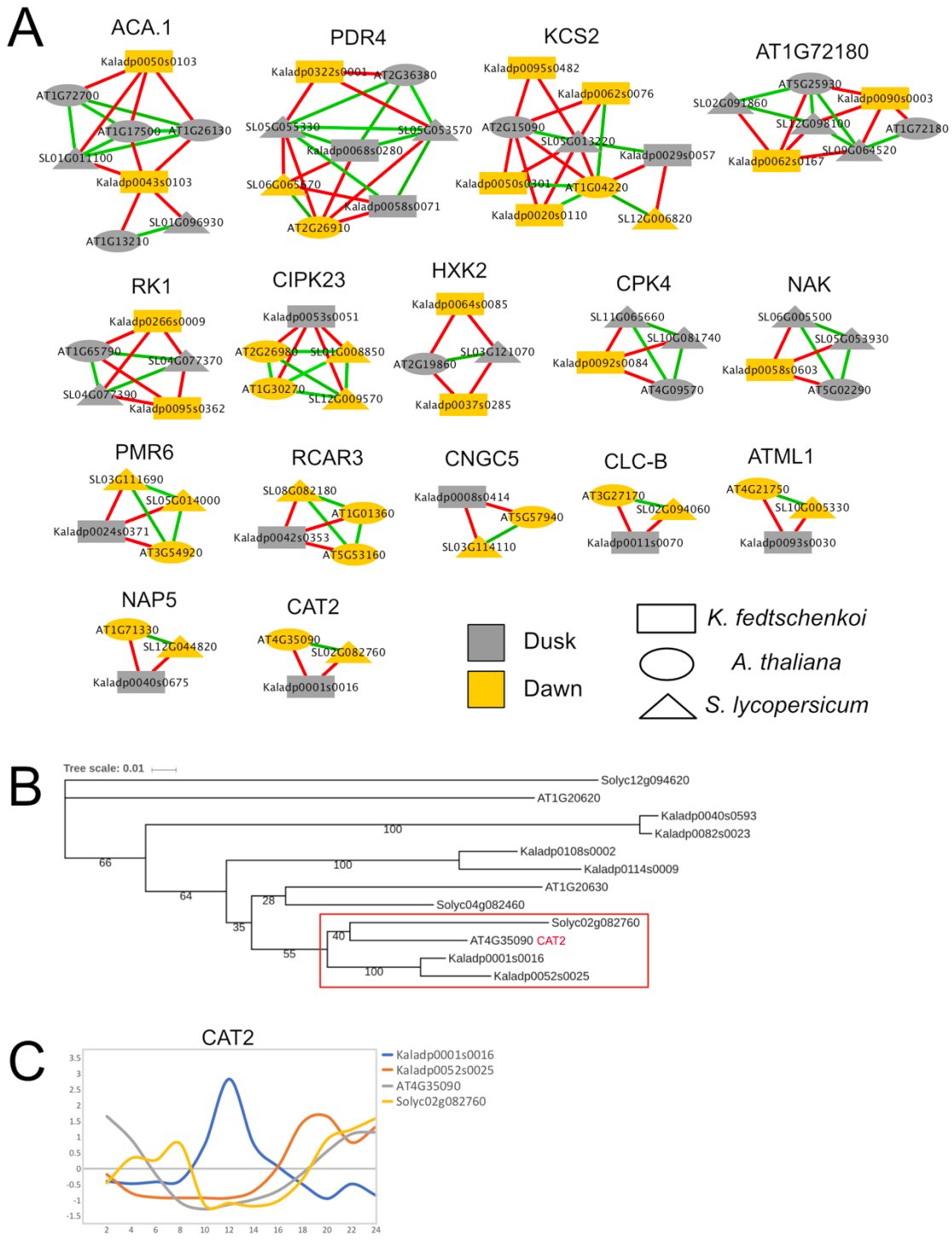


Figure 16. Overlap of ortholog groups between new stomata-related genes of *Arabidopsis thaliana* (Arth), *Solanum lycopersicum* (Soly), and *Kalanchoë fedtschenkoi* (Kafe).

Figure 17. Ortholog groups displaying enrichment time of each ortholog and phylogenetic tree and expression dynamics of catalase genes between *Arabidopsis thaliana*, *Solanum lycopersicum*, and *Kalanchoë fedtschenkoi*.

A) Genes in ortholog groups defined as rescheduled. Green edges represent rho ≥0.6 and red edges represent rho ≤-0.6. Rectangular, oval, and triangular nodes represent genes belonging to *K. fedtschenkoi*, *A. thaliana*, and *S. lycopersicum*, respectively. Nodes colored grey or yellow are genes that are enriched in dusk or dawn, respectively. **B)** Phylogenetic tree of catalase genes in *A. thaliana*, *S. lycopersicum*, and *K. fedtschenkoi*. **C)** Expression profiles of CAT2 genes in *A. thaliana*, *S. lycopersicum*, and *K. fedtschenkoi*. X-axis is z-score.

Figure 17. Continued.



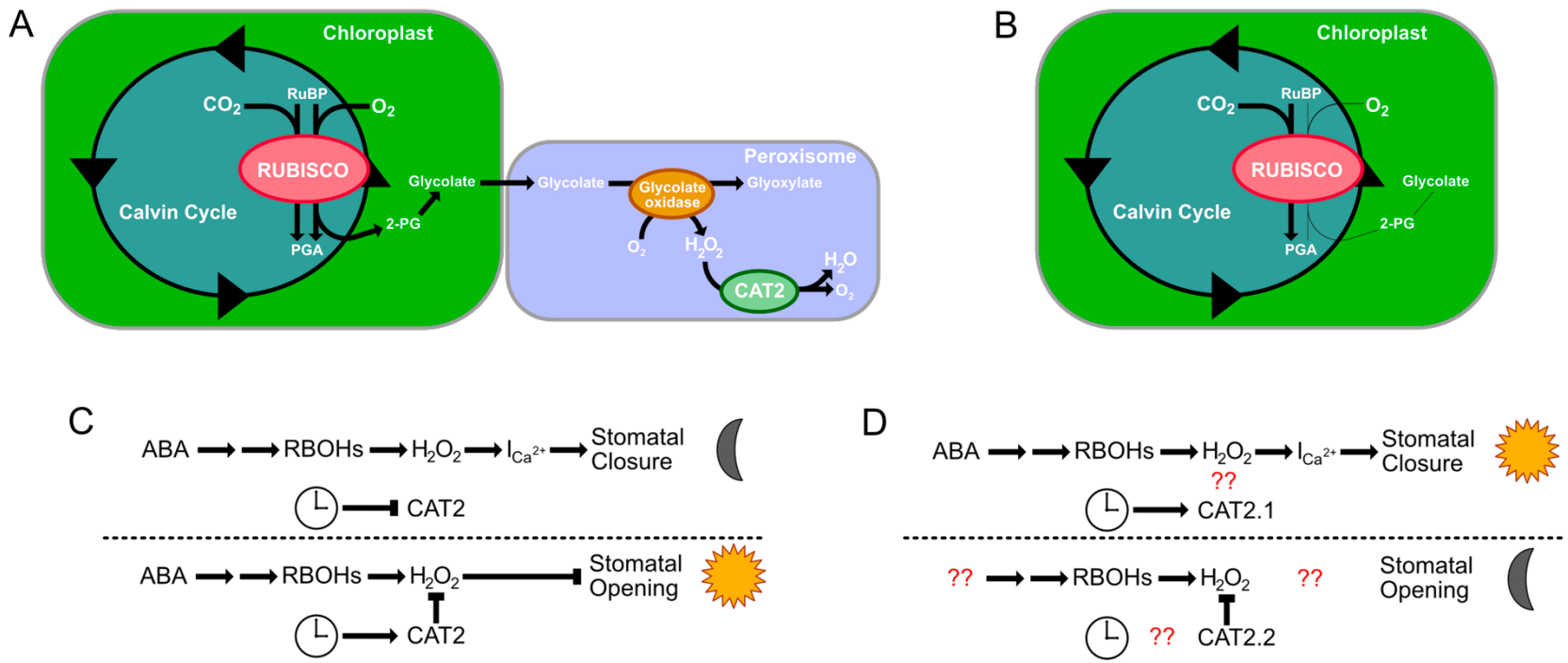


Figure 18. Catalase 2's (CAT2) roles in C_3 plants and predicted roles in CAM plants.

A) CAT2's role in photorespiration in C_3 plants. **B)** Photorespiration is reduced in CAM plants. **C)** CAT2's role in stomatal open closing (top) and opening (bottom) in C_3 plants. **D)** CAT2's predicted role in stomatal closing (top) and opening (bottom). Red question marks indict unknown interactions that warrant further investigation.

CHAPTER IV
INFERENCE OF GENE REGULATORY NETWORK UNCOVERS THE LINKAGE BETWEEN CIRCADIAN CLOCK AND CRASSULACEAN ACID METABOLISM IN *KALANCHOË FEDTSCHENKOI*

This chapter is derived from a manuscript currently in preparation: **Moseley R**, Motta F, Haase S, Yang X. Inference of gene regulatory network uncovers the linkage between circadian clock and CAM in *Kalanchoë fedtschenkoi.*

All work was conducted by Robert C. Moseley as well the writing of the manuscript. The remaining authors provided guidance and feedback on data analysis and manuscript preparation.

Abstract

The circadian clock enables plants to predict and anticipate periodic/diurnal changes in their environment, providing a great advantage in their sessile life. The molecular components at the core of clock have been found conserved across numerous plant species, in sequence and expression dynamics, and are hypothesized to control the majority of the genes expressed in plants. Recently, genes orthologous to the core circadian clock genes of *Arabidopsis thaliana* were found in *Kalanchoë fedtschenkoi*, an obligate CAM species. However, whether these clock orthologs regulate CAM processes, such as stomatal movement and CO_2 fixation, remains unknown. Here we show the circadian clock controls several genes involved stomata-related processes in *K. fedtschenkoi* and could have played a role in the inversion of day/night stomatal movement pattern in comparison with C_3 species. We first applied a new metric for identifying candidate clock-regulated genes using diel time-course RNA-Seq data, resulting in 593 clock-regulated genes, of which 69 were identified as transcription factors. Focusing on these clock-regulated transcription factors and reasoning that new clock genes regulate each other, we applied the Local Edge Machine (LEM) algorithm to infer regulatory relationships reciprocally between the transcription factors and known core clock genes in *K. fedtschenkoi*, identifying seven candidate clock genes that are closely connected with orthologous clock genes in *K. fedtschenkoi*. To test the hypothesis that the circadian clock controls the expression of stomata-related genes, we used LEM to identify stomata-related gene targets for known and new candidate clock genes and constructed a gene regulatory network for clock and stomata-related genes. Our results provide a reduction in the hypothesis space for investigating the architecture of the circadian clock in *K. fedtschenkoi*. Furthermore, our results provide new insight into the mechanism of circadian control of stomatal movement in *K. fedtschenkoi*, facilitating the engineering of drought tolerance into non-CAM plants.

Introduction

The circadian clock is a vital time-keeping mechanism that synchronizes periodic environmental signals to an organism's physiology, allowing for biological processes to function in an anticipatory manner. This mechanism is very important in plants due to their sessile nature. Numerous environmental signals and stressors to plants are cyclic, such as light availability, temperature, and predation. The circadian clock thus enables plants to anticipate these variables by activating the appropriate processes.

Plants that exhibit crassulacean acid metabolism (CAM) are examples of how plants synchronize biological processes to their environment, enabling an advantage.

CAM plants exhibit improved photosynthetic efficiency due to a temporal separation of CO_2 fixation and improved water-use efficiency due to inverted stomatal conductance, relative to C_3 plants (Ehleringer and Monson 1993). These traits coupled with the global issue of increased frequency and intensity of drought (Dai 2012; Cook et al. 2014) have generated an increase in CAM research with the goal of engineering these traits into C_3 plants, to enable better drought responses and/or improved drought tolerance (Borland et al. 2014; Yang et al. 2015). Currently, it is theorized that the temporal separation of CO_2 fixation is under control of the circadian clock (Hartwell 2005; Hartwell 2006) and that the inverted stomatal conductance could be a result of a change in clock regulation (Moseley et al., manuscript in preparation). In general, though, the extent of the role the circadian clock plays in the improved photosynthetic efficiency and water-use efficiency characteristics of CAM plants is not well understood.

