USE AND IMPROVEMENT OF TRANSCRIPTOMIC METHODS FOR THE TEMPORAL STUDY OF CRASSULACEAN ACID METABOLISM (CAM) IN THE GENUS *ERYCINA*

by

MICHELLE HWANG

(Under the Direction of Jim Leebens-Mack)

ABSTRACT

The RNA sequencing (RNA-Seq) analysis pipeline can be complex and variable between transcriptomic studies, and therefore necessitates the need for optimizing methodology to achieve the best results. Further, few transcriptomic methods have been used to analyze the regulation of the Crassulacean acid metabolism (CAM) photosynthesis pathway, much less comparatively against a non-CAM species or in the Orchidaceae. Here we perform a temporal gene expression analysis between CAM species *Erycina pusilla* and close C_3 relative *Erycina crista-galli*, and establish an optimal methodology to assemble and analyze time series RNA-Seq data for the purpose of studying the evolution and regulation of CAM photosynthesis. Gene clustering for temporal patterns reveals a network of genes associated with canonical CAM gene phosphoenolpyruvate carboxylase (PEPC) present in both *E. pusilla* and *E. crista-galli*, suggesting a possible ancestral predisposition for CAM.

INDEX WORDS: Crassulacean acid metabolism, transcriptomics, assembly, Erycina, gene clustering

USE AND IMPROVEMENT OF TRANSCRIPTOMIC METHODS FOR THE TEMPORAL STUDY OF CRASSULACEAN ACID METABOLISM (CAM) IN THE GENUS *ERYCINA*

by

MICHELLE HWANG

BS, University of Michigan, 2015

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2017

© 2017

Michelle Hwang

All Rights Reserved

USE AND IMPROVEMENT OF TRANSCRIPTOMIC METHODS FOR THE TEMPORAL STUDY OF CRASSULACEAN ACID METABOLISM (CAM) IN THE GENUS *ERYCINA*

by

MICHELLE HWANG

BS, The University of Michigan, 2015

Major Professor:	Jim Leebens-Mack
Committee:	Magdy Alabady
	Chung-Jui Tsai

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
May 2017

ACKNOWLEDGEMENTS

Orchid plants were generously provided by Ming-Tsair Chan at the Academia Sinica – Biotechnology Center in Southern Taiwan, Tainan, Taiwan. Tissue sampling for *E. pusilla* was conducted by our friends at the University of Buffalo, Victor Albert and Kimberly Farr. For the titratable acidity measurements and sampling of *E. crista-galli* plants, I thank Katia Silvera and the Smithsonian Tropical Research Institute in Panama. I'd like to thank the Georgia Genomics Facility (GGF) at the University of Georgia for performing our sequencing runs. Also, the members of the University of Georgia's Georgia Advanced Computing Resourcing Center (GACRC) worked tirelessly to ensure the computing clusters were working optimally. Thanks to Jeremy Ray, our lab technician, for help in sequencing preparation. Thank you to Karolina Heyduk for substantial advice on CAM and research methodology throughout the project. Funding was provided by the National Science Foundation (DEB 1442199 to Jim Leebens-mack). A special thank you to Jim Leebens-mack and my committee members Magdy Alabady and Chung-Jui Tsai for their time, support, and intellectual contributions.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

CRASSULACEAN ACID METABOLISM (CAM) AND THE ORCHIDACEAE

Crassulacean acid metabolism (CAM) photosynthesis evolved from C_3 photosynthesis as a means for plants to survive in drought stress or conditions with low CO_2:O_2 ratios. By limiting carbon dioxide uptake to the nighttime, CAM plants limit the loss of water due to photorespiration during the day, greatly increasing their water use efficiency. The carbon dioxide obtained at night is converted to oxaloacetate by phosphoenolpyruvate carboxylase (PEPC) and then malate by malate dehydrogenase (MDH) to be stored in the vacuole. During the day, the accumulated malate is released and converted back into carbon dioxide either by phosphoenolpyruvate carboxykinase (PEPCK) and MDH or by malic enzyme (ME), depending on the species. Therefore, the CAM cycle is temporal by nature and relies heavily on the circadian clock.

Several flowering plant taxa have been used as model systems to study CAM photosynthesis, including *Ananas comosus* (pineapple) (Ming et al., 2015) and *Agave* (Gross et al., 2013) and *Phalaenopsis equestris* in the Orchidaceae (Cai et al. 2015). The orchid family boasts a diverse range of species that include C_3, CAM, and intermediate photosynthetic phenotypes, even between closely related species (Silvera et al., 2010). In fact, there is evidence of multiple evolutionary origins of CAM photosynthesis and even reversals of CAM in the Orchidaceae (Silvera et al., 2009). *Erycina pusilla*, an epiphytic orchid that undergoes CAM, is a

strong candidate for a model orchid to study CAM photosynthesis due to its relatively short generation time and smaller genome (Lee et al., 2015). In addition, *E. pusilla* is a close relative of a C_3 species, *E. crista-galli*, which can be used in conjunction for a comparative study.

TEMPORAL GENE EXPRESSION ANALYSIS

Transcriptomic analyses can be an efficient way to study genes and biological processes that exhibit circadian rhythms. Plants that undergo CAM follow a day/night cycle, accumulating carbon dioxide and storing the carbon during the night to be used during the day. Titratable acidity measurements in the leaves of CAM plants were shown to be elevated when sampled during the night as opposed to the daytime (Silvera, Santiago, & Winter, 2005). Therefore, it is expected that genes involved in CAM photosynthesis follow temporal patterns that accompany the phenotype; therefore, one would expect to see temporal variation in expression levels of gene transcripts.

Few RNA-Seq studies in transcriptomics have analyzed CAM in a time series context. Traditional differential expression analyses are insufficient to capture the temporal component of time series data because traditional methods assume all samples are independent of one another. Therefore, temporal data necessitates an alternative method to compare expression between genes. The analysis is further complicated by the inability to sample the same *Erycina* plant over time, due to the small tissue availability of the plant. A variety of methods have been proposed to address this challenge, including Euclidean distance, model-based, and pattern recognition, or feature-based, algorithms (Androulakis, Yang, & Almon, 2007).

ASSEMBLY METHODS IN TRANSCRIPTOMICS

Since its introduction, high-throughput sequencing of cDNA (RNA-Seq) has been increasingly utilized in transcriptomics over traditional microarray technologies. RNA-Seq is a next-generation sequencing (NGS) method used to quantify transcript abundance in a transcriptome. There are many advantages of RNA-Seq including greater sensitivity, does not require any *a priori* assumptions about the gene regions, and increased resolution for transcriptional features, like alternative splicing. But, one of the biggest strengths of RNA-Seq technology is *de novo* transcript discovery. However, because of the relatively recent emergence of the technology, there is little consensus in best practices and methods for RNA-Seq analysis from platform and experimental design to assembly.

Transcriptome assembly is an essential, and complex step in the RNA-Seq pipeline. Optimizing the assembly for accuracy is of crucial importance in an RNA-Seq study, yet there are few established methods and tools to do so. Several studies have presented best strategies to sequence and evaluate *de novo* assemblies of RNA-Seq data. Most conclude that transcriptome assemblies are best evaluated by a combination of quality statistics (Soneson & Delorenzi, 2013; SEQC Consortium, 2014; Honaas et al., 2016) that include precision, recovery, and accuracy metrics. Although an increasing number of strategies have been developed to assess assembly quality, few studies have compared the results of these strategies across varying assemblies *within* a study. Frequently, RNA-Seq studies benchmark the quality of an assembly compared to other transcriptomic studies or a recommended score in literature. Some efforts to establish a more rigorous framework to evaluate transcriptome assemblies have been made. Recently, a multiple k-mer method in which different k-mer lengths were used to construct a *de novo* transcriptome assembly was shown to improve assembly quality (Surget-Groba & Montoya-

Burgos, 2010). The NOISeq package attempts to optimize a low-count filtering threshold by applying it to every experimental condition in the data set, incorporating the experimental design in the process (Tarazona et al., 2015).

SUMMARY

To perform an RNA-Seq analysis on *E. pusilla* and *E. crista-galli* to investigate CAM, I optimized and developed a transcriptome assembly for both species in my first chapter. I utilized various assessment strategies to evaluate assembly quality across different filtering thresholds and raw read counts. After producing a suitable assembly, I performed a comparative gene expression analysis on time series data for *E. pusilla* and *E. crista-galli* using a feature-based approach to identify temporal expression patterns associated with CAM behavior and candidate CAM regulators. Jointly, both chapters seek to establish a best methodology to perform an RNA-Seq analysis on a time series data set to explore CAM photosynthesis and study the evolution of CAM in the Orchidaceae.

REFERENCES

Androulakis, I., Yang, E., & Almon, R. (2007). Analysis of Time-Series Gene Expression Data: Methods, Challenges, and Opportunities. *Annu. Rev. Biomed*. Eng., 1–23. https://doi.org/10.1146/annurev.bioeng.9.060906.151904.Analysis

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W. C., Liu, K. W., ... & Zheng, Z. (2015). The genome sequence of the orchid *Phalaenopsis equestris*. *Nature Genetics*, 47(1), 65-72.

Consortium, S.-I. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control consortium. *Nature Biotechnology*, 32(9), 903–914. https://doi.org/10.1002/aur.1474.Replication

Gross, S. M., Martin, J. A., Simpson, J., Abraham-Juarez, M., Wang, Z., Visel, A., … Kuiper, M. (2013). *De novo* transcriptome assembly of drought tolerant CAM plants, Agave deserti and Agave tequilana. *BMC Genomics*, 14(1), 563. https://doi.org/10.1186/1471-2164-14-563

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., … DePamphilis, C. W. (2016). Selecting superior *de novo* transcriptome assemblies: Lessons learned by leveraging the best plant genome. *PLoS ONE*, 11(1), 1–42. https://doi.org/10.1371/journal.pone.0146062

Lee, S. H., Li, C. W., Liau, C. H., Chang, P. Y., Liao, L. J., Lin, C. S., & Chan, M. T. (2015). Establishment of an *Agrobacterium*-mediated genetic transformation procedure for the experimental model orchid *Erycina pusilla*. *Plant Cell*, Tissue and Organ Culture, 120(1), 211–220. https://doi.org/10.1007/s11240-014-0596-z

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., … Yu, Q. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nature Genetics*, 47(12), 1435–1442. https://doi.org/10.1038/ng.3435

Silvera, K., Neubig, K. M., Whitten, W. M., Williams, N. H., Winter, K., & Cushman, J. C. (2010). Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology*, 37(11), 995–1010. https://doi.org/10.1071/FP10084

Silvera, K., Santiago, L. S., Cushman, J. C., & Winter, K. (2009). Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. *Plant Physiology*, 149(4), 1838–47. https://doi.org/10.1104/pp.108.132555

Silvera, K., Santiago, L. S., & Winter, K. (2005). Distribution of crassulacean acid metabolism in orchids of Panama: Evidence of selection for weak and strong modes. *Functional Plant Biology*, 32(5), 397–407. https://doi.org/10.1071/FP04179

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14(1), 91. https://doi.org/1471-2105-14-91 [pii]\r10.1186/1471-2105-14-91

Surget-Groba, Y., & Montoya-Burgos, J. I. (2010). Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome Research*, 20(10), 1432–1440. https://doi.org/10.1101/gr.103846.109

Tarazona, S., Furio-Tari, P., Turra, D., Di Pietro, A., Nueda, M. J., Ferrer, A., & Conesa, A. (2015). Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21). https://doi.org/10.1093/nar/gkv711

CHAPTER II

OPTIMIZING A *DE NOVO* TRANSCRIPTOME ASSEMBLY FOR *ERYCINA PUSILLA* AND *ERYCINA CRISTA-GALLI*[1]

[1] Hwang, M., Albert, V., Silvera, K., Leebens-Mack, J. To be submitted to *BMC Plant Biology*.

ABSTRACT

Accurate assembly of RNA-Seq reads is essential for robust downstream assessment of gene expression, but *de novo* transcriptome assembly is a complex process (Martin & Wang, 2011). However, typical assembly assessment processes in RNA-Seq studies fail to compare quality across competing assembly strategies to evaluate the most appropriate assembly procedure. To obtain the best possible assembly for a subsequent temporal gene expression analysis to study CAM photosynthesis in orchid species *Erycina pusilla* and *E. crista-galli*, I generated multiple assemblies for both species varying depth of coverage filtering parameters. To assess quality, I used read mapping rate, N_x metrics, and percentage of conserved protein representation statistics. Methods included genome alignment using BLAT, assembly evaluation with TransRate, protein conservation using BUSCO, and read mapping alignment with Bowtie2 in the Trinity pipeline. Additionally, canonical CAM and clock genes were used as benchmark reference genes for assessment of transcript completeness. I found that using larger RNA-Seq read data sets to produce an assembly is does not necessarily improve quality. Furthermore, assessment with a single statistic is insufficient for capturing the entirety of the assembly quality and cannot be compared across different transcriptome studies. After careful evaluation of various filtering procedures assisted by CAM reference genes, the optimal assembly for our purposes utilized approximately 500k reads and was filtered by an fragments per kilboase of transcript per million mapped reads (FPKM) threshold of 1. Therefore, the final filtered assembly was tuned to performance in expression analyses for a CAM photosynthesis study.

INTRODUCTION

The emergence of massively parallel cDNA synthesis and sequencing (RNA-Seq) greatly accelerated the study of gene expression by providing an economical and efficient approach for quantifying transcripts. While an RNA-Seq study does not necessitate any *a priori* knowledge of a transcriptome, microarray-based approaches are biased by probes that only return gene expression results in regions from which they were designed. In addition, RNA-Seq offers the possibility of performing discovery-based experiments without the need for a reference genome, making it an attractive technology for non-model species. With RNA-Seq, the ability to determine transcript levels of tens of thousands of genes at once has allowed experiments to expand to a much larger scale but also necessitates new techniques to address the many computational challenges that come with processing the data (Vijay et al., 2013).

In the absence of a reference genome, a foundational stage in an RNA-Seq study involves the complex process of compiling sequencing reads into a *de novo* transcriptome assembly. Common RNA-Seq assembly techniques utilize the alignment of cDNA sequencing reads to a reference genome. While a mapping-based strategy for assembly can be advantageous in its precision, many organisms do not have high quality reference genomes available. Without the assistance of a reference, *de novo* assembly is a more difficult approach, that must be coupled with careful assessment to ensure a high quality assembly (Wang, Gerstein, & Snyder, 2009; Vijay et al., 2013). Because reads are mapped back to the assembly during transcript quantification to estimate gene expression levels, the accuracy of downstream analyses is fundamentally dependent on the quality of the *de novo* transcriptome assembly. Thus, assembly evaluation is a crucial and iterative step in the *de novo* assembly process.

Several criteria have been developed to evaluate transcriptome assemblies, including read mapping rate, N_x statistics (the length of the shortest contig in the set of longest contigs encompassing N% of the assembly's bases), recovery of conserved protein sequences, and the number of unique genes – all of which are valuable on their own but should be considered as a whole when assessing the quality of an assembly (Honaas et al., 2016). To improve assembly quality, these types of assembly summary statistics are used to diagnose problems in an assembly so that they may be resolved. Problems can arise during assembly because assembly software, while powerful, can assemble low quality or partial transcripts and even nonsense transcripts such as chimeric sequences that must be filtered out. Transcripts that are artifacts of assembly often have low read support and can be filtered based on normalized read support. Two common statistics used to remove likely erroneous sequences include fragments per kilobase of transcript per million mapped reads (FPKM) for paired-end reads, reads per kilobase of transcript per million mapped reads (RPKM) for single-end reads, and assembled transcript length. RPKM/FPKM is typically used as a normalization procedure for transcript count based on transcript length, allowing the more accurate comparison for transcripts with different length within a sample (Li et al., 2009). This statistic can also be used to filter out lowly expressed transcripts that could be artifacts of assembly, while transcript length is helpful in removing short reads or partial sequences. Filters that remove lowly expressed transcripts greatly improve assembly quality (SEQC Consortium, 2014), therefore choice of filtering threshold is an important assembly step. Filtering greatly affects downstream analyses using the assembly because reads are aligned to the assembly to estimate expression. If an assembly is not filtered properly, assembly artifacts may appear to be expressed genes, increasing noise in the data.

Therefore, filtering should be considered in conjunction with broader assembly-wide assessment criteria to improve overall assembly quality.

To study the evolution and regulation of Crassulacean acid metabolism (CAM) photosynthesis in the Orchidaceae, I am generating a transcriptome assembly for CAM species *Erycina pusilla* and its close C_3 relative *E. crista-galli* for a downstream temporal gene expression analysis in which genes will be clustered based on similar expression over a 24-hour period to identify master regulators in CAM. I am using a *de novo* assembly approach in the absence of a quality *E. pusilla* or *E. crista-galli* reference genome. Because we can expect canonical CAM genes to exhibit temporal variation in expression, we will use them as benchmark reference genes in order to assess and compare generated assemblies.

Every species manifests a unique biology that requires a slightly different approach in RNA-Seq experiments for assessing specific research questions. For example, one must consider that RNA-Seq only takes a snapshot of gene expression at a specific time point; however, many biological processes are correlated with time, like CAM. As a result, there is no one universal tool or set of rules for optimal transcriptome assembly. Optimizing an assembly necessitates fine tuning of software parameters, filtering procedures, and comparative testing to determine the best option. Here I present a methodology to determine an optimal *de novo* transcriptome assembly for a temporal gene expression analysis of *E. pusilla* and *E. crista-galli* through individual assessment and comparison of various assembly strategies described in the literature.

METHODS

Experimental Design and Sequencing

E. pusilla plants were grown in 12-hour day and 12-hour night conditions over three independent dates in a growth chamber at University of Buffalo. RNA was isolated from leaf tissue using the RNeasy Plant Mini Kit (Qiagen). Due to limited biomass availability, only one sample was taken from a plant; therefore, all tissue was collected from distinct, randomly selected orchids. The first two data sets include plants sampled every 4 hours for a total of 27 and 6 samples, and the third data set includes plants taken every 2 hours for a total of 36 samples. For *E. crista-galli,* plants were grown on an open-sided shadehouse, located in a mountain tropical humid rainforest in Central Panama, Republic of Panama. Plants were watered daily and fertilizer was applied once every 15 days using commercial 20-20-20 and 16-32-16 (N-P-K) fertilizer solutions. Daily temperatures ranged from 19C to 32C with an average of 24C, 2500 mm/year rainfall. Leaf tissue samples were subsequently quantified via nanodrop and checked for integrity with a Bioanalyzer v2100. RNA libraries were constructed using the Kapa mRNA stranded kit with a combinatorial barcoding scheme. Libraries were sequenced on an Illumina NextSeq500 with PE75 reads, pooling 30-32 samples per run. A summary of the data can be found in Table 1.1A.

De novo Transcriptome Assembly

Since there are three data sets for *E. pusilla*, we wanted to determine whether the assembly quality improved as more data were added to the assembly process. Transcriptomes were assembled and quality trimmed with Trinity v2.0.6 using the recommended *de novo* protocol with its *in-silico* normalization algorithm to a maximum coverage of 50x (Haas et al., 2013). Five transcriptome assemblies were named as follows: assembly A (*E. pusilla* sets 1 and

2), assembly B (*E. pusilla* sets 1, 2, and 3), assembly C (*E. pusilla* set 3), assembly D (*E. pusilla* set 3), and assembly E (*E. crista-galli* set 4), containing 540M, 540M, 1.1B, 566M, and 391M total reads before normalization, respectively. Trinity assembly divides output into transcripts and genes in which transcripts, or isoforms, are clustered into groups based on shared sequence content, loosely defined as 'components' or genes. To annotate the transcripts, we ran TransDecoder v3.0.1 (Haas & Papanicolaou et al.), retaining transcripts with open reading frames validated with NCBI BLAST+ (Camacho et al., 2009) and the hmmsearch tool in HMMER v3.1 (hmmer.org). Canonical CAM and clock genes were identified by best BLAST search against the UniProtKB/Swiss-Prot database (E-value < 1E-3, percent identity > 80%). Canonical CAM genes included phosphoenolpyruvate carboxylase (PEPC), phosphoenolpyruvate carboxylase kinase (PPCK), malate dehydrogenase (MDH), NADP-malic enzyme (NADP-me), NAD-malic enzyme (NAD-me), pyruvate phosphate dikinase (PPDK), phosphoenolpyruvate carboxykinase (PEPCK), and phosphoenolpyruvate carboxylase phosphatase (PEPC-phosphatase). *A complete list of canonical genes and citations can be found in Table S1.1.*

Transcript Quantification and Filtering

Abundance estimates of the assembly transcripts was calculated using RSEM v1.3.0 (Li & Dewey, 2011) from the Trinity pipeline. To filter out transcripts that are likely to be assembly artifacts, assemblies were filtered by FPKM count. If a transcript did not have an FPKM score at a certain threshold, it was removed from the assembly. FPKM filtering thresholds were applied at 0.25, 0.5, 1, 1.5, 2, 3, 4, and 5. To determine a reasonable FPKM threshold to filter by, we referenced the FPKM count of the canonical CAM genes. Transcript length was also used as a filtering parameter at 300, 400, 500, 600, 700, and 800 bases. (Trinity previously filtered out any

transcripts <200bp when building the assembly.) A histogram of transcript lengths <3000bp was generated to observe the distribution of sequence lengths in the assembly across varying FPKM thresholds (Fig. 2.2).

Assessment and Comparison of Assemblies

I used several methods to assess the quality of the transcriptome assembly: (1) TransRate (Smith-Unna, Boursnell, Patro, Habbard, & Kelly, 2016), (2) genome alignment with BLAT (Kent, 2002), and (3) BUSCO v2 (Simao, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015). The first software, TransRate, assesses *de novo* transcriptome assemblies and reports numerous quality scores for contigs and assemblies. A genome assembly exists for *E. pusilla* (Ming-Tsair Chan et al. in prep, Academia Sinica – Biotechnology Center in Southern Taiwan, Tainan, Taiwan), but it is highly fragmented. Despite having a low-quality genome, we can still gain information by aligning assemblies to the reference. Studies have shown that hybrid assemblies composed of a reference-based assembly and a *de novo* assembly yield the best results (Marchant et al., 2014), so mapping performance can be valuable in assessment. The assemblies were aligned to the genome using BLAT, an efficient BLAST-like alignment tool that specializes in genomes. I used default parameters for DNA-RNA searches, while including the "-fine" parameter, which searches more deeply for small initial and terminal exons. BUSCO v2, developed by the Zdobnov Computational Evolutionary Genomics group, is an open-source software that calculates transcriptome completeness based on a database of highly conserved single-copy orthologs selected from OrthoDB (Simao et al., 2015). Each quality assessment software pipeline was run on all five raw and filtered assemblies for comparison. *For a full list of software parameters, see Table S1.2.*

RESULTS

Assembling the Transcriptome

Species and data sets used to build each assembly as well as transcript and isoform counts are summarized in Table 2.1B. *E. pusilla* assemblies A and B show little difference in the number of transcripts and Trinity-defined genes. The addition of a third data set in *E. pusilla* assembly C doubled the number of reads that were used as input, but only gave ~27% increase in transcripts in the assembly. For *E. pusilla* data sets, set 3 alone has approximately the same amount of reads as Set 1 and 2 combined, but results in an assembly that has ~9% less transcripts when assembled alone.

Results of the initial assessment of the unfiltered assemblies are described in Table 2.2. All five assemblies have a mapping percentage of at least 80%, indicating that there is good RNA-Seq read representation in the assembly A (Conesa et al., 2016; Haas et al., 2013). The percentage of conserved proteins from BUSCO that were present in the assembled transcripts averaged ~76% with very little variation. Similarly, there was little difference in the mean percentage of transcripts covered by an open reading frame, although *E. crista-galli* performed the highest by six points at 49.26%. All *E. pusilla* assemblies scored ~77% when mapped to the *E. pusilla* genome with little variance. *E. crista-galli* only scored a 55% mapping rate, which is expected since the genome is of a different species. On average, all *E. pusilla* assemblies resolved 16 canonical CAM gene copies and 35 clock gene copies, while the assembly generated from *E. crista-galli* data recovered double the number of canonical CAM gene copies.