Omics data (e.g., genomic and transcriptomic data) provide a source for investigating the circadian clock in organisms. Several studies have utilized these data to identify clock-regulated genes (Covington et al. 2008; Michael et al. 2008; Sharma et al. 2017) and compare clock output across species (Filichkin et al. 2011; Boyle et al. 2017;Moseley et al., manuscript in preparation). To date, several time-course transcriptome datasets exist for various plant species, however, only three exist for CAM species, *Ananas comosus* (Ming et al. 2015), *Agave Americana* (Abraham et al. 2016), and *Kalanchoë fedtschenkoi* (Yang et al. 2017).

Analyses of the *A. comosus* 24-h leaf RNAseq data were able to reveal links between the circadian clock and CAM-related processes, as well conservation of core clock genes. Differential gene co-expression networks were constructed using the gene expression data from the photosynthetic green leaf tissue and the non-photosynthetic white leaf tissue, revealing two co-expression clusters containing most CAM pathway genes. Genes in these clusters displayed strong diurnal expression patterns and the clusters themselves were enriched for core circadian clock genes and circadian-related *cis*-acting elements, suggesting circadian control of gene expression of CAM-related genes (Wai et al. 2017). An additional study used the data as input into a rhythmic detection algorithm, HAYSTACK (Michael et al. 2008), to identify rhythmically expressed transcription factors and transcription coregulators and found that almost half of each group of factors were rhythmic (Sharma et al. 2017). Orthologs of core clock genes in *Arabidopsis thaliana* were found conserved in sequence and in expression dynamics.

Analyses of the *K. fedtschenkoi* dataset also provided evidence for circadian clock involvement of CAM expression and evolution. In combination with 24-h time-course datasets from *A. comosus* and *A. thaliana*, convergence in the rescheduling of gene expression between CAM-related orthologs of the two CAM species and *A. thaliana* orthologs was identified (Yang et al. 2017). No genes orthologous to *A. thaliana* circadian clock genes showed evidence of rescheduling; however, the authors do state that there could be unknown circadian genes between the two species that display rescheduling in gene expression or there are genes involved in circadian rhythm in *K. fedtschenkoi* that lack this function in *A. thaliana*. Further analysis between the *K. fedtschenkoi* and *A. thaliana* datasets revealed CAM-related orthologs displaying various

clock-related phenotypes in expression, e.g., phase shifts between rhythmic orthologs, suggesting the circadian clock could have played a role in the evolution and maintenance of CAM expression (Moseley et al., manuscript in preparation). Whether there are new clock genes in *K. fedtschenkoi* or if clock genes regulate different genes between *K. fedtschenkoi* and *A. thaliana* remains unanswered.

Numerous computational methods exist for investigating and inferring gene networks based on gene expression data that can help answer these questions. A new method has recently been published, which specializes in using time-course data to infer regulatory relationships and network dynamics, called the local edge machine (LEM) (McGoff et al. 2016). LEM is based on nonlinear ordinary differential equations (ODEs) for the chemical kinetics and a Bayesian formalism on the network structure (McGoff et al. 2016). The nonlinear kinetic equations employed are Hill functions that model the activation and repression interactions between a gene and its target gene. This allows for a more intuitive understanding of the resulting models as they provide realistic models of transcription. LEM's usefulness was demonstrated by applying it to identify new clock genes in mouse. (McGoff et al. 2016).

The circadian clock plays a crucial role in the physiological response to various environmental stresses in plants, such as drought (Grundy et al. 2015), and CAM is an adaptation to arid environments. Therefore, identifying new clock genes and establishing links between the clock and CAM will be key to fully elucidating the circadian regulation of CAM and successful engineering of CAM into C_3 plants for improved drought response and tolerance. The aim of this study is to address the question of whether new genes could play a role in the circadian clock in *K. fedtschenkoi* and if clock genes control stomata-related genes in *K. fedtschenkoi*. Leveraging LEM's capabilities, several genes were identified as potentially new clock genes in *K. fedtschenkoi*. Additionally, stomata-related genes, including genes with rescheduled gene expression, were predicted to be regulated by clock genes in *K. fedtschenkoi*.

Methods

Kalanchoë fedtschenkoi and Arabidopsis thaliana time-course data

The diurnal expression data for *K. fedtschenkoi* and *A. thaliana* were obtained from (Yang et al. 2017) and (Mockler et al. 2007), respectively. The *Kalanchoe* expression data were collected at 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, and 24 h whereas the *Arabidopsis* data were collected at 4, 8, 12, 16, 20, and 24 h after the starting of the light period. Since the *Arabidopsis* gene expression data was measured at 4-h intervals and the *Kalanchoe* data was measured at 2-h intervals, the *Arabidopsis* data was adjusted to arrive at expression profiles for all *Arabidopsis* and *Kalanchoe* genes on the same time scale. Here, the piecewise cubic Hermite interpolating polynomial (pchip) interpolation function in the pandas Python library was used to up sample the *Arabidopsis* data to simulate gene expression levels at additional time points so that both time-course data sets consisted of the same time intervals: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, and 24 h after the starting of the light period. Pchip was preferred over the more common method of cubic spline interpolation due to cubic spline's tendency to overshoot which introduces

oscillations. Additionally, pchip maintains the shape of the data as best as possible and has been used on microarray time course data sets (Dong et al. 2011; Luo et al. 2014). Additionally, *K. fedtschenkoi* genes with a max FPKM <1 were considered noise and removed. The rhythmic *K. fedtschenkoi* gene set was taken from Moseley et al. (manuscript in preparation) while the rhythmic *Arabidopsis* gene set was taken (Mockler et al. 2007) as well as from Moseley et al. (manuscript in preparation).

Identifying candidate clock genes

Clock genes in various species tend to have the highest amplitude and the most statistically significant rhythms (Anafi et al. 2014; McGoff et al. 2016; Sharma et al. 2017). The periodicity detection methods of de Lichtenberg (DL) (de Lichtenberg et al. 2005) and JTK-CYCLE (JTK) (Hughes et al. 2010) are assays that take into account the amplitude of time-course gene expression and if the period of expression matches to the period length in question, respectively. Therefore, to identify candidate clock genes, a new metric was made that combines these two features of DL and JTK (S. Haase and F. Motta, personal communication). This new metric is named DLxJTK and uses the p-values for amplitude from DL and for periodicity from JTK for each gene. The DLxJTK formula is

$$\text{DLxJTK} = P_{per}\, P_{amp} \left(\left(1 + \frac{P_{per}}{0.001}\right)^2\right)\left(\left(1 + \frac{P_{amp}}{0.001}\right)^2\right)$$

Where, P_{per} is the JTK p-value for periodicity and P_{amp} is the DL p-value for amplitude. DLxJTK will be applied to the *K. fedtschenkoi* ranked ordered rhythmic gene list from Moseley et al. (manuscript in preparation) to identify candidate clock-regulated genes. A p-value cutoff of 0.05 will be applied to create a set of high-confidence candidate clock-regulated genes.

Identifying candidate core clock genes

The first implementation of LEM will be to identify if any candidate clock-regulated genes could function in the clock network of *K. fedtschenkoi*, reasoning that a new core clock gene would have similar expression dynamics as other core clock genes. The majority of core clock genes identified in plants are TFs (Nohales and Kay 2016), therefore, to determine if any candidate clock-regulated genes could be involved in the core clock, only candidate clock-regulated genes identified as TFs were used. Of these clock TFs, lux arrhythmo (LUX) is only active after forming the evening complex (EC) with early flowering 3 (ELF3) and 4 (ELF4) (Nusinow et al. 2011; Mizuno et al. 2014), which have not been found to bind to DNA (Helfer et al. 2011). Due to this relationship, ELF3 and ELF4 were also included. The clock TFs will be used as potential regulators of the candidate clock-regulated TFs in the first run of LEM. Literature supports that RVE8, RVE6, LWD1, LWD2, LNK1, and LNK2 are activators of transcription and CCA1, LHY, TOC1, LUX, NOX, PRR9, PRR7, PRR5, ELF3, and ELF4 are repressors of transcription (Nohales and Kay 2016), therefore LEM will be set to only use the respective mode of gene regulation for each clock TF used. All candidate clock-regulated TFs will be set as targets for the clock TFs. Since core clock TFs are known to regulate

other known core clock genes, LEM will be run again but with the candidate clock-regulated TFs as potential regulators of the core clock TFs. To determine potentially new core clock TFs in *K. fedtschenkoi*, a measure of likelihood, described in (McGoff et al. 2016), will be used for each candidate clock-regulated TF. This measure is calculated by taking the maximum LEM probability that the candidate clock-regulated TF was a regulator of any known core clock TF and multiplying it by the LEM probability that the candidate clock-regulated TF was regulated by any core clock TF.

Identify clock-regulators of stomata-related genes

Various stomata-related genes have been found to display circadian expression in various CAM plants (Ming et al. 2015; Abraham et al. 2016), as well as show re-scheduling of transcript expression relative to their respective ortholog in *A. thaliana* (Abraham et al. 2016; Yang et al. 2017). Such evidence, coupled with the fact that most differences in gene expression are due to *trans*-acting regulatory variation (Brem et al. 2002), leads to the reasoning that the reprogramming of the stomata-related genes between *A. thaliana* and *K. fedtschenkoi* could be the result of a change in gene regulator, particularly a clock regulator. Therefore, LEM will be applied to identify regulatory relationships between core clock TFs and stomata-related genes. Known core-clock TFs plus the candidate core-clock TFs predicted earlier will be used as potential regulators of stomata-related genes. Stomata-related genes in *K. fedtschenkoi* are identified as genes that are orthologous to an *A. thaliana* that is either annotated as stomata-related or are known stomata-related genes. Orthology between species is based on placement within the same ortholog group. Additionally, new *K. fedtschenkoi* stomata-related genes identified in Chapter 3 will be included as well. Only stomata-related genes with an FPKM >0.01 were used.

Gene Ontology analysis

Gene Ontology (GO) terms for the *K. fedtschenkoi* and *A. thaliana* were obtained from Phytozome v12.1 (Goodstein et al. 2012). *K. fedtschenkoi* genes encoding putative transcription factors were retrieved from (Yang et al. 2017). Using ClueGO (Bindea et al. 2009), observed GO biological process were subjected to the right-sided hypergeometric enrichment test at medium network specificity selection and p-value correction was performed using the Holm-Bonferroni step-down method (Holm 1979). There was a minimum of 3 and a maximum of 8 selected GO tree levels, while each cluster was set to include a minimum of between 3% and 4% of genes associated with each term. GO term fusion and grouping settings were selected to minimize GO term redundancy and the term enriched at the highest level of significance was used as the representative term for each functional cluster. The GO terms with p-values less than or equal to 0.05 were considered significantly enriched.

Comparative analysis of gene expression

To calculate time-delay between time-course gene expression profiles, the diurnal expression data were normalized by Z-score transformation. Pair-wise correlation was calculated for the orthologous gene pairs of interest for all possible time delays using

circular cross correlation. Circular cross correlation produces a correlation coefficient between gene 1 and shifted, or lagged, copies of gene 2 as a function of the lag. With each correlation coefficient, a lag value was given. The lag values were then converted into hours, giving an estimate on time delay. The time delay at which the correlation was maximum was selected as the estimated delay between the two genes. Spearman's rank correlation coefficient was then calculated between gene 1's expression data and the shifted expression data of gene 2 by its estimated time delay.