Transcriptome Filtering

Filtering transcripts is an important step in generating a quality transcriptome assembly. By removing transcripts that are unlikely to be real due to misassembly or contamination, we can

reduce the number of transcripts for more efficient computations and easier downstream analyses. Previous studies in transcriptomics have set the FPKM threshold at 0.3 (L. Wang et al., 2013), 0.5 (Grabherr et al., 2011), 1 (Graveley, Brooks, & Carlson, 2011), 2 (Zeng et al., 2014), and 5 (Secco et al., 2014). To investigate how FPKM filtering affects the assembly quality, we chose to test a range of threshold values from 0.25 to 5 on the assemblies for *E. pusilla* and *E. crista-galli*, respectively. At a threshold of 0.25, the gene count drops drastically to ~50% of the unfiltered assembly for both species (Fig. 2.1A). Gene count continues to drop at a rapid rate until beginning to taper at a FPKM threshold of 1. Many of the sequenced transcripts in the unfiltered assembly were very lowly expressed. It is possible that some of the filtered genes were truly lowly expressed genes; however, these genes are not essential to our downstream analysis, since they are lowly expressed throughout the day.

Fig. 2.1B shows the subsequent loss of reference CAM and clock genes at different FPKM thresholds for both assemblies. *E. crista-galli* C_3 assembly E is robust against losing canonical CAM genes as the FPKM threshold increases and retains all of its reference genes at a threshold level as high as FPKM=5. On the other hand, the *E. pusilla* C_3 assembly B begins to lose genes annotated as canonical CAM genes when filtered by FPKM=1.5. Both assemblies retain canonical clock genes up to a threshold level of FPKM=1.5.

A right skewed histogram of the transcript lengths <3000bp in both the CAM and C_3 assemblies shows that there is a high number of short length transcripts primarily in the 300-400bp range (Fig. 2.2A). At increasing FPKM threshold levels, the number of these short length transcripts decreases dramatically in proportion to longer transcripts (Fig. 2.2B,C).

Assembly Assessment

While N_x statistics improve with higher transcript length thresholds as expected, a different pattern occurs with increasing FPKM threshold. Each N_x value tested peaks at a certain FPKM threshold ranging from FPKM=0.5 to FPKM=3 (Fig. 2.3A). Figure 2.3B and Figure 2.3C visualize the change in percentages of these statistics at increasing threshold levels of FPKM and transcript length, respectively. For both assemblies, increasing threshold reduces the percentage of BUSCO protein representation, especially at higher levels of FPKM threshold. On the other hand, the genome mapping percentage is increased at low FPKM thresholds but drops back to the percentage from the unfiltered assembly at approximately FPKM=1. At increasing thresholds of transcript length, the percentage of BUSCO representation changes very marginally, staying about the same until a threshold of 600bp. Mean percentage ORF remains stable across changes in transcript length threshold. The mean percentage open reading frame changes very little post-filtering by length and increases sharply only up to a threshold of FPKM=1 (Fig. 2.3B).

DISCUSSION

Determining the Optimal Raw Assembly

Assembly B, derived from 540,621,053 reads and 33 libraries, was chosen as the optimal assembly for *E. pusilla* to move forward with testing filtering parameters. Assembly B outperformed assembly A in read mapping percentage, percentage of conserved protein representation, and annotation of canonical CAM and clock gene copies despite having identical input reads. The addition of the third data set did not improve the assembly scores. Assembly D, which contained set 3 alone, resolved many fewer genes than assembly B, but did not suffer a loss of conserved proteins from BUSCO. However, assembly D has a lower percentage of

transcripts that map to the *E. pusilla* genome than assembly B. For subsequent discussion, assembly B for *E. pusilla* will be referred to as the CAM assembly, and assembly E for *E. crist-galli* will be referred to as the C_3 assembly.

Setting an FPKM Threshold

Setting a greater FPKM threshold value will remove assembly artifacts to improve the quality of the assembly. In addition, the genes of interest for our study should have reasonably high FPKM values in *E. pusilla* similar to those of our canonical CAM genes, since we are looking for regulators in the CAM photosynthesis pathway. The utilization of a reference gene set to benchmark our assessment methods in itself may introduce bias, as only those genes with expression similar to that of our reference genes are considered. The higher the FPKM threshold, the more likely we might be removing our reference canonical CAM genes, increasing the likelihood of removing legitimately expressed genes that may be important from the assembly. Therefore, opting to stay more on the conservative side is advisable in this case.

The large amount of short length transcripts is indicative of partially-sequenced transcripts or short RNA reads that are not interesting for our analysis. Canonical CAM and clock genes are almost perfectly retained up to a transcript length threshold of 800bp in both assemblies, demonstrating that our genes of interest are robust against transcript length filters. It is important to note that while a short transcript may be removed from the assembly, an alternative isoform of the same gene may still exist in the assembly.

The TransRate assembly score is a summary statistic calculated as the geometric mean of all four contig scores including: a measure of whether a base is called correctly, a measure of whether a base is truly part of a transcript, the probability a contig is derived from a single transcript, and the probability a contig is structurally complete and correct. As a result, any one

poor measurement can drastically affect the TransRate contig score. By combining measurements and calculating the product across contigs, it is difficult to determine why an assembly might have a low or high score and how much each contig or measurement is influencing the score. Thus, the overall assembly score is vague and uninformative.

Assessing Assembly Quality

The three methods used to assess assembly quality were executed across varying threshold levels of FPKM and transcript length in both assemblies. The initial software used to assess assembly quality was TransRate, in which we focused on N_x statistics, contig quality, and mean ORF %. Assembly and contig scores in TransRate are calculated by the product of four different statistics. Although a certain proportion of transcripts were determined to have a poor contig score by TransRate, we retained these transcripts because a single component of the score could drastically lower the result, penalizing a transcript that could still be useful in downstream analyses. The improvement of an N_x statistic from increasing the FPKM threshold indicates a correlation between short length transcripts and very lowly expressed transcripts that can be simultaneously removed. However, increasing the threshold above this value may result in filtering out full length transcripts that happen to be lowly expressed. Choosing a threshold within this range is ideal because it is at this threshold where the balance between filtering out short length transcripts while saving legitimate mostly full length sequences lies. Removal of extremely lowly expressed transcripts may improve the mean percentage ORF by the removal of assembly artifacts with nonsense open reading frames.

Each of the remaining assessment results can be summarized by a percentage statistic: 1) percentage of transcripts that mapped to the *E. pusilla* genome using BLAT, 2) percentage of transcripts that aligned to the original RNA-Seq reads with Bowtie2, and 3) percentage of

conserved proteins represented by the assembly in BUSCO. First, a higher mapping percentage to the *E. pusilla* genome implies a more accurate assembly, assuming the reference genome is representative of all genes in the species. Next, a quality assembly should have a high percentage of the original RNA-Seq reads aligning to the assembly, indicating that the assembly is representative of the reads used to build it. Finally, BUSCO's analysis is founded on the idea that a robust assembly should contain those proteins that are highly conserved across lineages. Therefore, subsequent filtering of the assembly by FPKM or transcript length should mostly retain these reference proteins.

Overall, it is expected that read mapping percentage and percentage of BUSCO representation decreases as the FPKM or length filter increases. The statistics tend to decrease more rapidly past a threshold of over FPKM=1.5 to FPKM=2. The percentage of transcripts that mapped to the reference genome, however, has a trend that increases slightly at low FPKM thresholds and then begins to decrease. For length, the percentage mapped to the reference continues to increase, which is either the result of removal of assembly artifacts or genes that are very lowly expressed. (BLAT will find matches that are of 80% or greater similarity to the query.)

Many studies utilize high BUCSO protein representation as a statistic indicating a quality transcriptome assembly. However, while a complete protein representation may be a positive assessment of a transcriptome assembly, it does not confirm whether transcripts of interest are present. The only assumption that can be made is that the assembly is comprehensive enough to include a highly conserved protein set.

Assessing Strategies for Assembly Evaluation

While many studies apply FPKM filtering thresholds for transcriptome assemblies, few evaluate the effect different FPKM thresholds have on the quality of the transcriptome and instead refer to previous literature on an appropriate threshold value. However, a universal FPKM threshold is impossible, since we cannot compare across other transcriptome assemblies due to enormous variability between species, studies, and even sequencing runs. Thus, we have determined the best filtering approaches for our own purposes of studying the CAM photosynthesis pathway.

CONCLUSION

Assembly quality assessment is best performed via the combination of multiple methods to ensure a more comprehensive perspective. In my case, I utilized conserved protein enrichment, genome and RNA-Seq read alignment, and summary statistics while using canonical CAM genes as reference. Moving forward, we will be utilizing assemblies B and E for *E. pusilla* and *E. crista-galli*, filtering by an FPKM threshold of 1 and a transcript length of 600bp, for our downstream temporal gene expression analysis. Both assemblies will undergo identical filtering parameters to ensure consistency when comparing between species. The decision to use these filtering parameters was determined by an evaluation of assembly quality under varying levels of filtering, which allows for consistent comparison as opposed to comparing across other RNA-Seq studies.

REFERENCES

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST plus: architecture andti applications. *BMC Bioinformatics*, 10(421), 1. https://doi.org/Artn 421\nDoi 10.1186/1471-2105-10-421

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., … Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol*, 17(1), 13. https://doi.org/10.1186/s13059-016-0881-8

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, 29(7), 644–52. https://doi.org/10.1038/nbt.1883

Graveley, B., Brooks, A., & Carlson, J. (2011). The developmental transcriptome of *Drosophila melanogaster*. *Nature*, 471(7339), 473–479. https://doi.org/10.1038/nature09715

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Philip, D., Bowden, J., … Pochet, N. (2013). *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protocols* (Vol. 8). https://doi.org/10.1038/nprot.2013.084

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., … DePamphilis, C. W. (2016). Selecting superior *de novo* transcriptome assemblies: Lessons learned by leveraging the best plant genome. *PLoS ONE*, 11(1), 1–42. https://doi.org/10.1371/journal.pone.0146062

Kent, W. J. (2002). BLAT — The BLAST -Like Alignment Tool. *Genome Research*, 12, 656–664. https://doi.org/10.1101/gr.229202.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323. https://doi.org/10.1186/1471-2105-12-323

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., & Dewey, C. N. (2009). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4), 493–500. https://doi.org/10.1093/bioinformatics/btp692

Marchant, A., Mougel, F., Mendonça, V., Quartier, M., Jacquin-Joly, E., da Rosa, J. A., … Harry, M. (2014). Comparing *de novo* and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug *Rhodnius prolixus*. *Insect Biochemistry and Molecular Biology*, 69, 25–33. https://doi.org/10.1016/j.ibmb.2015.05.009

Martin, J. A., & Wang, Z. (2011). Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12(10), 671–682. https://doi.org/10.1038/nrg3068

Secco, D., Shou, H., Whelan, J., & Berkowitz, O. (2014). RNA-seq analysis identifies an intricate regulatory network controlling cluster root development in white lupin. *BMC Genomics*, 15(1), 230. https://doi.org/10.1186/1471-2164-15-230

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Smith-Unna, R., Boursnell, C., Patro, R., Habbard, J., & Kelly, S. (2016). TransRate: reference free quality assessment of *de novo* transcriptome assemblies. *Genome Research*, 26(8), 1134–1144. https://doi.org/10.1101/gr.196469.115

Vijay, N., Poelstra, J. W., Künstner, A., & Wolf, J. B. W. (2013). Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. *Molecular Ecology*, 22(3), 620–634. https://doi.org/10.1111/mec.12014

Wang, L., Zhao, S., Gu, C., Zhou, Y., Zhou, H., Ma, J., … Han, Y. (2013). Deep RNA-Seq uncovers the peach transcriptome landscape. *Plant Molecular Biology*, 83(4–5), 365–377. https://doi.org/10.1007/s11103-013-0093-5

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 57–63.

Zeng, J., He, X., Wu, D., Zhu, B., Cai, S., Nadira, U. A., … Zhang, G. (2014). Comparative transcriptome profiling of two Tibetan wild barley genotypes in responses to low potassium. *PLoS ONE*, 9(6). https://doi.org/10.1371/journal.pone.0100567

Table 2.1 – Summary tables of A) sampled data and B) raw assemblies.

A

Set	Species	Phenotype	Reads	Samples	Sample Interval	Average Replicates
1	*E. pusilla*	CAM	346,366,044	22	2 hours	2
2	*E. pusilla*	CAM	220,467,196	11	2 hours	1
3	*E. pusilla*	CAM	566,833,240	24	2 hours	2
4	*E. crista-galli*	C_3	391,848,695	36	4 hours	3

B

Assembly	Species	Data	Reads	Transcripts	Trinity "Genes"	%GC
A	*E. pusilla*	Sets 1 & 2	540,621,053	310,313	198,358	36.18
B	*E. pusilla*	Sets 1 & 2	540,621,053	310,763	198,476	36.20
C	*E. pusilla*	Sets 1, 2, & 3	1,107,454,293	395,726	243,281	35.84
D	*E. pusilla*	Set 3	566,833,240	282,651	179,146	36.61
E	*E. crista-galli*	Set 4	391,848,695	189,386	136,155	37.70

Table 2.2 – Assessment statistics for each assembly.
[1] Mean percent of transcripts covered by an open reading frame
[2] Genes were annotated with BLASTP E-value<1E-5. Total number of unique CAM and clock genes searched were 7 and 16, respectively.

Assembly	**% Mapped RNA-Seq Reads**	**% Mapped Proteins w/ BUSCO**	**% ORF[1]**	**% Mapped to E. *pusilla* Genome w/ BLAT**	**Unique CAM gene[2]**	**Unique Clock genes[2]**	**N50**	**TransRate Assembly Score**
A	82.93%	76.80	40.40	77.12	7	16	1509	0.18921
B	83.04%	77.04	40.44	77.78	7	16	1512	0.18812
C	84.15%	74.94	38.33	78.79	7	16	1388	0.17364
D	85.73%	78.68	43.28	76.37	7	16	1337	0.13769
E	87.63%	74.51	49.26	55.0	7	16	1293	0.28532

A **B**

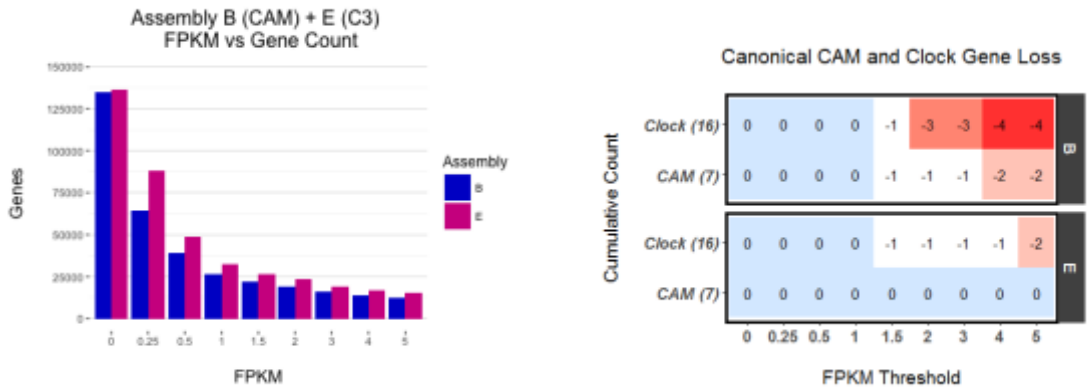
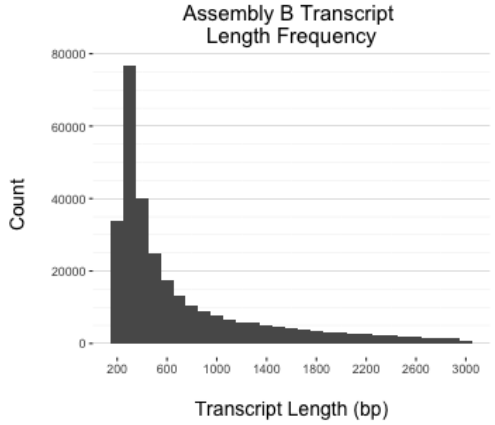
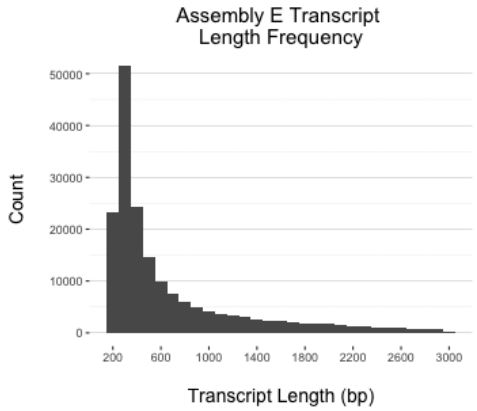


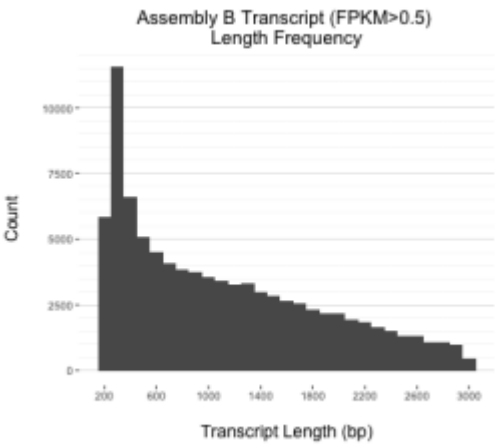
Figure 2.1 – Gene counts across varying FPKM thresholds for *E. pusilla* (CAM) assembly B and *E. crista-galli* (C_3) assembly E. A) Assembly gene count across varying FPKM threshold levels, B) Reference gene loss across varying FPKM threshold levels.

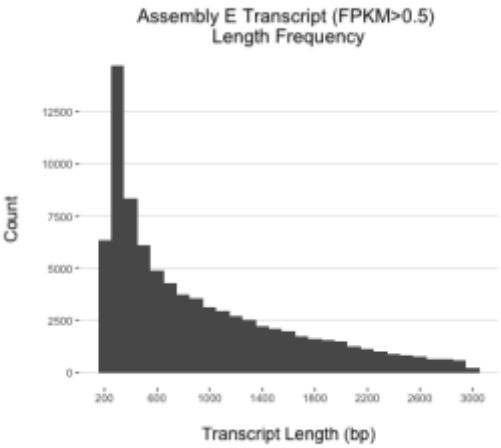
A



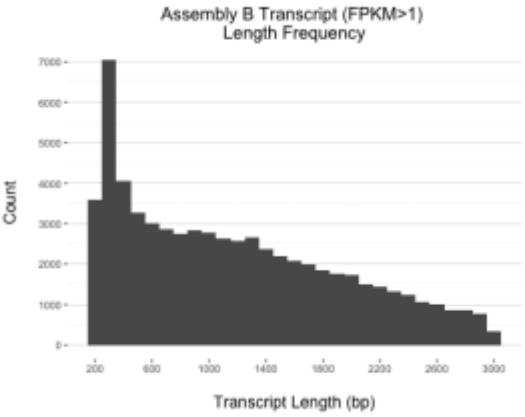


B





C



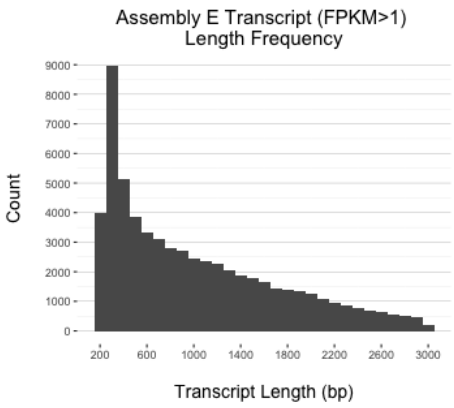
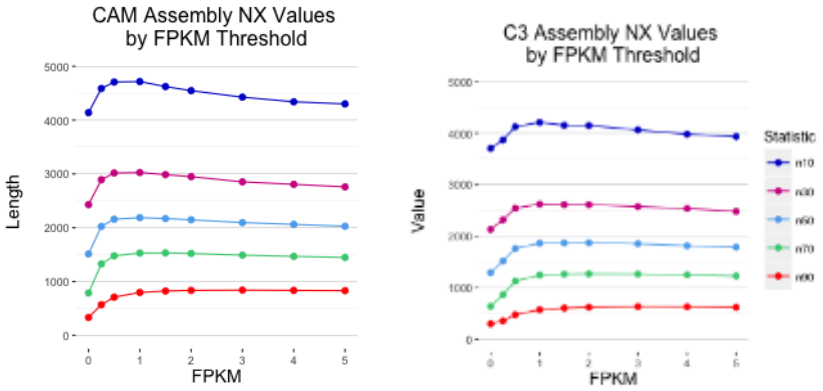
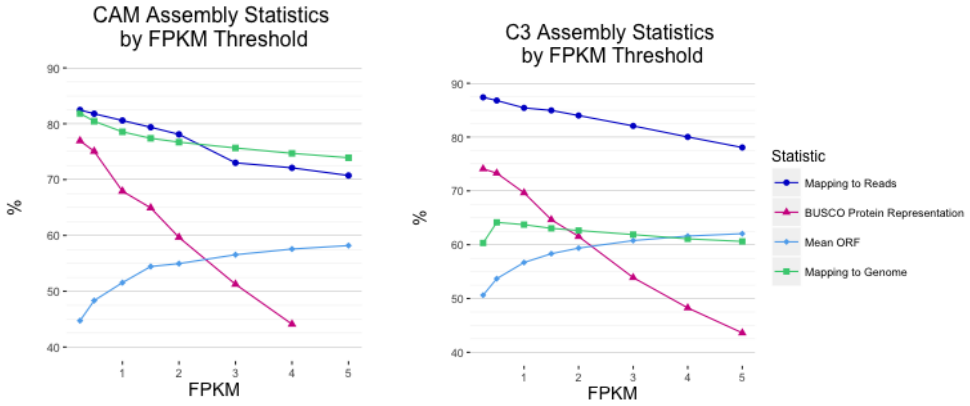


Figure 2.2 – Frequency distributions for transcript lengths <3000bp in *E. pusilla* and *E. crista-galli.* A) unfiltered assemblies, B) assemblies filtered by FPKM>0.5, and C) assemblies filtered by FPKM>1.

A



B



C

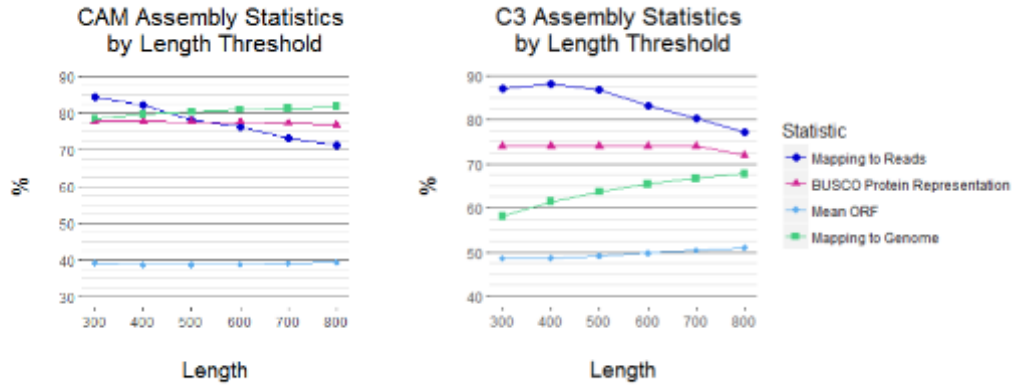


Figure 2.3 – Assembly quality assessment statistics. A) N_X values across FPKM thresholds, B) assessment statistics across varying FPKM thresholds, and C) assessment statistics across varying length thresholds. Assessment statistics include RNA-Seq read mapping percentage, percentage of BUSCO proteins represented in the assembly, mean percentage of transcripts covered by an open reading frame, and read mapping percentage to the *E. pusilla* genome.

CHAPTER III

A COMPARATIVE APPROACH TO TIME-DEPENDENT GENE EXPRESSION ANALYSIS OF CRASSULACEAN ACID METABOLISM (CAM) IN *ERYCINA*

[1] Hwang, M., Albert, V., Silvera, K., Heyduk, K., Leebens-Mack, J. To be submitted to *BMC Plant Biology*.