Results

Candidate clock-regulated genes in K. fedtschenkoi

A set of 593 candidate clock-regulated genes were identified in *K. fedtschenkoi* with the applied DLxJTK cutoff (Figure 19A and Table S10). The candidate clock-regulated genes covered all phases of the day and displayed a bimodal distribution with peaks occurring before subjective night and before subjective morning (Figure 19B). This is consistent with phase call distributions seen in other plant species, as well as non-plant species (Michael et al. 2008; Filichkin et al. 2011; Zhang et al. 2014). Several biological processes were found to be enriched in the gene set, such as lipid biosynthetic process, photosynthesis (light harvesting), and starch metabolic process (Figure 19C). Of the 593 genes, 69 (11.6%) were predicted to be TFs. Several clock genes in *A. thaliana* are known to be TFs. To determine if any of the *K. fedtschenkoi* TFs were orthologous to *A. thaliana* TFs that have been annotated as circadian-related, ortholog groups (OG) constructed in (Yang et al. 2017) were used to obtain candidate clock-regulated TFs in the OGs. Only 38 of the 69 *K. fedtschenkoi* TFs were placed in an OG with an *A. thaliana* gene, which totaled 77 *A. thaliana* genes. Using DAVID v6.8 (Huang da et al. 2009a; Huang da et al. 2009b), the 77 *A. thaliana* gene IDs were used to extract gene annotations, if available. Of the 77 *A. thaliana* genes, 13 were found to be associated with circadian rhythm (Table 15 and Table S11). Several TF families were represented with the C2H2, MYB-HB, and C2C2-CO families containing 15, 9, and 9 genes, respectively (Figure 19D). CO-like and MYB-related TF families have been linked to the circadian clock in previous studies (Ledger et al. 2001; Cheng and Wang 2005; Hazen et al. 2005; Imaizumi and Kay 2006; Zhang et al. 2007). Other TF families represented, such bZIP and JmjC, have also been reported to be linked to the circadian clock in *A. thaliana* (Ledger et al. 2001; Covington et al. 2008) as well as in *Drosophila* (Shalaby et al. 2018). Candidate clock-regulated TFs did not encompass all phases of the day and were mainly concentrated before subjective night and before subjective morning (Figure 19D E-F).

Candidate clock transcription factors

After applying a cutoff of 0.7 to remove low probability regulatory relationships, all core clock TFs were predicted to regulate at least one candidate clock-regulated TF (Figure 20). One ortholog of LNK2 (Kaladp0099s0129) was predicted to regulate the most candidate clock-regulated TFs at 20 genes while the orthologs of LUX (Kaladp0033s0047) and LNK1 (Kaladp0607s0046) were predicted to regulate eight different candidate clock-regulated TFs each. All but 6 candidate clock-regulated TFs

were found to be activated or repressed by clock TFs, while 13 candidate clock-regulated TFs were found to activate or repress core clock TFs. Only one candidate clock-regulated TF (Kaladp0748s0043) was predicted to regulate more than one core clock TF (Figure 20). A likelihood ranking was created by ranking the product of each candidate clock-regulated TF's max probability of being a regulator of and being regulated by a clock TF. Using a likelihood ranking cutoff of 0.7, eight TFs were identified to serve as candidates for new clock TFs in *K. fedtschenkoi* (Figure 21 and Table S12). The eight candidate core clock TFs were phased to 3 separate phases of the day. To annotate each of the eight candidate core clock TFs, *A. thaliana* orthologs were identified by placement in OGs. All *A. thaliana* orthologs identified were rhythmic (Mockler et al. 2007) and (Moseley et al., manuscript in preparation). Descriptions of the genes are below and separated into three categories corresponding to a phase of the day of max gene expression.

Morning phased candidate clock transcription factors

Three of the candidate clock TFs (Kaladp0011s1342, Kaladp0009s0042, and Kaladp1154s0002) in *K. fedtschenkoi* were phased to the morning (Table 16). Among these three *K. fedtschenkoi* TF genes, Kaladp0011s1342 and Kaladp0009s0042 were not placed in an OG with an *A. thaliana* gene. Therefore, their respective protein sequence was used to search the NCBI non-redundant protein BLAST database using a E value cut off of 1e-5. Kaladp0011s1342 was phased to 22 h after light and found to have a similar protein sequence with two *A. thaliana* proteins, AT3G58120 (BZIP61) and AT2G42380 (BZIP34) (Table 16 and Figure 22). BZIP TFs are known to regulate pathogen defense, light and stress signaling, seed maturation and flower development (Jakoby et al. 2001). BZIP34 has been predicted to be involved in the regulation of lipid metabolism and/or cellular transport (Gibalova et al. 2009). BZIP34 and BZIP61 were both rhythmic and were phased to four and eight h after light, respectively (Table 16 and Figure 22). Kaladp0009s0042's protein sequence lacked homology with any protein sequences in *A. thaliana*. The protein sequence of Kaladp0009s0042 was found to contain a Dof domain which is a zinc finger DNA-binding domain were found. Additionally, the remaining *K. fedtschenkoi* gene in this group, Kaladp1154s0002, was found in an OG containing three *A. thaliana* genes encoding for the Dof domain-containing proteins, including cycling DOF factor 1 (AT5G62430; CDF1), 2 (AT5G39660; CDF2), and 3 (AT3G47500; CDF3). CDF1, CDF2, and CDF3 are involved in various signaling pathways, including photoperiodic and light signaling, stress responses and circadian clock regulation (Fornara et al. 2009). CDF1 transcription has been reported to be repressed by the circadian clock pseudo-response regulator protein family (Nakamichi et al. 2007; Niwa et al. 2007; Ito et al. 2008; Huang et al. 2012; Nakamichi et al. 2012) and activated by the circadian clock genes circadian clock associated 1 (CCA1) and late elongated hypocotyl (LHY) (Seaton et al. 2015), resulting in CDF1 gene expression at dawn. All three *A. thaliana* orthologs were rhythmic and phased to dawn (Table 16 and Figure 22). CDF1 protein accumulation is also regulated by the circadian clock through protein stability via complex formation with gigantean (GI) or flavin-binding, kelch repeat, F-box 1 (FKF1) (Imaizumi et al. 2005; Fornara et al. 2009; Song et al. 2012). However, feedback into the

clock has not been reported for the CDFs. Both Kaladp0009s0042 and Kaladp1154s0002 gene expression peaked at dawn (Table 16 and Figure 22).

Midday phased candidate clock transcription factors

Four of the candidate clock TFs (Kaladp0878s0025, Kaladp0674s0030, Kaladp0748s0043, and Kaladp0674s0182) in *K. fedtschenkoi* were phased to midday (Table 16 and Figure 22). Kaladp0878s0025 had one *A. thaliana* ortholog (AT1G07050) in its respective OG, AT1G07050. AT1G07050 encodes for a constans-like protein encoding gene. Little data is available on AT1G07050; however, (Keurentjes et al. 2007) conducted a genome-wide expression variation analysis in a recombinant inbred line population of *A. thaliana* and constructed a genetic regulatory network by combining eQTL mapping and regulator candidate gene selection into a subnetwork for flowering time. Through their analysis, they predicted AT1G07050 to be a target of the clock regulator GI and was found within 2.5 Mbp of the GI locus. Kaladp0878s0025 gene expression was phased to 8 h after light, whereas the *A. thaliana* ortholog had gene expression phased to 12 h after light (Table 16 and Figure 22). Kaladp0674s0030's OG contained two *A. thaliana* orthologs, AT5G63160 and AT3G48360, both encode for members of the Bric-a-Brac/Tramtrack/Broad (BTB) gene family, specifically BT1 and BT2, respectively. Only BT2 had gene expression data in ??? and was found to be rhythmic with gene expression phased to 20 h after light. Kaladp0674s0030 was phased to six h after light, displaying a shift in expression between the two species (Table 16 and Figure 22). BT2 is known to activate telomerase expression in mature *A. thaliana* leaves, play a critical role in nitrogen-use efficiency in *A. thaliana* and *Oryza sativa,* suppress sugar and ABA responses, and positively regulate certain auxin responses in plants (Mandadi et al. 2009; Araus et al. 2016). Additionally, BT2 is regulated diurnally and controlled by the circadian clock, with maximum expression in the dark. It has been suggested that the pattern of gene expression for BT2 mRNA could be linked to the availability of photosynthate, which is a product of photosynthesis (Araus et al. 2016). Kaladp0748s0043 was found orthologous to the *A. thaliana* plant homeobox family protein BELL1 (BEL1), which is a key regulator of ovule development and needed for auxin and cytokinin signaling pathways for correct patterning of the ovule (Bencivenga et al. 2012). Kaladp0748s0043's gene expression was phased to 6 h after light, whereas BEL1 in *A. thaliana* was phased to 12 h after light (Table 16 and Figure 22). Lastly, Kaladp0674s0182 had only one *A. thaliana* ortholog, AT3G29270, which encodes for a ring/U-box superfamily protein. Proteins in this superfamily are involved in protein ubiquitination. Kaladp0674s0182 was annotated as a C2H2 TF in (Yang et al. 2017), but searching the Pfam database with its amino acid sequence revealed a zinc finger, ring-type domain (Figure S3). Since the focus of this study is on TFs, Kaladp0674s0182 will not be used in further analysis, however, its role in circadian post-translational regulation is worth investigation.

Evening phased candidate clock transcription factors

Only one candidate core clock TF (Kaladp0007s0017) in *K. fedtschenkoi* had gene expression phased to the evening (Table 16 and Figure 22). Kaladp0007s0017 had 2

A. thaliana orthologs in its respective OG, which encode for jasmonate (JA)-associated MYC2-like proteins 1 (AT2G46510; JAM1) and 2 (AT1G01260; JAM2). JAM1 has been reported as the balancing component opposite of the MYC2 TF in the JA signaling pathway (Fonseca et al. 2014). Specifically, JAM1 and MYC2 are induced by JA and share many of the same target genes. Where MYC2 activates transcription of multiple genes, including JAM1 which negatively influences gene expression by physically inferring with MYC2 binding to promoter regions of target genes (Fonseca et al. 2014). These target genes for both species are considered "early-responsive JA genes" as changes in gene expression of target genes occur within 1 h of JA detection (Nakata et al. 2013). JA signaling is linked to activation of defense pathways and has been reported to under the control of the circadian clock through regulation of MYC2 via repression of transcription by time for coffee (TIC) (Li et al. 2007). Additionally, JAM1 has been reported to participate in ABA signaling as a positive regulator as overexpression of the gene in *A. thaliana* increased drought tolerance (Li et al. 2007). JAM1 and JAM2 in *A. thaliana* had gene expression phased to 12 and 9 h after light, respectively, whereas Kaladp0007s0017 had gene expression phased to 14 h after light (Table 16 and Figure 22).

Clock regulation of stomata-related genes in K. fedtschenkoi

Stomata-related processes have been shown to be under circadian clock control via regulation of gene expression, allowing plants to anticipate the oncoming day and night. Adding the fact that stomatal movement has been inverted in CAM plants raises the question if the circadian clock, specifically clock TFs, played a role in this inversion through rescheduling of gene expression. To investigate this question, regulatory relationships were inferred between core clock TFs and the candidate clock TFs identified in this study and rhythmic genes involved in stomata-related processes in *K. fedtschenkoi*. Rhythmic genes were those identified as rhythmic in Moseley et al. (manuscript in preparation) and resulted in 1,605 rhythmic stomata-related genes. *K. fedtschenkoi* genes orthologous to two clock TFs (Kalado0033s0047: LUX/NOX and Kaladp0032s0115: PRR9/5) were found within the target list of rhythmic stomata-related genes and were subsequently removed as targets. Additionally, candidate clock TFs 2, 4, 5, 6, and 8 were also in the target list and removed as targets.

LEM was used to infer regulatory relationships between clock TFs and rhythmic stomata-related genes. Using a cutoff of 0.7 on the LEM output related to the probability of a TF regulating a gene, 582 of the 1,605 stomata-related genes were inferred to be regulated by core clock TFs (Table 17 and Table S13). A visualization of the overall network can be seen in Figure 23. Clock genes are known to activate or repress genes, depending on their mode of regulation, during specific phases of the day. For instance, CCA1 and LHY repress genes that are expressed during the evening (Nohales and Kay 2016). To determine if LEM predicts the appropriate phase of regulation for clock genes, the phase calls of target genes for core clock TFs with ≥10 target genes were plotted (Figure 24). Night light-inducible and clock-regulated 1 (LNK1) and 2 (LNK2) are thought to activate gene expression of targets during the afternoon and evening (Nohales and Kay 2016), and in line with this, a majority of the predicted stomata-related gene

targets for both LNK genes are phased to the afternoon and evening (Figure 24). Pseudo response regulator 7 (PRR7) is known to repress genes during dawn and in the morning (Nohales and Kay 2016). Indeed, gene targets predicted by LEM for PRR7 in *K. fedtschenkoi* were phased to dawn and the morning (Figure 24). Lastly, early flowering 4 (ELF4) is known to repress genes in the morning and evening by forming a complex with other ELF3 and LUX (Nusinow et al. 2011). LEM only predicted morning-phased gene targets for ELF4 in *K. fedtschenkoi* (Figure 24). The remaining components of the EC were examined as well to see if any of their targets were phased to in the evening. None of the targets of the remaining EC components, including the two other ELF4 genes had genes phased to the evening (Table S14). Additionally, ELF4, ELF3, and LUX did not share similar targets.