ABSTRACT

Crassulacean acid metabolism (CAM) photosynthesis, an alternative to C_3 photosynthesis, has been implicated in the evolution of flowering plants as an adaptation for survival in arid environments. While many of the major genes involved in the CAM pathway are known, for example phosphoenolpyruvate carboxylase (PEPC), little is known about their regulators or evolution in the Orchidaceae, a diverse family of flowering plants that contains a mix of both C_3 and CAM species (e.g. Silvera et al., 2005). *Erycina pusilla*, an orchid that uses CAM photosynthesis, is a relatively fast-growing species with the potential for functional genomic characterization of CAM genes. To develop a list of genes that may be contributing to CAM and in order to study CAM evolution in the Orchidaeae, we compared *E. pusilla* gene expression profiles to those of a close relative, *E. crista-galli*; such a comparison highlights changes in regulation that occurred during the evolution of C_3 to CAM in *Erycina*. We focus on putative CAM genes that follow a clock-like pattern in the CAM species but not in the C_3, and we predict that regulators of those genes will exhibit similar patterns of time-dependent expression. Genes that exhibit similar time-structured expression patterns were clustered and compared with the clusters exhibiting similar time-structured expression in *E. crista-galli*. A network analysis gene expression correlations with each cluster revealed candidate regulators of the CAM photosynthesis pathway, as well as CAM-like expression patterns in canonical CAM genes present in both species, implying that C_3 ancestors of CAM species may have been pre-disposed for CAM.

INTRODUCTION

Plants exhibiting Crassulacean acid metabolism (CAM) photosynthesis evolved from C_3 ancestors as an adaptation to environments with limited water availability or low CO_2:O_2 ratios. Compared to C_3 plants, CAM plants have an improved water-use efficiency (WUE), or more carbon dioxide acquired per unit of water lost. CAM species have as high as a 50% increase in WUE relative to C_3 and C_4 photosynthetic species (Graham & Nobel, 1996; Drennan & Nobel, 2000). The inefficiency of C_3 photosynthesis is largely a result of excess transpiration, in which an overwhelming majority of the water plants absorb through the roots is lost through stomata. Stomatal movements control both the rate of CO_2 assimilation and transpiration via changes in partial pressures of both water vapor and CO_2 inside the leaf (Farquhar & Sharkey, 1982). In C_3 plants, stomata do not close under normal daytime conditions; therefore, there is constant collection and immediate carbon fixation of CO_2 and consequential water loss, which is unsustainable under prolonged periods of drought. Unlike C_3 plants, CAM plants close their stomata during the daytime, reducing water loss by up to six-fold relative to C_3 plants (Drennan & Nobel, 2000). The intake of carbon dioxide for CAM plants is largely limited to nocturnal hours where it is converted to a four-carbon molecule in the form of malate in reactions catalyzed by first phosphoenolpyruvate carboxylase (PEPC) and then malate dehydrogenase (MDH). Malate is then stored in the vacuole (Fig 3.1A). The accumulated malate is transported from the vacuole during the day, decarboxylated by MDH and converted back to a four-carbon molecule by phosphoenolpyruvate and carbon dioxide by phosphoenolpyruvate carboxykinase (PEPCK) to be used in the Calvin Cycle. An alternative decarboxylation pathway is utilized by some plants in which malate is decarboxylated by NADP-malic enzyme (NADP-me) or NAD-malic enzyme (NAD-me) and converted back to a four-carbon molecule by pyruvate dikinase

(PPDK). By opening stomata at night when transpiration rates will generally be lower than during the day, CAM plants reduce water loss by up to six fold relative to C_3 plants (Drennan & Nobel, 2000).

Many previous phylogenetic analyses have analyzed the evolution of PEPC in CAM species (Lepiniec et al., 1994; Gehrig et al., 2005) because it is expected that the CAM-specific PEPC isoforms evolved from C_3 isoforms. Two major factors are indicative of a CAM-specific PEPC isoform in a plant: 1) a CAM PEPC that exhibits diurnal expression, or relatively higher activity during the nighttime relative to the daytime, and 2) there is elevated mRNA and protein expression levels in leaf tissues, where photosynthesis occurs (Borland et al., 1999). In addition, there is evidence of differential expression between CAM-specific PEPC isoforms and C_3 PEPC isoforms in CAM species (Gehrig, Heute, & Kluge, 2001; Gehrig et al., 2005) as well as a correlation between CAM phenotype and PEPC activity (Ceusters et al., 2014). Silvera et al. describes this relationship in a proposed model for the evolutionary progression of PEPC genes within orchids in which those species exhibiting a stronger CAM phenotype contain a CAM PEPC with a more abundant expression relative to its C_3 isogene (2010). Therefore, it is important to compare the expression profiles of PEPC and all CAM genes between CAM and related C_3 species so we can begin to identify differences in regulation to identify CAM regulators that are not temporally regulated in the C_3 species.

While many of the core players of the CAM pathway are known, and have been well-studied for decades, little has been established about the details of CAM gene regulation (DePaoli et al., 2014; Ming et al., 2015). Due to its day and night separation, it is evident that CAM follows a clock-like cycle, in which certain genes are expressed during the day as opposed to the night, and vice versa. Previous studies have reported a nocturnal pattern for PEPC and

PEPC kinase (Nimmo et al., 1987; Taybi et al., 2000) and a diurnal pattern for phosphoenolpyruvate carboxykinase (PEPCK) (Walker, Trevanion, & Leegood, 1995; Leegood & Walker, 1996) in CAM plants. Since we know that these major genes involved in CAM follow a temporal expression pattern, it is likely that the regulators of those genes are similarly expressed. We expect that genes involved in the CAM regulatory pathway should be controlled by a co-expressed master regulator (Borland et al., 1999; Nimmo, 2000; Taybi et al., 2000).

Many methods have been developed to analyze differentially expressed genes between samples from RNA-Seq, often adapted from strategies previously used for microarrays. DESeq (Anders & Huber, 2012), edgeR (Robinson, McCarthy, & Smyth, 2009), and limma/voom (Law et al., 2014) are popular R packages used for differential expression analysis, each with their own advantages and disadvantages. These algorithms calculate correlation statistics between pairs of genes. When traditional differential expression approaches are applied to transcriptomic analyses of time-course data, valuable temporal information is lost. Those analyses often rely on the assumption that samples are independent of each other, whereas samples in a time series study are ordered in time. Here we apply a time-dependent clustering approach (Conesa et al., 2006) for analyzing gene expression in *Erycina* to study the regulation and evolution of CAM photosynthesis in Orchidaceae, as well as identify candidates for previously unknown master regulators of the CAM pathway.

E. pusilla and its C_3 relative, *E. crista-galli*, offer valuable insight into studying the evolution and regulation of CAM photosynthesis in the Orchidaceae. Utilizing a time series based approach takes advantage of the diel expression of genes characteristic of CAM plants, offering an additional layer of information when analyzing RNA-Seq data.

METHODS

Studying CAM in the Orchidaceae

Due to the competitive advantage of WUE in the CAM system, CAM is widely distributed across angiosperms, having evolved independently in at least 35 families (Silvera et al., 2010). CAM is found in diverse lineages that experience insufficient access to water due to environmental conditions such as aridity and epiphytism, in which plants rely on another plant for structural support and direct rainfall as its primary source of water. Epiphytic species constitute more than 70% of the Orchidaceae family (Gravendeel et al., 2004; Chase et al., 2015), a quarter of which undergo CAM (Silvera et al., 2005). Orchids are well known for their taxonomic diversity, numbering over 25,000 species and 736 genera (Chase et al., 2015), with a mixed distribution of C_3 and CAM traits (Silvera et al., 2013). Because many genera constitute both C_3 and CAM species, the orchids are an attractive family to study the evolution of CAM photosynthesis.

Previous studies have found evidence for multiple independent origins of CAM across the orchid phylogeny (Silvera et al., 2010). The genus *Erycina* represents a clade of epiphytic plants composed of both CAM plants, such as our study species *Erycina pusilla*, and C_3 plants, like its close relative, *Erycina crista-galli*. Faster growing and containing a smaller genome than many of its orchid relatives, *E. pusilla* has the potential to be a model system for studying CAM photosynthesis in the Orchidaceae (Lee et al., 2015), while *E. crista-galli* provides a favorable subject for a comparative analysis with a C_3 species.

Titratable Acidity

To verify the photosynthetic pathway of *E. pusilla*, titratable acidity was measured on an additional set of plants. In plants using CAM photosynthesis, we expect leaf acidity to peak in

the early morning hours just before the onset of light, as CAM plants store nocturnally assimilated CO2 in the leaf as malic acid until stomata close during the day. Leaf samples were collected from healthy mature plants of the miniature orchid species *E. pusilla* and *E. crista-galli*. For *E. pusilla,* one plant per time point was collected, and roughly half the plant was used for titratable acidity. For *E. crista-galli,* mature plants were transported to a growth chamber, and kept under low light conditions (~70 umol.m2.sec-1), 12h light/12 dark conditions, 60% humidity, 26C during the daytime and 24C during nighttime. Plants were acclimated for 48 hours before collecting tissue. Leaves from mature plants were collected every 4 hours, flashed frozen in liquid nitrogen, weighed, and boiled in 20% ethanol and deionized water. Titratable acidity was measured as the amount of NaOH required to neutralize the extract to a pH of 7. Two samples in *E. pusilla,* one at 12 AM and one at 2 AM, showed an abnormally low titratable acidity during the nighttime and were removed from the analysis to avoid compromising variability (Table. S3.1).

Library Construction, Sequencing, Assembly, and Abundance Estimation

Library construction, sequencing, assembly and abundance estimation was conducted as described in Chapter 1. Briefly, RNA was isolated from the leaf tissue of plants grown in 12-hour day and 12-hour night conditions. The first two E. pusilla data sets contain samples taken every 4 hours for a total of 27 and 6 samples, while the third *E. pusilla* data set, the set used in the expression analysis, contains samples taken every 2 hours for a total of 24 samples. The *E. crista-galli* data set contains samples taken every 4 hours for a total of 24 samples. Assemblies were generated following the Trinity v2.0.6 (Haas et al., 2013) pipeline, utilizing both its in-silico normalization algorithm and Trimmomatic implementation. Transcript annotation was determined using Trinotate v3.0.1 in the Trinity pipeline. Annotations were only retained at e-

value <= 1E-5 and percent identity >= 50. The best hit to an Arabidopsis or plant species was used as the primary annotation. Read mapping and abundance estimation was conducted using RSEM v1.3.0 (Li & Dewey, 2011) and Bowtie2 v2.2.9 (Langmead, 2013). Genes less than 600bp or FPKM of 1 were removed.

Time-dependent Clustering Analysis

To incorporate time into our clustering analysis, we used R software package maSigPro v1.46.0 (Conesa et al., 2006), which specifically analyzes expression data for patterns across time by fitting each gene's expression pattern to a polynomial using stepwise regression. We used a generalized linear model under the Gaussian distribution to the 4^{th} degree for the regression fit over using a negative binomial model, which had resulted in issues with convergence. Genes whose expression did not exhibit time structured variation with a $p < 0.05$, or those with a "flat" polynomial, were filtered out. Any genes considered outliers based on DFBETAS diagnostic, or the studentized change (estimated standard deviation of the fit at that point) in the predicted value for a point when left out of the regression (Belsley, Kuuh, & E., 1980), were removed from the set to avoid possible bias. 873 transcripts from *E. pusilla* and 58 transcripts from *E. crista-galli* were removed. The remaining isoforms that did show variation across time with a significance < 0.05 were clustered by fuzzy clustering into a specified number of groups based on similarity in pattern. An optimal fuzzifier m, a parameter that determines how much clusters overlap to control noise in the data, for the clustering was calculated as specified by Schwämmle & Jensen (2010). The number of groups was determined by observing the percentage of within-group variance explained against different number of groups and choosing the point at which the marginal gain drops with increase in group number, also known as the elbow method (Fig. S3.1). Summary distributions of each resolved cluster was calculated by

taking the median value of each replicate at each time point and plotting a line across the mean of that median profile. To quantify similarity between clusters, we calculated Pearson's correlation coefficient between the median profiles of pairs of clusters in *E. pusilla* and *E. crista-galli* (Fig. 3.3A).

Orthology and Gene Trees

To determine gene family relationships, genes showing time-dependent expression by maSigPro were sorted into 14 orthogroup gene families from the genomes of the following: *Amborella trichopoda*, *Ananas* comosus (pineapple*), Arabidopsis thaliana*, *Asparagus officinalis*, *Brachypodium distachyon*, *Carica papaya*, *Dendrobium catenatum* (orchid), *Elaeis guineesis* (oil palm), *Musa acuminate* (banana), *Oryza sativa (*rice), *Phalaenopsis equestris* (orchid), *Solanum lycopersium* (tomato), *Sorghum bicolor*, *Spirodella polyrhiza* (duckweed), *Vitis vinifera* (grape vine), and *Zostera marina* (seagrass). The gene family annotations were used to characterize clusters by which gene families were present and at what percentage. Cross-cluster comparisons highlighted which gene groups were enriched in certain clusters, and whether these patterns existed in both *E. pusilla* and *E. crista-galli*.