Since candidate core clock TFs were set in LEM to be used as activators and repressors, all candidate core clock TFs were predicted as both activators and repressors of gene expression (Table 17). However, candidate clock TFs 2 (CTF2) and 6 (CTF6) were primarily predicted as activators of gene expression and CTF 4 was predicted primarily as a repressor of gene expression (Table 17). A majority of the targets for all candidate TFs were phased to either dusk or dawn (Figure 24).

To determine what biological functions in stomata-related processes are under the control of the circadian clock, enrichment of associated gene ontology terms was performed. A majority of the biological functions enriched in the 582 rhythmic stomata-related genes were associated to protein phosphorylation (Figure 25). To determine if any *K. fedtschenkoi* genes are related to *A. thaliana* genes annotated or known as stomata-related, OGs were examined again. Within the 582 *K. fedtschenkoi* genes, 49 were placed in OGs that contained *A. thaliana* genes that were either annotated or known as stomata-related genes (Table 17, Table S13). All the candidate core clock TFs and four core clock TFs were predicted to regulate at least one of the 49 *K. fedtschenkoi* genes. The remaining stomata-related genes were identified in Moseley et al. (manuscript in preparation) as new stomata-related genes and all core clock and new candidate core clock TFs were predicted to regulate at least one new stomata-related gene (Table 17, Table S13).

Regulation of rescheduled stomata-related genes

Two studies have investigated the molecular components involved in the inversion of stomatal movement in the CAM species *K. fedtschenkoi*, (Yang et al. 2017) and Moseley et al. (manuscript in preparation). Five genes identified in (Yang et al. 2017) were inferred to be regulated by the clock TFs and with only one being orthologous to a known stomata-related gene in *A. thaliana* (Table 18). Twelve genes identified in Moseley et al. (manuscript in preparation) were inferred to be clock regulated with only three being orthologous to annotated or known stomata-related genes in *A. thaliana* (Table 18). Several of these genes were predicted to encode for protein kinases and transporters. CTF3 was predicted to regulate the most stomata-related genes that displayed re-scheduling (Table 18). Interestingly, CTF3 also displayed rescheduling of gene expression relative to its two orthologs in *A. thaliana* (Figure 22) and one of the

targets of CTF3 was the rescheduled catalase 2 (CAT2) gene identified in Moseley et al. (manuscript in preparation).

Discussion

The aim of this study was to first infer new clock genes in *K. fedtschenkoi* and then determine which stomata-related genes are controlled by the circadian clock. Through a combination of a new metric to identify clock-regulated genes and LEM, several TFs were classified as candidate core clock TFs in *K. fedtschenkoi* with high confidence. Using the candidate core clock TFs with known core-clock TFs in *K. fedtschenkoi*, several stomata-related genes were found to be influenced by the clock at the transcriptional level. These clock-controlled stomata-related genes included several genes identified in previous studies to have re-scheduled gene expression relative to their *A. thaliana* ortholog's gene expression (Yang et al. 2017;Moseley et al., manuscript in preparation).

LEM was employed to answer 2 questions: 1) are there new clock TFs in *K. fedtschenkoi* and 2) what are the stomata-related genes regulated by the clock in *K. fedtschenkoi*. Application 1 followed the same method of using LEM to identify new core clock genes in mouse (McGoff et al. 2016). In this study, 4 out of the top 10 genes identified as clock-regulated were found required for normal circadian rhythms in mouse after siRNA knockdowns of each gene in NIH 3T3 cells. The success of this application of LEM in a complex system makes the candidate core clock TFs identified here in the same manner good candidates for future experimental work. Confidence in results produced from application 2 of LEM was established from examining the phase calls of target genes of known core clock TFs used as regulators in LEM. All known core-clock TFs examined had targets with gene expression phased to the expected time of day as reported in the literature for each TF, except for genes orthologous to the components of the EC. Additionally, EC components did not have overlapping gene targets. LEM assumes that each gene in a regulatory network has only one dominant regulator controlling its expression level. However, LEM has been proven to perform well on networks that do contain combinatorial regulation of genes, which is suggested to be caused by LEM identifying at least one of the regulators as the dominant regulator of the target gene (McGoff et al. 2016). Therefore, the target genes predicted for each EC gene can be combined and thought of as target genes for the EC. Evening-phased genes only being predicted as targets for the EC could suggest that the EC regulates stomatal closure by repression of genes involved in stomatal opening. The role of the EC in stomatal closure of *K. fedtschenkoi* needs further examination to test this hypothesis.

Seven TFs were identified that could play a role in the circadian clock of *K. fedtschenkoi* as the transcriptional feedback loops they have with the core clock network is consistent with the feedback-loop architecture of the core clock. Several of the candidate core clock TFs were found to be orthologous to *A. thaliana* TFs that have been reported to be connected to the circadian clock and are involved in various signaling pathways. Only one TF, CTF7, did not have an ortholog in *A. thaliana*, however, it did contain a protein domain known to be associated with proteins involved in circadian rhythms. Additionally, all of the candidate clock TFs were either phased to the morning

or evening which is consistent with a majority of the circadian genes being expressed during these times (Michael et al. 2008; Filichkin et al. 2011). Moreover, a majority (68%) of the rhythmic stomata-related genes were predicted to be regulated by candidate clock TFs (Table 17). It has been suggested that the circadian clock plays a reduced, but still key role in stomatal movement in CAM plants relative to its role in C_3 plants (von Caemmerer and Griffiths 2009). This hypothesis could account for why the known core clock TFs were predicted to regulate only 24% of the rhythmic stomata-related genes. Another reason could be that the candidate core clock TFs are just output from the clock and are used by the clock to integrate the various signaling pathways they control in the clock output network. Nevertheless, the 7 candidate clock TFs serve as good starting points to further investigate the link between the clock and stomatal movement in CAM. Experimental work is needed to determine their essentiality in stomatal movement and the type of relationship they have with the clock. For instance, if the candidate clock TFs feedback into the clock. Protoplasts serve as a great system to test such hypotheses as they can provide a quick measure of circadian rhythm (Hansen and van Ooijen 2016).

The known and candidate core-clock TFs were next used to determine what stomata-related genes are subordinate to the circadian clock and if any of the genes have been previously described as having undergone gene expression rescheduling. Hundreds of genes were predicted to be regulated by the TFs and several genes previously identified as rescheduled (Yang et al. 2017; Moseley et al., manuscript in preparation) were found. Within the clock-controlled stomata-related genes, several GO terms associated with phosphorylation were significantly enriched (Figure 25). Additionally, several rescheduled stomata-related genes were identified as protein kinases (Table 18). Phosphorylation allows for rapid regulation of protein function and is known to play a significant role in stomatal movement. Furthermore, an extensive array of phosphorylation and dephosphorylation events occur in guard cells (Wang et al. 2011). Signaling pathways for stomatal closure, e.g., ABA, and for stomatal opening, e.g., blue light, both involve protein kinases phosphorylating anion channels (ABA signaling: (Mustilli 2002; Vahisalu et al. 2008; Sirichandra et al. 2009; Grondin et al. 2015)) and H^+-ATPases (blue light signaling: (Kinoshita and Shimazaki 1999; Inoue et al. 2008; Wang et al. 2014)). Evidence of gene expression rewiring of protein kinases involved in stomatal movement has also been identified in the CAM plant *Agave Americana* (Abraham et al. 2016). Taken together, these results suggest that the clock played a role in the inversion of stomatal movement, potentially by rescheduling phosphorylation events of stomata-related genes. The next questions to ask are what the direct substrates of these kinases and how do they effect stomatal movement. The channels and ATPase identified in Table 18 serve as good candidates to test as substrates for the protein kinases identified here.

Conclusion

This study presents several gene candidates for further elucidating the circadian clock in the CAM species *K. fedtschenkoi* and suggest that clock-facilitated rescheduling of protein kinases involved in stomatal movement aided the inverted stomatal movement seen in CAM plants. Several questions remain, though, such as how the clock conveys

the circadian signal to the candidate clock TFs, e.g., through direct or indirect regulation, and if the candidate clock TFs are integral to proper functioning of the clock, e.g., would knockouts of the candidates disrupt circadian rhythms in *K. fedtschenkoi.*

References

Abraham PE, Yin H, Borland AM, Weighill D, Lim SD, De Paoli HC, Engle N, Jones PC, Agh R, Weston DJ et al. 2016. Transcript, protein and metabolite temporal dynamics in the CAM plant Agave. *Nat Plants* **2**: 16178.

Anafi RC, Lee Y, Sato TK, Venkataraman A, Ramanathan C, Kavakli IH, Hughes ME, Baggs JE, Growe J, Liu AC et al. 2014. Machine learning helps identify CHRONO as a circadian clock component. *PLoS Biol* **12**: e1001840.

Araus V, Vidal EA, Puelma T, Alamos S, Mieulet D, Guiderdoni E, Gutierrez RA. 2016. Members of BTB Gene Family of Scaffold Proteins Suppress Nitrate Uptake and Nitrogen Use Efficiency. *Plant Physiol* **171**: 1523-1532.

Bencivenga S, Simonini S, Benkova E, Colombo L. 2012. The transcription factors BEL1 and SPL are required for cytokinin and auxin signaling during ovule development in Arabidopsis. *Plant Cell* **24**: 2886-2897.

Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pages F, Trajanoski Z, Galon J. 2009. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**: 1091-1093.

Borland AM, Hartwell J, Weston DJ, Schlauch KA, Tschaplinski TJ, Tuskan GA, Yang X, Cushman JC. 2014. Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci* **19**: 327-338.

Boyle G, Richter K, Priest HD, Traver D, Mockler TC, Chang JT, Kay SA, Breton G. 2017. Comparative Analysis of Vertebrate Diurnal/Circadian Transcriptomes. *PLoS One* **12**: e0169923.

Brem RB, Yvert G, Clinton R, Kruglyak L. 2002. Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* **296**: 752-755.

Cheng XF, Wang ZY. 2005. Overexpression of COL9, a CONSTANS-LIKE gene, delays flowering by reducing expression of CO and FT in Arabidopsis thaliana. *Plant J* **43**: 758-768.

Cook B, Smerdon JE, Seager R, Coats S. 2014. Global warming and 21st century drying. *Climate Dynamics* **43**: 2607-2627.

Covington MF, Maloof JN, Straume M, Kay SA, Harmer SL. 2008. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol* **9**: R130.

Dai A. 2012. Increasing drought under global warming in observations and models. *Nature Climate Change* **3**: 52-58.

de Lichtenberg U, Jensen LJ, Fausboll A, Jensen TS, Bork P, Brunak S. 2005. Comparison of computational methods for the identification of cell cycle-regulated genes. *Bioinformatics* **21**: 1164-1171.

Dong B, Zhang P, Chen X, Liu L, Wang Y, He S, Chen R. 2011. Predicting Housekeeping Genes Based on Fourier Analysis. *PloS ONE* **6**: 1-11.

Ehleringer JR, Monson RK. 1993. EVOLUTIONARY AND ECOLOGICAL ASPECTS OF PHOTOSYNTHETIC PATHWAY VARIATION. *Annu Rev Ecol Syst* **24**.

Filichkin SA, Breton G, Priest HD, Dharmawardhana P, Jaiswal P, Fox SE, Michael TP, Chory J, Kay SA, Mockler TC. 2011. Global profiling of rice and poplar

transcriptomes highlights key conserved circadian-controlled pathways and cis-regulatory modules. *PLoS One* **6**: e16907.
Fonseca S, Fernandez-Calvo P, Fernandez GM, Diez-Diaz M, Gimenez-Ibanez S, Lopez-Vidriero I, Godoy M, Fernandez-Barbero G, Van Leene J, De Jaeger G et al. 2014. bHLH003, bHLH013 and bHLH017 are new targets of JAZ repressors negatively regulating JA responses. *PLoS One* **9**: e86182.
Fornara F, Panigrahi K, Gissot L, Sauerbrunn N, Ruhl M, Jarillo J, Coupland G. 2009. Arabidopsis DOF transcription factors act redundantly to reduce CONSTANS expression and are essential for a photoperiodic flowering response. *Developmental Cell* **17**: 75-86.
Gibalova A, Renak D, Matczuk K, Duplakova N, Chab D, Twell D, Honys D. 2009. AtbZIP34 is required for Arabidopsis pollen wall patterning and the control of several metabolic pathways in developing pollen. *Plant Molecular Biology* **70**: 581-601.
Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* **40**: D1178-1186.
Grondin A, Rodrigues O, Verdoucq L, Merlot S, Leonhardt N, Maurel C. 2015. Aquaporins Contribute to ABA-Triggered Stomatal Closure through OST1-Mediated Phosphorylation. *Plant Cell* **27**: 1945-1954.
Grundy J, Stoker C, Carre IA. 2015. Circadian regulation of abiotic stress tolerance in plants. *Front Plant Sci* **6**: 648.
Hansen LL, van Ooijen G. 2016. Rapid Analysis of Circadian Phenotypes in Arabidopsis Protoplasts Transfected with a Luminescent Clock Reporter. *J Vis Exp* doi:10.3791/54586.
Hartwell J. 2005. The co-ordination of central plant metabolism by the circadian clock. *Mechanistic and Functional Studies of Proteins* **33**: 945-948.
Hartwell J. 2006. The circadian clock in CAM plants. *Endogenous Plant Rhythms*: 211-236.
Hazen SP, Schultz J, Pruneda-Paz J, Borevitz J, Ecker J, Kay SA. 2005. LUX ARRHYTHMO encodes a Myb domain protein essential for circadian rhythms. *PNAS* **102**: 10387-10392.
Helfer A, Nusinow DA, Chow BY, Gehrke AR, Bulyk ML, Kay SA. 2011. LUX ARRHYTHMO encodes a nighttime repressor of circadian gene expression in the Arabidopsis core clock. *Curr Biol* **21**: 126-133.
Holm S. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6**: 65-70.
Huang da W, Sherman BT, Lempicki RA. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**: 1-13.
Huang da W, Sherman BT, Lempicki RA. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**: 44-57.

Huang W, Perez-Garcia P, Pokhilko A, Millar AJ, Antoschechkin I, Riechmann J, Mas P. 2012. Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator. *Science* **336**: 75-79.

Hughes ME, Hogenesch JB, Kornacker K. 2010. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms* **25**: 372-380.

Imaizumi T, Kay SA. 2006. Photoperiodic control of flowering: not only by coincidence. *Trends in Plant Science* **11**: 550-558.

Imaizumi T, Schultz T, Harmon F, Ho L, Kay S. 2005. FKF1 F-box protein mediates cyclic degradation of a repressor of CONSTANS in Arabidopsis. *Science* **309**: 293-297.

Inoue S, Kinoshita T, Matasumoto M, Nakayama K, Doi M, Shimazaki K. 2008. Blue light-induced autophosphorylation of phototropin is a primary step for signaling. *PNAS* **105**: 5626-5631.

Ito S, Niwa T, Nakamichi N, Kawamura H, Yamashino T, Mizuno T. 2008. Insight into missing genetic links between two evening-expressed pseudo-response regulator genes TOC1 and PRR5 in the circadian clock-controlled circuitry in Arabidopsis thaliana. *Plant Cell Physiol* **49**: 201-213.

Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F. 2001. bZIP transcription factors in Arabidopsis. *Trends in Plant Science* **7**: 106-111.

Keurentjes J, Fu J, Terpstra I, Garcia J, can den Ackerveken G, Snoek L, Peeters A, Vreugdenhil D, Koornneef M, Jansen R. 2007. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *PNAS* **104**: 1708-1713.

Kinoshita T, Shimazaki K. 1999. Blue light activates the plasma membrane H+-ATPase by phosphorylation of the C-terminus in stomatal guard cells. *The EMBO Journal* **18**: 5548-5558.

Ledger S, Strayer C, Ashton F, Kay SA, Putterill J. 2001. Analysis of the function of two circadian-regulated CONSTANS-LIKE genes. *The Plant Journal* **26**: 15-22.

Li H, Sun J, Xu Y, Jiang H, Li C. 2007. The bHLH-type transcription factor AtAIB positively regulates ABA response in Arabidopsis. *Plant Molecular Biology* **65**: 655-665.

Luo Z, Azencott R, Zhao Y. 2014. Modeling miRNA-mRNA interactions: fitting chemical kinetics equations to microarray data. *BMC Systems Biology* **9**: 1-15.

Mandadi KK, Misra A, Ren S, McKnight TD. 2009. BT2, a BTB protein, mediates multiple responses to nutrients, stresses, and hormones in Arabidopsis. *Plant Physiol* **150**: 1930-1939.

McGoff KA, Guo X, Deckard A, Kelliher CM, Leman AR, Francey LJ, Hogenesch JB, Haase SB, Harer JL. 2016. The Local Edge Machine: inference of dynamic models of gene regulation. *Genome Biol* **17**: 214.

Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM et al. 2008. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. *PLoS Genet* **4**: e14.

Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, Lyons E, Wang ML, Chen J, Biggers E et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* **47**: 1435-1442.

Mizuno T, Nomoto Y, Oka H, Kitayama M, Takeuchi A, Tsubouchi M, Yamashino T. 2014. Ambient temperature signal feeds into the circadian clock transcriptional circuitry through the EC night-time repressor in Arabidopsis thaliana. *Plant Cell Physiol* **55**: 958-976.

Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J. 2007. The Diurnal Project: Diurnal and Circadian Expression Profiling, Model-based Pattern Matching, and Promoter Analysis. *Cold Spring Harbor Symposia on Quantitative Biology* **72**: 353-363.

Mustilli AC. 2002. Arabidopsis OST1 Protein Kinase Mediates the Regulation of Stomatal Aperture by Abscisic Acid and Acts Upstream of Reactive Oxygen Species Production. *The Plant Cell Online* **14**: 3089-3099.

Nakamichi N, Kiba T, Kamioka M, suzuki T, Yamashino T, Higashiyama T, Skakibara H, Mizuno T. 2012. Transcriptional repressor PRR5 directly regulates clock-output pathways. *PNAS* **109**: 17123-17128.

Nakamichi N, Kita M, Niinuma K, Ito S, Yamashino T, Moizgouchi T, Mizuno T. 2007. Arabidopsis clock-associated pseudo-response regulators PRR9, PRR7 and PRR5 coordinately and positively regulate flowering time through the canonical CONSTANS-dependent photoperiodic pathway. *Plant Cell Physiology* **48**: 822-832.

Nakata M, Mitsuda N, Herde M, Koo AJ, Moreno JE, Suzuki K, Howe GA, Ohme-Takagi M. 2013. A bHLH-type transcription factor, ABA-INDUCIBLE BHLH-TYPE TRANSCRIPTION FACTOR/JA-ASSOCIATED MYC2-LIKE1, acts as a repressor to negatively regulate jasmonate signaling in arabidopsis. *Plant Cell* **25**: 1641-1656.

Niwa T, Ito S, Nakamichi N, Mizoguchi T, Niinuma K, Yamashino T, Mizuno T. 2007. Genetic linkages of the circadian-clock associated genes, TOC1, CCA1 and LHY, in the photoperiodic control of flowering time in Arabidopsis thaliana. *Plant Cell Physiol* **48**: 925-937.

Nohales MA, Kay SA. 2016. Molecular mechanisms at the core of the plant circadian oscillator. *Nat Struct Mol Biol* **23**: 1061-1069.

Nusinow DA, Helfer A, Hamilton EE, King JJ, Imaizumi T, Schultz TF, Farre EM, Kay SA. 2011. The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature* **475**: 398-402.

Seaton DD, Smith RW, Song YH, MacGregor DR, Stewart K, Steel G, Foreman J, Penfield S, Imaizumi T, Millar AJ et al. 2015. Linked circadian outputs control elongation growth and flowering in response to photoperiod and temperature. *Mol Syst Biol* **11**: 776.

Shalaby N, Pinzon J, Narayanan A, Jin E, Ritz M, Dove R, Wolfenberg H, Rodan A, Buszczak M, Rothenfluh A. 2018. JmjC domain proteins modulate circadian behaviors and sleep in Drosophila. *Scientific Reports* **8**: 1-9.

Sharma A, Wai CM, Ming R, Yu Q. 2017. Diurnal Cycling Transcription Factors of Pineapple Revealed by Genome-Wide Annotation and Global Transcriptomic Analysis. *Genome Biol Evol* **9**: 2170-2190.

Sirichandra C, Gu D, Hu HC, Davanture M, Lee S, Djaoui M, Valot B, Zivy M, Leung J, Merlot S et al. 2009. Phosphorylation of the Arabidopsis AtrbohF NADPH oxidase by OST1 protein kinase. *FEBS Lett* **583**: 2982-2986.

Song YH, Smith RW, To BJ, Millar AJ, Imaizumi T. 2012. FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. *Science* **336**: 1045-1049.

Vahisalu T, Kollist H, Wang YF, Nishimura N, Chan WY, Valerio G, Lamminmaki A, Brosche M, Moldau H, Desikan R et al. 2008. SLAC1 is required for plant guard cell S-type anion channel function in stomatal signalling. *Nature* **452**: 487-491.

von Caemmerer S, Griffiths H. 2009. Stomatal responses to CO2 during a diel Crassulacean acid metabolism cycle in Kalanchoe daigremontiana and Kalanchoe pinnata. *Plant Cell Environ* **32**: 567-576.

Wai CM, VanBuren R, Zhang J, Huang L, Miao W, Edger PP, Yim WC, Priest HD, Meyers BC, Mockler T et al. 2017. Temporal and spatial transcriptomic and microRNA dynamics of CAM photosynthesis in pineapple. *Plant J* **92**: 19-30.

Wang R, Pandey S, Li S, Gookin T, Zhao Z, Albert R, Assman S. 2011. Common and unique elements of the ABA-regulated transcriptome of Arabidopsis guard cells. *BMC Genomics* **12**.

Wang Y, Noguchi K, Ono N, Inoue S, Terashima I, Kinoshita T. 2014. Overexpression of plasma membrane H+-ATPase in guard cells promotes light-induced stomatal opening and enhances plant growth. *Proc Natl Acad Sci U S A* **111**: 533-538.

Yang X, Cushman JC, Borland AM, Edwards EJ, Wullschleger SD, Tuskan GA, Owen NA, Griffiths H, Smith JA, de Paoli HC et al. 2015. A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytologist* **207**: 491-504.

Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, Liu D, Weigel D, Cheol Yim W, Ha J et al. 2017. The Kalanchoë genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. *Nature Communications* **8**: 1-15.

Zhang R, Lahens N, Ballance H, Hughes ME, Hogenesch JB. 2014. A circadian gene expression atlas in mammals: Implications for biology and medicine. *PNAS* **111**: 16219-16224.

Zhang X, Chen Y, Wang ZY, Chen Z, Gu H, Qu LJ. 2007. Constitutive expression of CIR1 (RVE2) affects several circadian-regulated processes and seed germination in Arabidopsis. *Plant J* **51**: 512-525.

Appendix

Table 15. Genes identified as clock-regulated in *Kalanchoë fedtschenkoi* that are orthologous to circadian-related *Arabidopsis thaliana* genes.

K. fedtschenkoi Gene ID	*A. thaliana* Gene ID	*A. thaliana* Gene Symbol	*A. thaliana* Gene Description	PubMed ID
Kaladp0674s0030	AT3G48360	BT2	BTB and TAZ domain protein 2	PMC2719139
Kaladp0032s0115	AT2G46790	PRR9	pseudo-response regulator 9	15705949
Kaladp0032s0115	AT5G24470	PRR5	two-component response regulator-like protein	20233950
Kaladp1154s0002	AT5G62430	CDF1	cycling DOF factor 1	25600997
Kaladp0550s0052	AT2G27990	BLH8	BEL1-like homeodomain 8	21115819
Kaladp0151s0002	AT3G20810	JMJD5	2-oxoglutarate (2OG) and Fe (II)-dependent oxygenase superfamily protein	17587236
Kaladp0036s0161	AT5G37260	RVE2/CIR1	Homeodomain-like superfamily protein	21205033
Kaladp0577s0020	AT3G09600	RVE8	Homeodomain-like superfamily protein	PMC3639509
Kaladp0095s0568	AT5G62470	MYB96	myb domain protein 96	PMC3057456
Kaladp0033s0047	AT3G46640	LUX (PCL1)	Homeodomain-like superfamily protein	PMC3082275
Kaladp0088s0066	AT5G43270	SPL2	squamosa promoter binding protein-like 2	PMC4807768
Kaladp0039s0012	AT3G21890	BBX31	B-box type zinc finger family protein	PMC4807768
Kaladp0039s0012	AT4G15248	BBX30	B-box type zinc finger family protein	PMC4213167

Table 16. Candidate clock transcription factors in *Kalanchoë fedtschenkoi.*

"Kafe shift" is the number of h the *K. fedtschenkoi* gene expression profile shifted from its *A. thaliana* ortholog's gene expression profile. The shift was calculated by subtracting the phases calls of each ortholog. "Spear shift" is the Spearman rank correlation coefficient between orthologs after shifting the *K. fedtschenkoi* gene expression profile by the "Kafe shift".