A consensus gene trees of PEPC was generated using RAxML v8.2.4 (Stamatakis, 2006) using a JTT substitution matrix under the GAMMA model with 100 bootstraps to determine orthologous relationships between gene copies in *E. pusilla* and *E. crista-galli*. Outgroup PEPC genes used were from species *Dendrobium catenatum* and *x Mokara cv. 'Yellow'* on NCBI. Multiple sequence alignments used were created using CLUSTAL Omega (Sievers et al., 2011). All resulting trees were viewed in FigTree v1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Cluster Analysis

To identify notable regulatory candidates possibly involved in CAM, and examine the relationships between genes within clusters, we used the ARACNe-AP v1.4 (Lachmann et al., 2016) algorithm. The algorithm randomly samples gene pairs and uses an adaptive partitioning approach to infer a pairwise mutual information (MI) statistic, or measure of statistical dependence, between them, which is used to construct a network. This process is repeated iteratively for a specified number of bootstraps, while at each step removing those interactions ruled as indirect based on a data processing inequality (DPI) tolerance filter. At the construction of each network, the DPI method removes the edge with the smallest interaction in a set of three genes, ensuring that mutually independent genes are never connected by an edge. A final network is built based on the consensus of all bootstrap runs over a Poisson distribution. Although ARACNe provides an option to specify transcription factors to generate a directed network by only considering interactions with a transcription factor source, we chose to generate an undirected network.

A network was generated using ARACNe-AP on all four clusters in *E. pusilla*. The network were further divided into cluster pairs with antagonistic temporal patterns based on a high cluster correlation (Table S3.2, S3.3): EP1 with EP3 (N-CAM-1-3) and EP2 with EP4 (N-CAM-2-4). To look at direct interactions with PEPC, a sub-network for PEPC (N-PEPC) was created by including only those genes that are up to 4th neighbors to the PEPC node in network EP1-EP3, which was the smallest neighbor rank that included genes from cluster EP3. Node A's neighbor is defined as a node directly connected to node A; node A's 1st neighbor is defined as a the neighbor of a node directly connected to node A.

We imported the data into Cytoscape (Shannon et al., 2005) to generate visualizations of the network and calculate network statistics. To identify significant nodes in the network, we determined hub nodes based on high degree, bridging, and radiality rankings. The degree of a node is defined as the number of edges linked to a node, the connectivity is defined as the number of nodes it is connected to, and an indication of bridging is the number of nodes in the shortest path between other nodes. The radiality of a node is also called its node centrality index and is calculated by subtracting the average shortest path length of the node from the diameter of a connected node plus 1. A high radiality is a good measure of a node that bridges, or connects, other nodes together.

RESULTS

Identification of PEPC Genes

There are three PEPC copies in *E. pusilla*: TR37067|c1_g1 (EP-PEPC1; 474 aa), TR37067|c1_g4 (EP-PEPC2; 380 aa), and TR38975|c0_g1 (EP-PEPC3; 981 aa). EP-PEPC3 has the longest peptide and both EP-PEPC1 and EP-PEPC2 uniquely span its sequence. Likewise, there are three PEPC isoforms in *E. crista-galli*: TR57775|c0_g1 (EC-PEP1; 484 aa), TR38587|c6_g1 (EP-PEPC2; 677 aa), TR38707_c1_g2 (EC-PEPC3; 436 aa). Based on the RAxML gene tree (Fig. 3.2B), EP-PEPC1 and EC-PEPC1, EP-PEPC2 and EC-PEPC2, EP-PEPC3 and EC-PEPC3 are ortholog pairs.

Clustering of genes with time-structured expression profiles

After running maSigPro with optimal parameters, 4 clusters (EP1-4) were resolved in *E. pusilla* and 4 clusters (EC1-4) in *E. crista-galli* (Fig. 3.3A). From the original set of 17,074 transcripts in our *E. pusilla* assembly, 2,373 of those isoforms were determined to have time

structured variation (Table 3.1). Of the 17,257 transcripts in our *E. crista-galli* assembly, we calculated 4,775 of them as having time-dependent expression. Each of the clusters in *E. pusilla* correlate with a cluster in *E. crista-galli*: EP1 and EC1 (r=0.91), EP2 and EC3 (r=0.99), EP3 and EC4 (r=0.89), and EP4 and EC2 (r=0.90). Cluster EP3 in *E. pusilla* contained CAM genes EP-PEPC1 and MDH. In *E. crista-galli*, cluster EC4 contained EC-PEPC1 and PPDK, while cluster EC2 contained PPCK. The clusters containing PEPC in both *E. pusilla* and *E. crista-galli*, EP3 and EC4, had a correlation coefficient of 0.89, which is the highest value for cluster A3 (Table S3.1). While PEPC is shown to have no correlation with time of day in species that undergo C_3 photosynthesis, one copy of PEPC in *E. crista-galli* exhibits CAM-like expression. Moreover, it falls in cluster EC4, which resembles the cluster expression profile in cluster EP3 in *E. pusilla* that also contains PEPC.

Orthogroup Comparison

A total of 1,957 ortho groups, of which 1,797 (91.82%) are unique, were represented by all four *E. pusilla* clusters. The genes in the *E. crista-galli* clusters sorted into 3,569 orthogroups, of which 3,163 (88.62%) are unique and shares 931 groups with *E. pusilla*. Clusters EP1 and EC1 share 71 unique orthogroups, EP2 and EC3 share 129 unique orthogroups, EP3 and EC4 share 147 unique orthogroups, and EP4 and EC2 share 169 unique orthogroups (Fig 3.3B). While the cluster pairs between the two species share high correlation coefficients, the genes within them share <50% of the same orthogroups when sorted with the 14 species list (Fig. 3.3B). The top 20 GO terms enriched in clusters of both species showed little to no difference between and within species (Fig. S3.2)

Network Analysis

The network containing all *E. pusilla* clusters resolved 1,914 nodes and 10,355 edges (Table 3.2). Genes grouped together according to their cluster assignment with few cross-cluster connections (Fig. 3.4A). While ARACNe-AP does not incorporate temporal information within the analysis, we found that the generated networks grouped very well within their assigned gene clusters in *E. pusilla* (Fig. 4A). Those gene clusters with antagonistic expression patterns, like EP1 and EP3, show much fewer cross-cluster connections than with other clusters. Since all genes were included in the analysis, the presence of isolated nodes, or nodes with no significant edges, appear in the network. Out of all the paired sub-networks, the *E. pusilla* network containing PEPC, N-CAM-1-3, achieved the highest average clustering coefficient (Table 3.2).

Sub-network N-PEPC contains 282 genes, including MDH, of which 15 are from cluster EP3. Only 9 of the genes in this network directly interact with PEPC. Two additional sub-networks were created: network containing only those genes shared with *E. crista-galli* (N-PEPC-shared) and a network containing those genes only in *E. pusilla* (N-PEPC-not-shared). PPDK was also present in the full ARACNe-AP network, but only had one weak connection with a single gene; therefore, a PPDK sub-network was not created. A summary of the generated networks can be found in Table 3.3.

To identify possible regulators of CAM, the nodes within each network and sub-network were ranked based on degree, connectivity, stress, and radiality. Many of the genes ranked in the top 10 for all four metrics for the PEPC networks. The annotations of the top genes were used to recognize likely candidates (Table 3.4). A full list of ranked genes for each network can be found in Tables S3.6 - S3.9.

DISCUSSION

PEPC shows CAM-like expression in both species

Not only do *E. pusilla* and *E. crista-galli* share many of the same genes expressing time structured variation, but they also share key CAM genes with temporal patterns as well. The most notable example is PEPC, whose individual and summary expression distributions indicate the gene follows diurnal expression (Fig 3.2A) in both species regardless of photosynthetic pathway. The gene tree analysis confirmed that these two PEPC copies from *E. pusilla* and *E. crista-galli* are orthologs (Fig 3.2B), and thus share an ancestral CAM-like gene expression pattern. Additionally, the expression of both PEPC copies follow their cluster distributions closely, while the other two PEPC copies do not show temporal variation.

It is unusual that a C_3 plant that gains no apparent utility from the CAM process would have a PEPC isoform that mirrors the patterns of its ortholog in a closely related CAM plant. The production of oxaloacetate, the byproduct of PEPC at night, is unnecessary for a plant that steadily collects CO_2 in a diel fashion, especially when there is no increased MDH for conversion to malate to be stored in the vacuole. There are several possible evolutionary explanations as to why this phenomenon might occur, given that we know C_3 photosynthesis is the ancestral state across the Orchidaceae family (Silvera et al., 2009). It is possible that the ancestor of *E. pusilla* and *E. crista-galli* is a CAM plant and CAM was lost in *E. crista-galli*, leaving behind residual CAM gene expression profiles. Therefore, we are seeing a CAM reversal event in *E. crista-galli*. Phylogenetic analysis shows evidence of possible CAM to C_3 reversal events within the Orchidaceae (Silvera et al., 2009). If a CAM-like behavior produced no adverse consequences to the successful functionality of C_3 photosynthesis in *E. crista-galli*, there would be no selective pressure to eradicate the ancestral gene expression. In addition, it is

possible that not enough evolutionary time has passed between the divergence of *E. crista-galli* and its CAM ancestor to completely remove all traces of CAM. Another possibility is that the ancestor of *E. pusilla* and *E. crista-galli* is a C_3 plant that exhibits traits that are pre-disposed for CAM, such as the temporal regulation of canonical CAM genes that we see in *E. crista-galli*. Previously, leaves in C_3 ancestors of a CAM species in the *Agavoideae* was shown to have a predisposition for CAM-like morphology (Heyduk et al., 2016). A predisposition for CAM in the ancestor would suggest that a regulatory rewiring of genes in the CAM pathway led to a susceptibility of an independent evolution of CAM between the C_3 ancestor and *E. pusilla* after the divergence of *E. pusilla* and *E. crista-galli*. This hypothesis correlates with evidence that CAM evolution in orchids transitioned from C_3 ancestors to weak CAM to strong CAM phenotypes (Silvera et al., 2009). On the other hand, *E. crista-galli* did not experience a transition to CAM after its divergence from *E. crista-galli*.

Decarboxylation in E. pusilla

Another canonical CAM gene present in both species is PPDK; however, PPDK in *E. pusilla* falls in EP4 and EC1 in *E. crista-galli*. Despite that, both clusters share a similar diel expression pattern. Following the reduction of malate, PPDK catalyzes the conversion of pyruvate to PEP to eventually be used in sugar production pathways. Presence of PPDK in the diel clusters and absence of PEPCK in any temporal cluster suggests that *E. pusilla* utilizes the NADP-me or NAD-me CAM decarboxylation pathway as opposed to the MDH pathway for the breakdown of malate. Both NADP-me and NAD-me are present in the *E. pusilla* assembly, but were not annotated as a gene that showed time-structured expression by maSigPro. The absence of a malic enzyme in the temporal analysis may be from noisy expression values as a result of sample variability. NAD-me is expressed at a greater abundance than NADP-me, indicating that

E. pusilla is likely to utilize the NAD-me pathway for decarboxylation of malate. However, further investigation is required for confirmation.

Genes cluster into groups based on temporal pattern

Differentiable temporal expression patterns were clearly identified in both *Erycina* species, as evidenced by the gene clusters derived by maSigPro. Each cluster's gene set can be summarized in expression distributions distinct from those of the others; therefore, the total temporal variation of genes and their possible expression patterns can be fully represented by the four clusters for each species. The appearance of PEPC, MDH, PPDK, and PEPCK within the clusters indicate that the analysis isolates temporally-regulated CAM genes from the data set. On the other hand, no assumptions can be made from the absence of key CAM genes from *E. pusilla* clusters, which could have been filtered out as flatly or lowly expressed transcripts due to the high variability between samples and replicates.

While each cluster's temporal pattern is unique, a corresponding cluster in the group exists with a directly opposing pattern. This is the case for both *E. pusilla* and *E. crista-galli*. Furthermore, each cluster in *E. pusilla* can be reasonably paired with a similar cluster in *E. crista-galli*, which indicates that both species share an identical set of temporal patterns, allowing simple cross-cluster comparisons between the two species.