Candidate #	Kafe Gene ID	Kafe Phase Call	Arth Ortholog ID	Arth Gene Symbol	Arth Gene Desc.	Arth Phase Call	Kafe Shift (hrs)	Spear Shift
1	Kaladp0748s0043	6	AT5G41410	BEL1	POX (plant homeobox) family protein	12	-6	0.97
2	Kaladp0007s0017	14	AT2G46510	JAM1	ABA-inducible BHLH-type transcription factor	12	+2	0.91
2	Kaladp0007s0017	14	AT1G01260	JAM2	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	8	+6	0.83
3	Kaladp0011s1342	22	AT2G42380	BZIP34	Basic-leucine zipper (bZIP) transcription factor family protein	4	-8	0.96
3	Kaladp0011s1342	22	AT3G58120	BZIP61	Basic-leucine zipper (bZIP) transcription factor family protein	8	-8	0.97
4	Kaladp1154s0002	22	AT3G47500	CDF3	cycling DOF factor 3	2	-4	0.97
4	Kaladp1154s0002	22	AT5G62430	CDF1	cycling DOF factor 1	24	-2	0.97
4	Kaladp1154s0002	22	AT5G39660	CDF2	cycling DOF factor 2	24	-2	0.87
5	Kaladp0674s0030	6	AT5G63160	BT1	BTB and TAZ domain protein 1			
5	Kaladp0674s0030	6	AT3G48360	BT2	BTB and TAZ domain protein 2	20	+10	0.94
6	Kaladp0878s0025	8	AT1G07050		CCT motif family protein	12	-2	0.92
7	Kaladp0009s0042	22						

Table 17. Regulation relationships between clock transcription factors and stomata-related genes in *Kalanchoë fedtschenkoi.*

Clock TFs	Annotated or Known Stomata-related Genes			New Stomata-related Genes			Grand Total
	Activates	Represses	Total	Activates	Represses	Total	
Kaladp0007s0017 (CTF2)	5	1	6	83	4	87	93
Kaladp0674s0030 (CTF5)	5	6	11	43	32	75	86
Kaladp0878s0025 (CTF6)	6		6	52	1	53	59
Kaladp0748s0043 (CTF1)	1	1	2	37	16	53	55
Kaladp0060s0264 (LNK2)	5		5	50		50	55
Kaladp0099s0129 (LNK2)	6		6	45		45	51
Kaladp1154s0002 (CTF4)		4	4	1	35	36	40
Kaladp0009s0042 (CTF7)	1	1	2	9	22	31	33
Kaladp0011s1342 (CTF3)	3	2	5	7	16	23	28
Kaladp0607s0046 (LNK1)	1		1	16		16	17
Kaladp0047s0123 (LNK1)				16		16	16
Kaladp0101s0041 (PRR7)		1	1		12	12	13
Kaladp0059s0037 (ELF4)					10	10	10
Kaladp0496s0018 (CCA1/LHY)					4	4	4
Kaladp0032s0115 (PRR9/5)					4	4	4
Kaladp0066s0115 (CCA1/LHY)					4	4	4
Kaladp0033s0047. (LUX/NOX)					3	3	3
Kaladp0055s0349 (RVE6)				2		2	2
Kaladp0022s0168 (RVE6)				2		2	2
Kaladp0577s0020 (RVE8)				1		1	1
Kaladp0037s0163 (ELF4)					1	1	1
Kaladp0040s0446 (TOC1)					1	1	1
Kaladp0039s0732 (ELF3)					1	1	1
Kaladp0032s0054 (CHE)					1	1	1
Kaladp0019s0045 (RVE6)				1		1	1
Kaladp0045s0206 (ELF4)					1	1	1
Grand Total	33	16	49	365	168	533	582

Table 18. Clock-controlled stomata-related genes with rescheduled gene expression from different studies.

Clock TF	Relationship	Study	Target	Stomata-related	Gene Description
Kaladp0748s0043 (CTF1)	Represses	Yang et al., 2017	Kaladp0059s0048	New	aquaporin pip1-2
Kaladp0748s0043 (CTF1)	Represses	Moseley et al.	Kaladp0062s0167	New	receptor-like protein kinase haiku2
Kaladp0007s0017 (CTF2)	Activates	Yang et al., 2017	Kaladp0011s0363	New	trehalose-phosphate synthase
Kaladp0007s0017 (CTF2)	Activates	Moseley et al.	Kaladp0092s0084	Known/GO	calcium-dependent protein kinase 26
Kaladp0011s1342 (CTF3)	Represses	Yang et al., 2017	Kaladp0040s0264	New	btb poz domain-containing protein npy2-like
Kaladp0011s1342 (CTF3)	Represses	Yang et al., 2017	Kaladp0008s0539	New	mitogen-activated protein kinase
Kaladp0011s1342 (CTF3)	Represses	Yang et al., 2017	Kaladp0033s0113	Known/GO	phototropin-2
Kaladp0011s1342 (CTF3)	Represses	Moseley et al.	Kaladp0001s0016	New	catalase isozyme 1
Kaladp0011s1342 (CTF3)	Represses	Moseley et al.	Kaladp0093s0030	New	(TF) homeobox-leucine zipper protein anthocyaninless 2 isoform x1
Kaladp0059s0037 (ELF4)	Represses	Moseley et al.	Kaladp0062s0076	New	3-ketoacyl- synthase 19-like
Kaladp0047s0123 (LNK1)	Activates	Moseley et al.	Kaladp0008s0414	New	cyclic nucleotide-gated ion channel 15
Kaladp0060s0264 (LNK2)	Activates	Moseley et al.	Kaladp0024s0371	New	pectin lyase-like superfamily protein isoform 1
Kaladp0099s0129 (LNK2)	Activates	Moseley et al.	Kaladp0042s0353	Known	abscisic acid receptor pyl8-like
Kaladp0060s0264 (LNK2)	Activates	Moseley et al.	Kaladp0095s0634	GO	mitogen-activated protein kinase homolog mmk2
Kaladp0101s0041 (PRR7)	Represses	Moseley et al.	Kaladp0092s0115	New	pleiotropic drug resistance protein 1-like
Kaladp0055s0349 (RVE6)	Activates	Moseley et al.	Kaladp0043s0103	New	phospholipid-transporting atpase 3
Kaladp0055s0349 (RVE6)	Activates	Moseley et al.	Kaladp0090s0003	New	receptor-like protein kinase

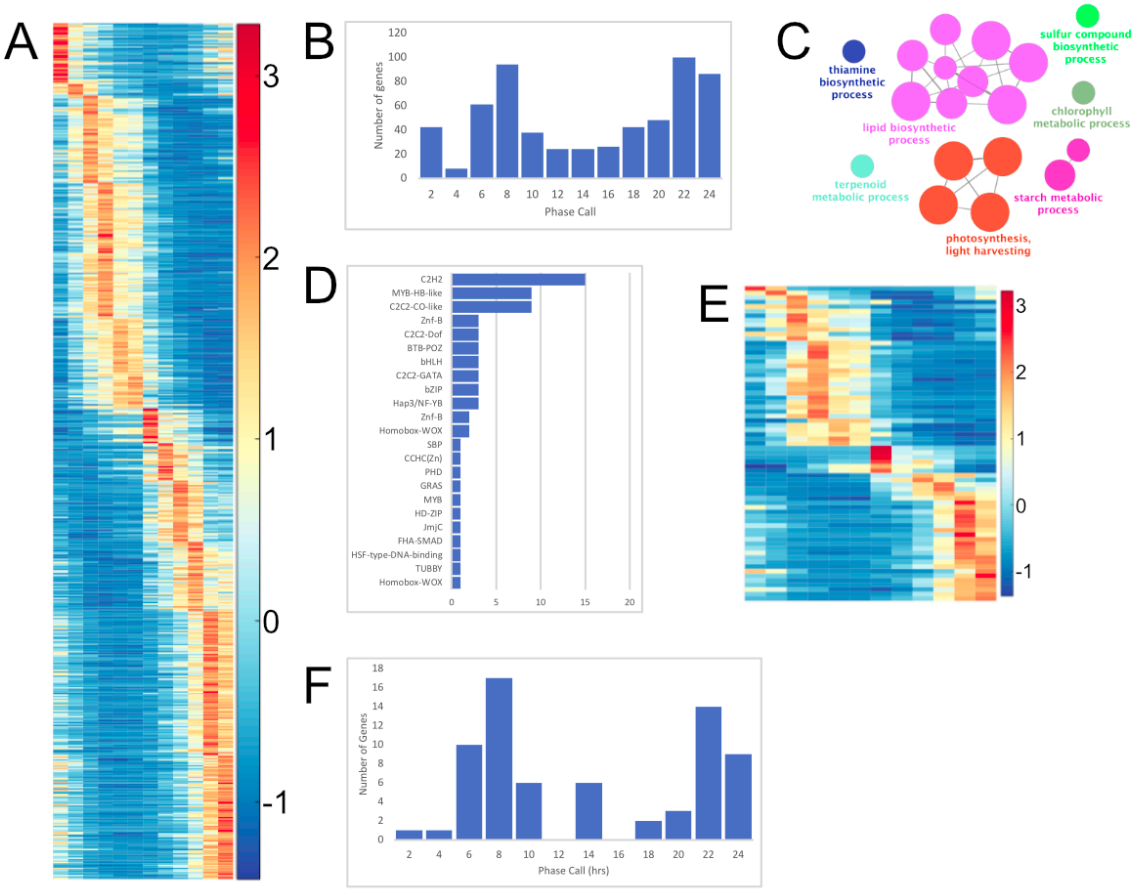


Figure 19. DLxJTK rank ordered *Kalanchoë fedtschenkoi* genes.

A) Heatmap displaying z-score standardized expression profiles of the top 582 genes ranked ordered by DLxJTK. **B)** Phase call distribution of the top 582 genes ranked ordered by DLxJTK. **C)** Enriched functional groups within the top 582 genes ranked ordered by DLxJTK. **D)** Abundance of protein domains in the 68 transcription factors identified in the top 582 genes ranked ordered by DLxJTK. **E)** Heatmap displaying z-score standardized expression profiles of 68 transcription factors. **F)** Phase call distributions of the 68 transcription factors.