Network analysis

Networks are often used to model regulatory interactions between proteins or other molecules in biological systems. Hub nodes, or nodes that play central nodes in networks, have been found to play significant roles in biological networks. Loss of a hub node within a network can be detrimental or even lethal (Jeong et al., 2001). In addition, hubs occupy advantageous positions in networks because of their high connectivity, which allow them to spread information

quickly and efficiently. Because hub nodes serve as bridges between nodes with low connectivity, they are powerful communicators and likely master regulators of biological processes that the network is involved in. Therefore, applying a network-based approach to the analysis of gene clusters containing canonical CAM genes in *E. pusilla* allows easy identification of possible CAM gene regulators.

Isolating the PEPC network by those genes that are not found in the corresponding *E. crista-galli* cluster but including PEPC, we find that the top ranked gene for high connectivity, bridging, and bottlenecks codes for protein PLASTID TRANSCRIPTIONALLY ACTIVE 12 (PTAC12), involved in the initiation of photomorphogenesis in the phytochrome signaling pathway (Chen et al., 2010). Other highly ranked genes include homologs of plastid-lipid-associated protein 4 (PAP4), PLASTID TRANSCRIPTIONALLY ACTIVE 14 (PTAC14), and pheophorbide a oxygenase (PAO) and pheophytinase (PPH) which are involved in the breakdown of chlorophyll molecules.

Conversely, isolating the PEPC network by those genes that are shared between *E. pusilla* and *E. crista-galli* and retrieving the top ranked nodes by degree, stress, and radiality yields a higher proportion of uncharacterized proteins, many of which are pentatricopeptide repeat-containing proteins. However, a few genes are involved in the chloroplast: aluminum-activated malate transporter 9 (ALMT9) and adagio protein 3 (ADO3). ALMT9 is a malate transporter in the chloroplast that is likely to be responsible for shuttling the malate converted by MDH into the vacuole during the night (Kovermann et al., 2007), which correlates with the expression pattern of the cluster ALMT9 is assigned. It is unusual, though, that this gene shows a CAM-like expression pattern in both species, despite *E. crista-galli* lacking a temporal MDH. Perhaps the diel expression of ALMT9 is the result of the same evolutionary forces that acted on

PEPC in *Erycina*. Although ALMT9 is more likely to be a member of the CAM pathway and not a major regulator, the second gene, ADO3/FKF1, shows more promise. ADO3 is a blue light photoreceptor involved in flowering that mediates the degradation of CONSTANS (CO) expression alongside GIGANTEA (GI) in the presence of FMN (Nelson, Lasswell, Rogg, Cohen, & Bartel, 2000; Takase et al., 2011). On long days, FKF1 was shown to stabilize CO protein in the late afternoon (Song et al., 2012), which also coincides with the temporal pattern of the cluster it is contained. While ADO3 is primarily attributed to regulation of the flowering pathway, it is possible that it is involved in the CAM regulatory pathway, as CAM is innately dependent on the circadian clock. In fact, it is highly likely that there is overlap between the circadian clock and CAM pathways.

The ranked hub candidates in network N-CAM-2-4 were much less annotated than that of N-CAM-1-3. The top candidate based on rank in this network is 25.3 kDA heat shock protein, chloroplastic (HS25P), whose GO terms included "response to high light intensity." Although lacking in functional and experimental annotation, HS25P is a strong hub node candidate, and therefore plays a central role in network N-CAM-2-4. Another candidate is CHLOROPLAST IMPORT APPARATUS 2 (CIA2), which upregulates TOC33 and TOC75 leaf expression, two proteins involved in protein import into chloroplasts (Sun, Huang, & Chang, 2009). Finally, there are also three genes involved in ABA biosynthesis. There is also U-box domain-containing protein 44 (PUB44) which represses abscisic acid (ABA) biosynthesis (Raab et al., 2009), fimbrin-5 (FIMB5) which stabilizes and prevents actin depolymerization, and abscisic acid-insensitive 5-like protein (AI5L5), which also regulates ABA signaling. Elevated ABA levels can result in stomatal closure (Daszkowska-Golec & Szarejko, 2013). Therefore, when these

genes are down-regulated during the day, the ABA synthesis pathway is likely up-regulated, increasing ABA levels that result in the closing of stomata.

FUTURE WORK AND CONCLUSION

A comparative time-series expression analysis is a novel and effective method for studying CAM photosynthesis. While the expression of CAM genes relies heavily on time of day, previous CAM studies have not utilized transcriptomics with a time series design. We have demonstrated here that taking an approach that incorporates temporal information can isolate genes with distinct patterns across time and that a gene network analysis on temporal clusters can highlight candidate genes or regulators involved in the CAM pathway. Additionally, having a CAM species and a C_3 species within the same genus is advantageous in studying the evolution of CAM on a more phylogenetically narrow scale. Direct comparison between gene clusters or networks can elucidate changes in expression patterns since the divergence of the two species.

It is likely that humidity variation in the flasks in which *E. pusilla* plants were grown, resulting in among-sample variation in carbon metabolism and gene expression, limited our ability to identify common circadian gene expression profiles. As a result, many genes are likely absent from the maSigPro reduced set. Genes not determined to be temporal included canonical CAM genes like NAD-me and NADP-me and other reference clock genes that showed time-structured expression profiles in *E. crista-galli* but not in *E. pusilla*. Therefore, the only conclusions we could infer about gene expression had to be based on the presence of a gene in the set showing diel expression profiles, and not their absence.

Here we presented an exploratory framework to study CAM photosynthesis in *Erycina* using time series RNA-Seq data. By combining a temporal clustering and network analysis

approach, we expanded the knowledge of what we know about CAM evolution in *Erycina* and provided a first step to discovering novel regulators in the CAM pathway. Future work could focus on applying this approach to other relatives of *E. pusilla* or in similar families in the Orchidaceae that exhibit both C_3 and CAM traits. Even further investigations could involve comparison with species that exhibit strong or weak CAM phenotypes. Studies in proteomics on these CAM genes can also confirm the temporal expression patterns that we are seeing.

REFERENCES

Anders, S., & Huber, W. (2012). Differential expression of RNA-Seq data at the gene level. *EMBL*.

Belsley, D. A., Kuuh, E., & E., W. R. (1980). Wiley Series in Probability and Statistics. *Wiley Series in Probability and Statistics*. https://doi.org/10.1002/9780470316559.oth1

Borland, A. M., Hartwell, J., Jenkins, G. I., Wilkins, M. B., Nimmo, H. G., Borland, A. M., … Nimmo, H. G. (1999). Metabolite Control Overrides Circadian Regulation of Phosphoenolpyruvate Carboxylase Kinase and CO_2 Fixation in Crassulacean Acid Metabolism. *Plant Physiology*, 121(3), 889–896. Retrieved from http://www.jstor.org/stable/4279010

Ceusters, J., Borland, A. M., Taybi, T., Frans, M., Godts, C., & De Proft, M. P. (2014). Light quality modulates metabolic synchronization over the diel phases of crassulacean acid metabolism. *Journal of Experimental Botany*, 65(13), 3705–3714. https://doi.org/10.1093/jxb/eru185

Chase, M. W., Cameron, K. M., Freudenstein, J. V., Pridgeon, A. M., Salazar, G., van den Berg, C., & Schuiteman, A. (2015). An updated classification of Orchidaceae. *Botanical Journal of the Linnean Society*, 177(2), 151–174. https://doi.org/10.1111/boj.12234

Chen, M., Galvao, R. M., Li, M., Burger, B., Bugea, J., Bolado, J., & Chory, J. (2010). Arabidopsis HEMERA/pTAC12 initiates photomorphogenesis by phytochromes. *Cell*, 141(7), 1230–1240. https://doi.org/10.3816/CLM.2009.n.003.Novel

Christmas, Rowan; Avila-Campillo, Iliana; Bolouri, Hamid; Schwikowski, Benno; Anderson, Mark; Kelley, Ryan; Landys, Nerius; Workman, Chris; Ideker, Trey; Cerami, Ethan; Sheridan, Rob; Bader, Gary D.; Sander, C. (2005). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *American Association for Cancer Research Education Book*, (Karp 2001), 12–16. https://doi.org/10.1101/gr.1239303.metabolite

Conesa, A., Nueda, M. J., Ferrer, A., & Talón, M. (2006). maSigPro: A method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, 22(9), 1096–1102. https://doi.org/10.1093/bioinformatics/btl056

Daszkowska-Golec, A., & Szarejko, I. (2013). Open or close the gate - stomata action under the control of phytohormones in drought stress conditions. *Frontiers in Plant Science*, 4(May), 138. https://doi.org/10.3389/fpls.2013.00138

DePaoli, H. C., Borland, A. M., Tuskan, G. A., Cushman, J. C., & Yang, X. (2014). Synthetic biology as it relates to CAM photosynthesis: Challenges and opportunities. *Journal of Experimental Botany*, 65(13), 3381–3393. https://doi.org/10.1093/jxb/eru038

Drennan, P. M., & Nobel, P. S. (2000). Responses of CAM species to increasing atmospheric CO2 concentrations. *Plant, Cell and Environment*, 23(8), 767–781. https://doi.org/10.1046/j.1365-3040.2000.00588.x

Farquhar, G. D., & Sharkey, T. D. (1982). Stomatal Conductance and Photosynthesis. *Annual Review of Plant Physiology*, 33(1), 317–345. *https://doi.org/10.1146/annurev.pp.33.060182.001533*

Gehrig, H. H., Wood, J. A., Cushman, M. A., Virgo, A., Cushman, J. C., & Winter, K. (2005). Large gene family of phosphoenolpyruvate carboxylase in the crassulacean acid metabolism plant Kalanchoe pinnata (Crassulaceae) characterised by partial cDNA sequence analysis. *Functional Plant Biology*, 32, 467. https://doi.org/10.1071/FP05079

Gehrig, H., Heute, V., & Kluge, M. (2001). New partial sequences of phosphoenolpyruvate carboxylase as molecular phylogenetic markers. *Molecular Phylogenetics and Evolution*, 20(2), 262–274. https://doi.org/10.1006/mpev.2001.0973

Graham, E. A., & Nobel, P. S. (1996). Long-term effects of a doubled atmospheric CO2 concentration on the CAM species *Agave deserti*. *Journal of Experimental Botany*, 47(294), 61–69. https://doi.org/10.1093/jxb/47.1.61

Gravendeel, B., Smithson, A., Slik, F. J. W., & Schuiteman, A. (2004). Epiphytism and pollinator specialization: drivers for orchid diversity? Philosophical Transactions of the Royal Society of London. *Series B, Biological Sciences*, 359(1450), 1523–1535. https://doi.org/10.1098/rstb.2004.1529

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Philip, D., Bowden, J., … Pochet, N. (2013). *De novo* transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat Protocols* (Vol. 8). Nature Protocols. https://doi.org/10.1038/nprot.2013.084.De

Jeong, H., Mason, S. P., Barabási, a L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833), 41–42. https://doi.org/10.1038/35075138

Kovermann, P., Meyer, S., Hörtensteiner, S., Picco, C., Scholz-Starke, J., Ravera, S., … Martinoia, E. (2007). The Arabidopsis vacuolar malate channel is a member of the ALMT family. *Plant Journal*, 52(6), 1169–1180. https://doi.org/10.1111/j.1365-313X.2007.03367.x

Lachmann, A., Giorgi, F. M., Lopez, G., & Califano, A. (2016). ARACNe-AP: Gene network reverse engineering through adaptive partitioning inference of mutual information. *Bioinformatics*, 32(14), 2233–2235. https://doi.org/10.1093/bioinformatics/btw216

Langmead. (2013). Bowtie2. *Nature Methods*, 9(4), 357–359. https://doi.org/10.1038/nmeth.1923.Fast

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15, R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lee, S. H., Li, C. W., Liau, C. H., Chang, P. Y., Liao, L. J., Lin, C. S., & Chan, M. T. (2015). Establishment of an *Agrobacterium*-mediated genetic transformation procedure for the experimental model orchid *Erycina pusilla*. *Plant Cell*, Tissue and Organ Culture, 120(1), 211–220. https://doi.org/10.1007/s11240-014-0596-z

Leegood, R. C., & Walker, R. P. (1996). Phosphorylation of phosphoenolpyruvate carboxykinase in plants. *Biochemical Journal*, 317, 653–658.