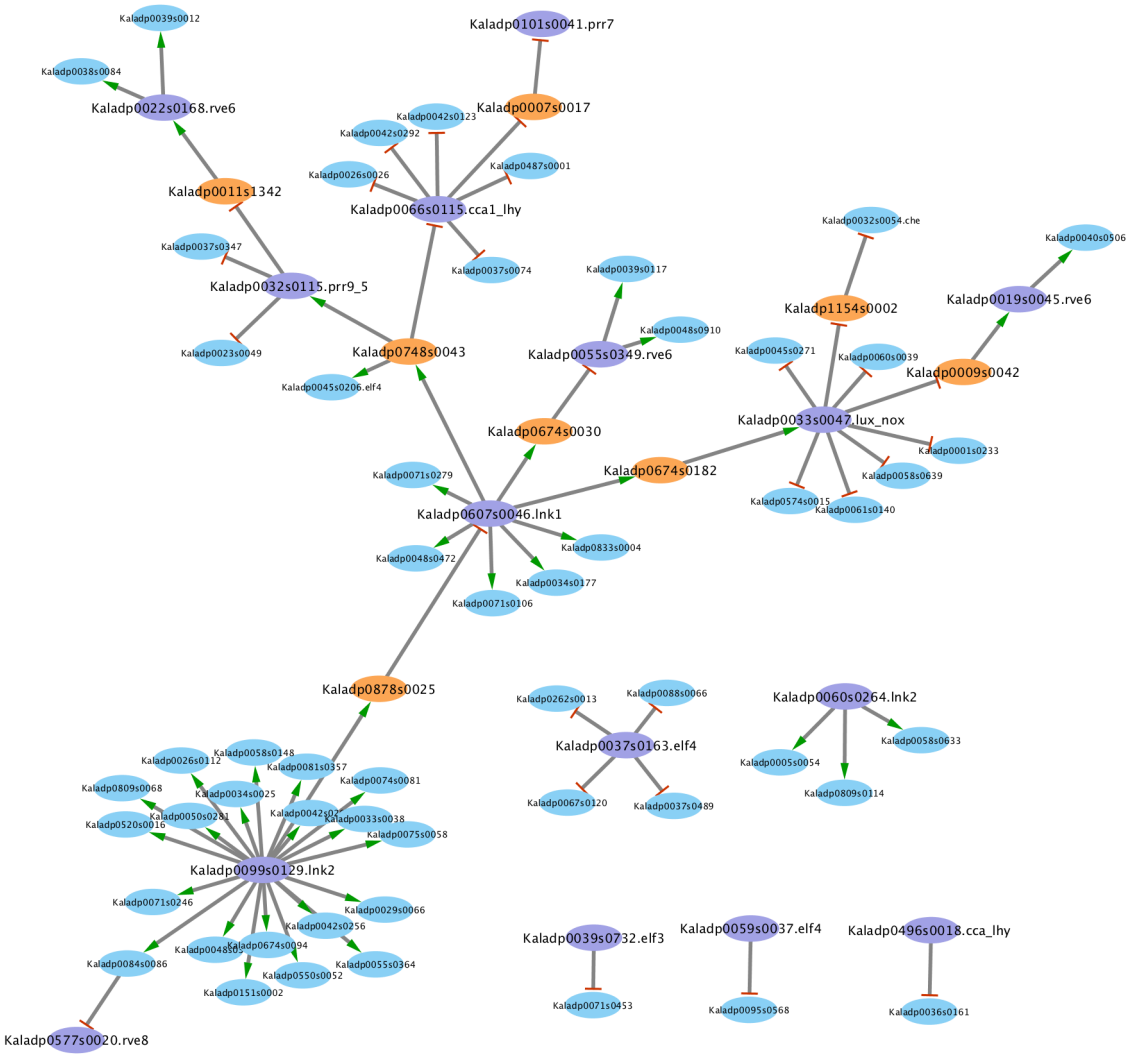


Figure 20. Several candidate clock transcription factors are predicted to regulate and be regulated by known core clock transcription factors.

Orange ovals represent candidate clock transcription factors predicted to regulate and be regulated by known core clock transcription factors. Purple ovals represent known clock transcription factors. Blue ovals represent candidate clock transcription factors that are predicted to only be regulated by known core clock transcription factors. Edges with a green arrow represent activation of gene expression. Edges with red lines represent repression of gene expression.

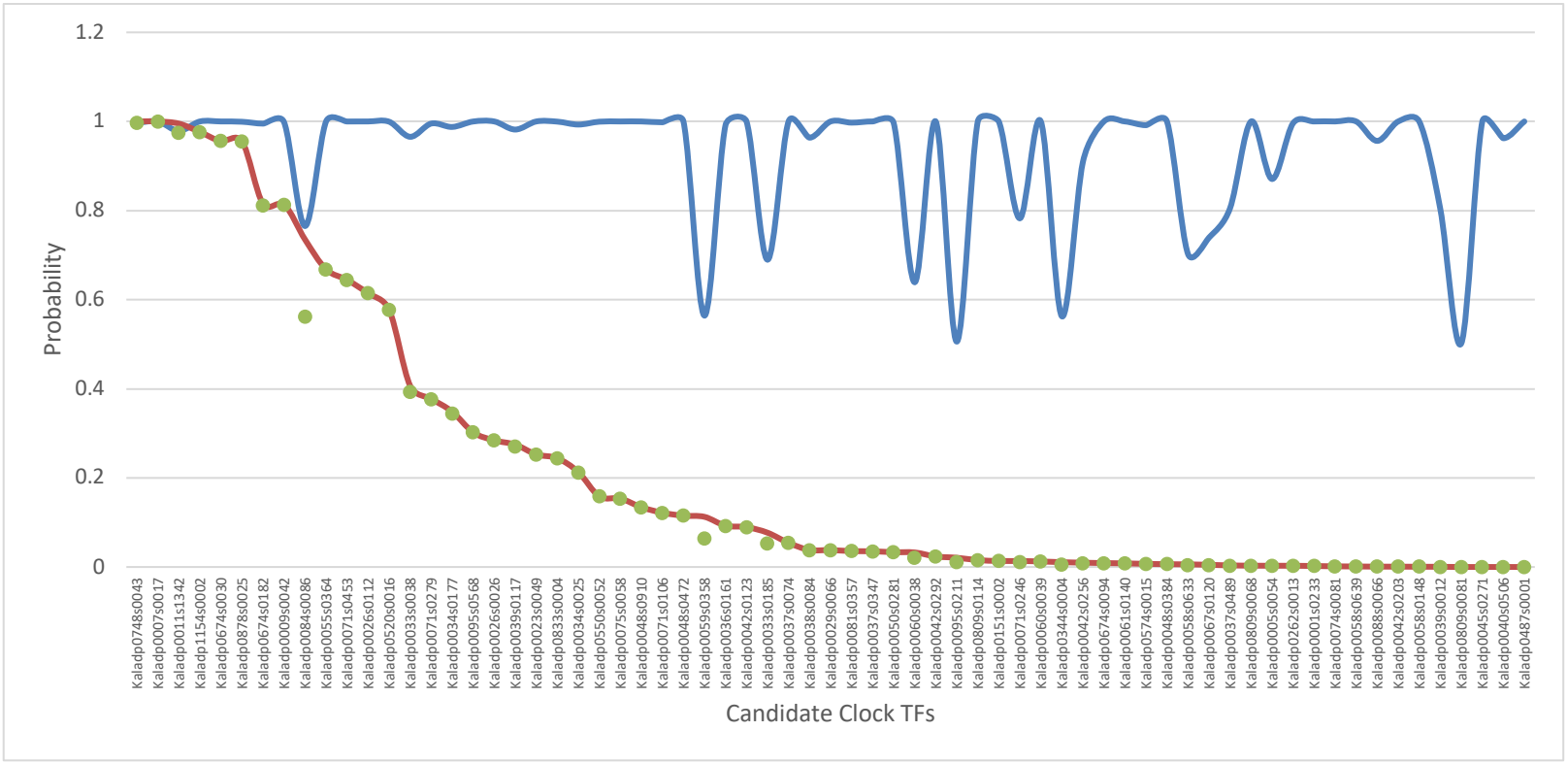


Figure 21. Likelihood rankings of candidate clock transcription factors.

Blue line is the probability that any known core clock transcription factor regulates the candidate clock transcription factor. Red line is the probability that the candidate clock transcription factor regulates any known core-clock transcription factor. Green circle is the likelihood probability.

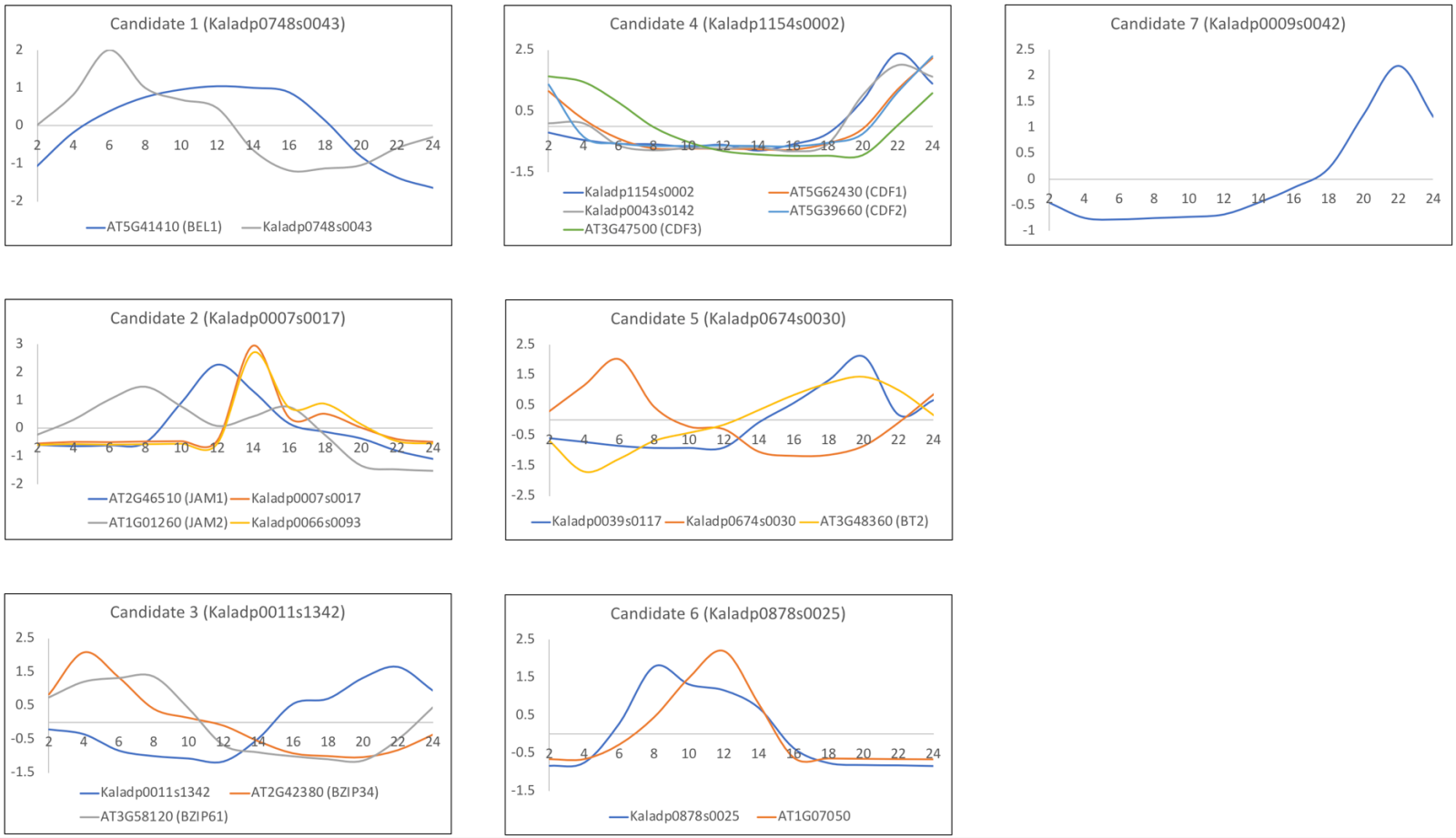


Figure 22. Gene expression of *Kalanchoë fedtschenkoi* candidate transcriptions and their orthologs in *Arabidopsis thaliana*.
Y-axis is z-score standardized gene expression and x-axis is hours after first light.

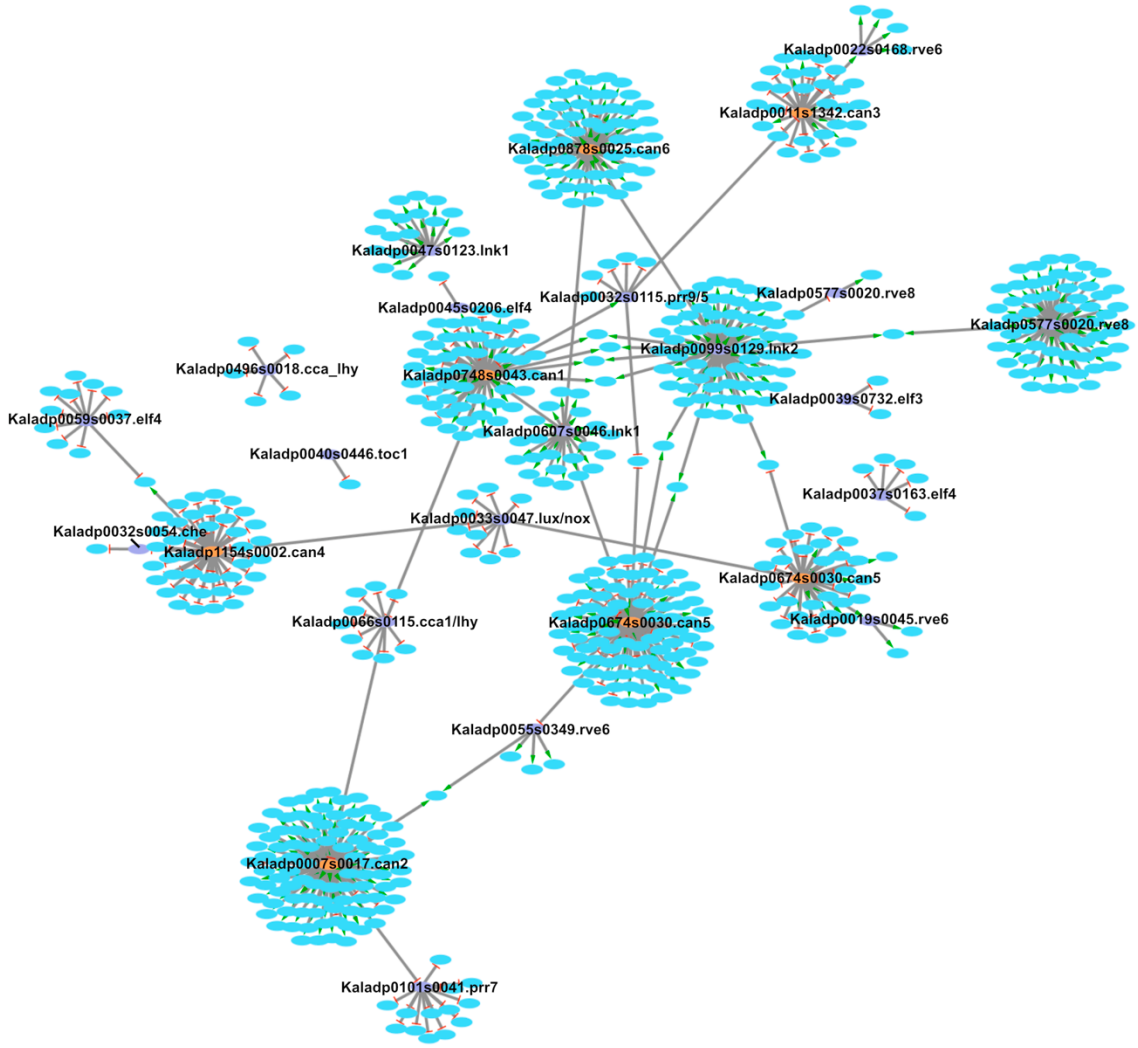


Figure 23. *Kalanchoë fedtschenkoi* clock transcription factors and stomata-related genes regulatory network.

Orange ovals represent candidate clock transcription factors predicted to regulate and be regulated by known core clock transcription factors. Purple ovals represent known clock transcription factors. Blue ovals represent stomata-related genes. Edges with a green arrow represent activation of gene expression. Edges with red lines represent repression of gene expression.

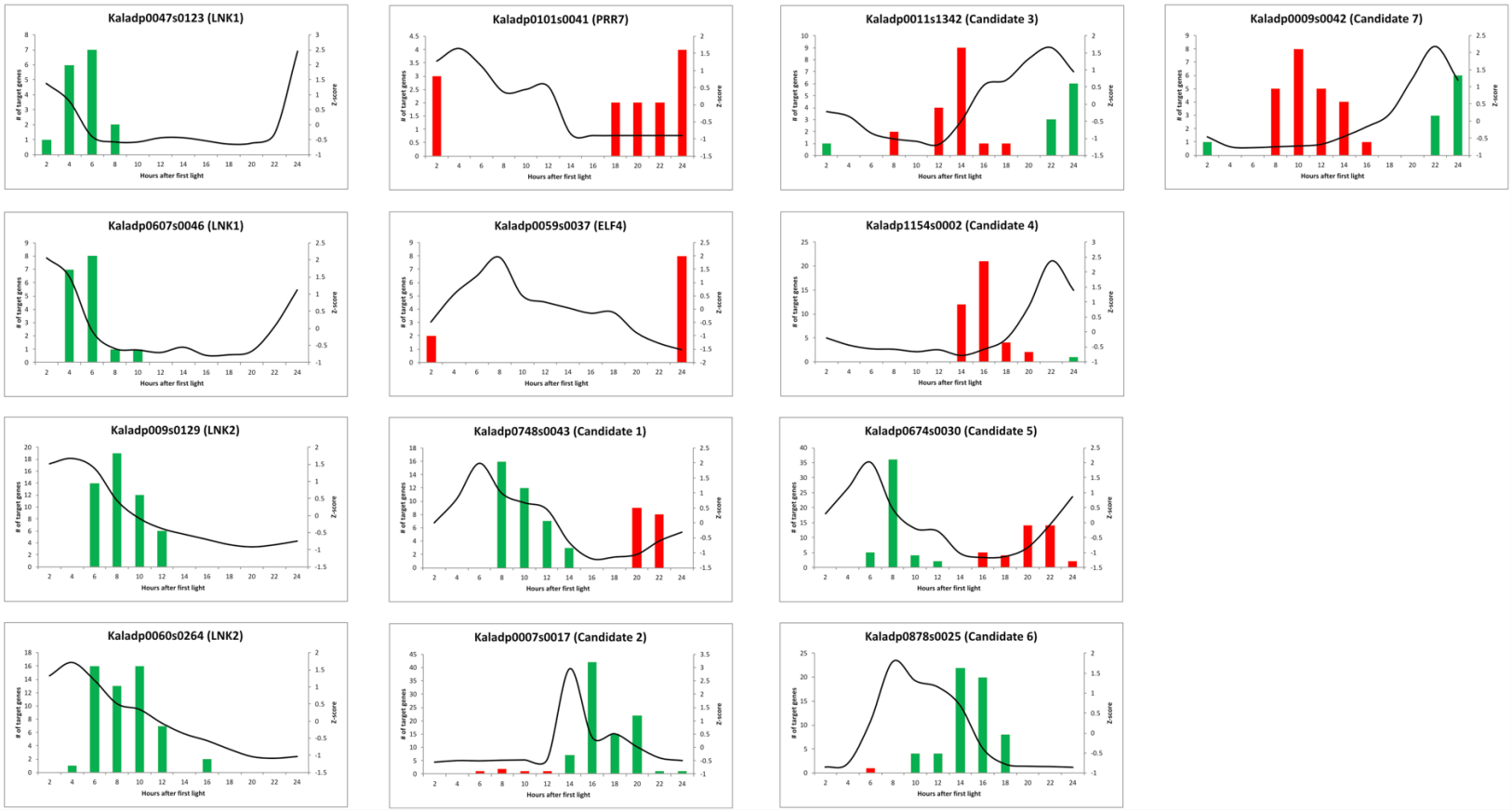


Figure 24. *Kalanchoë fedtschenkoi* candidate transcription gene expression and phase calls of their respective targets.

Red bars signify that the target genes are repressed by the respective regulator and green bars signify that the target genes are activated by the respective regulator.

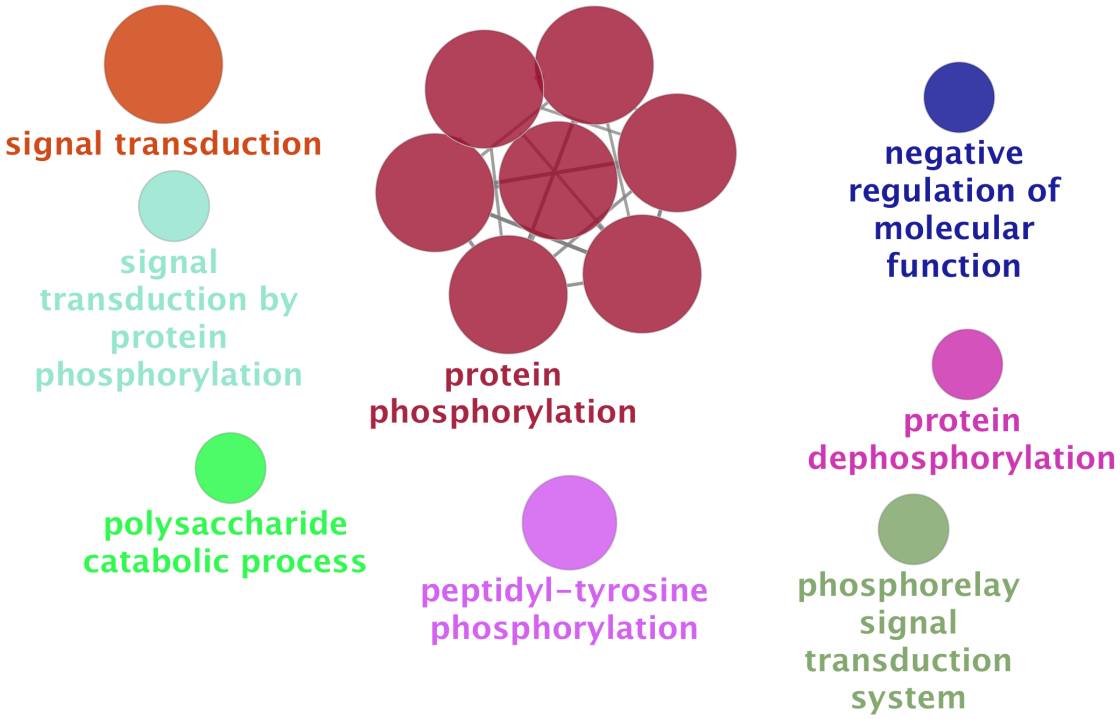


Figure 25. Enriched functional groups within of clock-controlled stomata-related genes in *Kalanchoë fedtschenkoi*.

CONCLUSION

The focus of the dissertation was on the circadian clock in the CAM plant *Kalanchoë fedtschenkoi* and how it controls the physiology of the plant. Through use of various omics resources and computational methods and tools, several testable hypotheses on how the clock impacted CAM evolution are presented, along with several resources to facilitate future experiments. Additional resources are recommended for generation, such as 48-h time-course RNAseq for model CAM and C_3 plants, following the newly defined guidelines. These resources will be of great importance and use for the plant circadian community as they will allow for improved statistical power of downstream analyses thus giving more confidence in the insights derived.

Moreover, the core-clock network of *K. fedtschenkoi* needs further examination to confirm if expression dynamics of its components are truly altered and then to determine the hierarchal structure of tissue-specific clocks. Recent studies examining *A. thaliana* tissue-specific clocks provide promising methods to apply to *K. fedtschenkoi*. Additionally, candidate core clock genes are presented that warrant validation. Protoplasts offer a quick, scalable expression system to test circadian gene expression, as well as to test feedback into the core clock network. The results produced from these studies will help further define the control mechanism of the circadian output network opening avenues for new studies into differentiating circadian-controlled CAM processes from the circadian-controlled C_3 background.

Lastly, future work on stomata in CAM plants should probe deeper into the impact of protein kinases on stomatal movement. Furthermore, elucidating their role in the inverted stomatal movement of CAM plants will help guide efforts into engineering CAM properties, such as drought tolerance, into non-CAM plants. In conclusion, the work in this dissertation helps further our understanding of how CAM evolved in the context of the circadian clock and in stomatal movement and will aid efforts into engineering CAM pathways into non-CAM bioenergy and agricultural crops.

VITA

Robert C. Moseley was born in Stone Mountain, GA to Nanci and Clifford Moseley and has one older brother, Scott McElroy, and one older sister Holly Wendel. He and his family later moved to Snellville, GA, where Robert attended Shiloh High School. He received his undergraduate degree in Environmental Science in 2010 and a Master's degree in Biology in 2012 at Georgia College. For the following year, he was a fulltime lecturer at Georgia College where he met his fiancé, Cierra Cathey. After the one-year term, Robert accepted a fellowship through the Bredesen Center for Interdisciplinary Research and Education at the University of Tennessee in 2013 where he earned his Doctor of Philosophy in Energy Science and Engineering. During this time, he worked in the lab of Dr. Xiaohan Yang at Oak Ridge National Laboratory ORNL in the Plant Systems Biology group. He studied the circadian clock and its relationship to crassulacean acid metabolism and how it played a role in the evolution of CAM. Also, during this time, Robert created a startup, SimPath, that leverages a cloning technology created at ORNL. Robert has a passion for science and entrepreneurship and plans to pursue a post-doctorate and continue running his company.