Lepiniec, L., Vidal, J., Chollet, R., Gadal, P., & Cretin, C. (1994). Phosphoenolpyruvate carboxylase: structure, regulation and evolution. *Plant Science*, 99(2), 111–124. https://doi.org/10.1016/0168-9452(94)90168-6

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, 12, 323. https://doi.org/10.1186/1471-2105-12-323

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., … Yu, Q. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nature Genetics*, 47(12), 1435–1442. https://doi.org/10.1038/ng.3435

Nelson, D. C., Lasswell, J., Rogg, L. E., Cohen, M. a, & Bartel, B. (2000). FKF1, a clock-controlled gene that regulates the transition to flowering in *Arabidopsis*. *Cell*, 101(3), 331–340. https://doi.org/10.1016/S0092-8674(00)80842-9

Nimmo, G. A., Mcnaughton, G. A. L., Fewson, C. A., Wilkins, M. B., & Nimmo, H. G. (1987). Changes in the kinetic properties and phosphorylation state of phosphoenolpyruvate carboxylase in *Zea mays* leaves in reponse to light and dark. FEBS Letters, 213(1), 18–22. https://doi.org/10.1016/0014-5793(87)81457-6

Nimmo, H. G. (2000). The regulation of phosphoenolpyruvate carboxylase in CAM plants. *Trends in Plant Science*, 5(2), 75–80. https://doi.org/10.1016/S1360-1385(99)01543-5

Raab, S., Drechsel, G., Zarepour, M., Hartung, W., Koshiba, T., Bittner, F., & Hoth, S. (2009). Identification of a novel E3 ubiquitin ligase that is required for suppression of premature senescence in *Arabidopsis*. *Plant Journal*, 59(1), 39–51. https://doi.org/10.1111/j.1365-313X.2009.03846.x

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Schwämmle, V., & Jensen, O. N. (2010). A simple and fast method to determine the parameters for fuzzy c-means cluster analysis. *Bioinformatics*, 26(22), 2841–2848. https://doi.org/10.1093/bioinformatics/btq534

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., … Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), 539. https://doi.org/10.1038/msb.2011.75

Silvera, K., Neubig, K. M., Whitten, W. M., Williams, N. H., Winter, K., & Cushman, J. C. (2010). Evolution along the crassulacean acid metabolism continuum. *Functional Plant Biology*, 37(11), 995–1010. https://doi.org/10.1071/FP10084

Silvera, K., Santiago, L. S., Cushman, J. C., Winter, K., & Rica, C. (2013). Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. *Plant Physiology*, 149(4), 1838–1847. https://doi.org/10.1104/pp.l08.132555

Song, Y. H., Smith, R. W., To, B. J., Millar, A. J., & Imaizumi, T. (2012). FKF1 conveys crucial timing information for CONSTANS stabilization in the photoperiodic flowering. *Science*, 336(6084), 1045–1049. https://doi.org/10.1126/science.1219644.FKF1

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21), 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Sun, C., Huang, Y., & Chang, H. (2009). CIA2 Coordinately Up-Regulates Protein Import and Synthesis in Leaf Chloroplasts. *Plant Physiology*, 150(2), 879–888. https://doi.org/10.1104/pp.109.137240

Takase, T., Nishiyama, Y., Tanihigashi, H., Ogura, Y., Miyazaki, Y., Yamada, Y., & Kiyosue, T. (2011). LOV KELCH PROTEIN2 and ZEITLUPE repress Arabidopsis photoperiodic flowering under non-inductive conditions, dependent on FLAVIN-BINDING KELCH REPEAT F - BOX1. *Plant Journal*, 67(4), 608–621. https://doi.org/10.1111/j.1365-313X.2011.04618.x

Taybi, T., Patil, S., Chollet, R., Cushman, J. C., Taybi, T., Patil, S., … Cushman, J. C. (2000). A Minimal Serine / Threonine Protein Kinase Circadianly Regulates Phosphoenolpyruvate

Carboxylase Activity in Crassulacean Acid Metabolism-Induced Leaves of the Common Ice Plant. *Plant Physiology*, 123(4), 1471–1481. Retrieved from http://www.jstor.org/stable/4279381

Walker, R. P., Trevanion, S. J., & Leegood, R. C. (1995). Phosphoenolpyruvate Carboxykinase from Higher-Plants - Purification from Cucumber and Evidence of Rapid Proteolytic Cleavage in Extracts from a Range of Plant-Tissues. *Planta*, 196(1), 58–63.

Table 3.1 – Gene counts for gene clusters resolved by maSigPro.

	E. pusilla	***E. crista-galli***
Filtered Assembly	17,074	17,257
Temporal	2276	4775
Outliers (Removed)	873	58
Cluster 1	626	760
Cluster 2	489	1413
Cluster 3	630	1438
Cluster 4	531	1164

Table 3.2 – Gene networks of clusters resolved by ARACNe-AP in *E. pusilla*. Genes shared/non shared between *E. pusilla* and *E. crista-galli* were determined using shared gene families.

Network	**Description**	**Nodes**	**Edges**	**Clustering Coefficient**	**Heterogeneity**
N-CAM	All *E. pusilla* clusters	1914	10355	0.187	1.134
N-CAM-1-3	EP1, EP3	1106	6026	0.204	1.044
N-CAM-2-4	EP2, EP4	808	2709	0.181	1.179
N-PEPC	PEPC + 1^{st}, 2^{nd}, 3^{rd}, and 4^{th} neighbors	282	1299	0.257	0.888
N-PEPC-not-shared	N-PEPC nodes not in *E. crista-galli*	160	342	0.200	0.927
N-PEPC-shared	N-PEPC nodes shared with *E. crista-galli*	112	382	0.312	0.907

Table 3.3 – Possible CAM regulators derived from hub nodes found in ARACNe-AP networks.

Network	Gene	Cluster	Degree Rank	Stress Rank	Radiality Rank
N-PEPC-not-shared	PTA12	3	1	3	2
N-PEPC-not-shared	PAP4	3	4	11	7
N-PEPC-not-shared	PAO	3	6	17	8
N-PEPC-not-shared	PPH	3	11	48	50
N-PEPC-shared	ALMT9	3	18	28	18
N-PEPC-shared	ADO3	3	22	17	19
N-PEPC-2-4	HS25P	2	1	4	5
N-PEPC-2-4	CIA2	2	<50	12	7
N-PEPC-2-4	PUB44	4	1	7	12

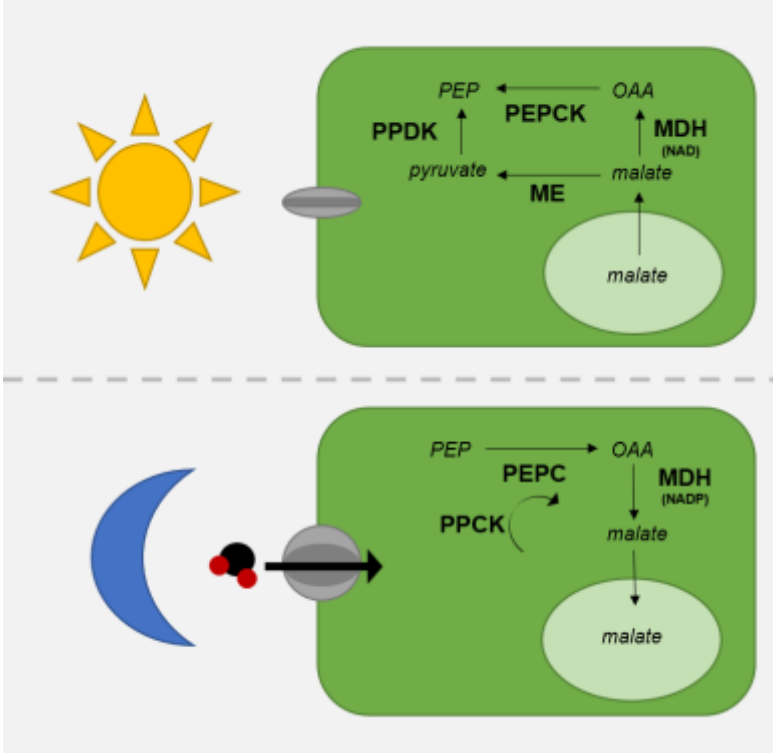


Figure 3.1 - A simplified diagram of the Crassalucean acid metabolism (CAM) pathway under day and night conditions.

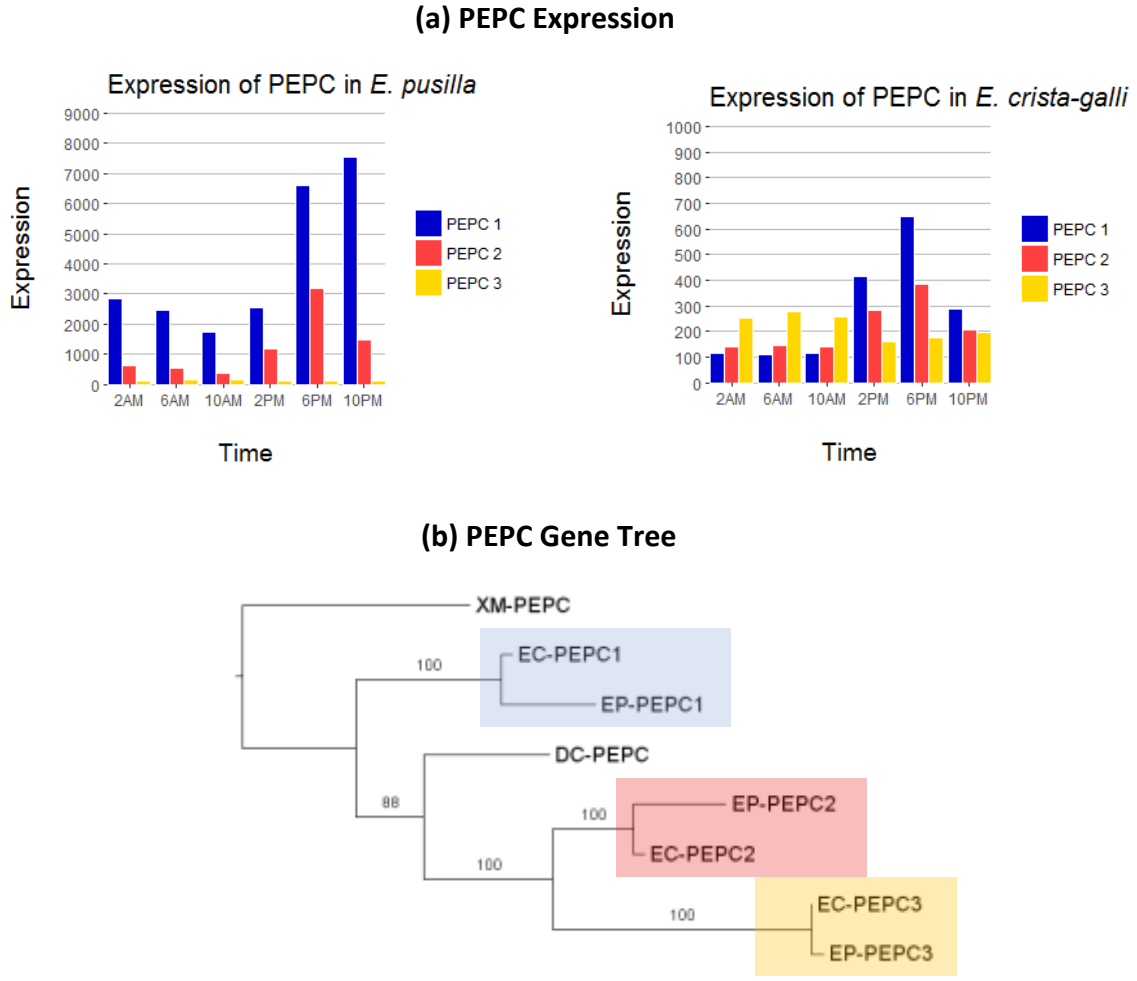


Figure 3.2 – Phosphoenlpyruvate carboxylase (PEPC) orthologs in *E. pusilla* and *E. crista-galli***.** A) Median expression of three PEPC isoforms within *E. pusilla* and *E. crista-galli* across time; time points in *E. pusilla* not in *E. crista-galli* were removed for better comparison (Note the difference in the y-axis) B) A RAxML gene tree of the PEPC isoforms in *E. pusilla* and *E. crista-galli* shown in (a); *X Mokara cv. 'Yellow'* (XM), *D. catenatum* (DC), *E. pusilla* (EP), *E. crista-galli* (EC).

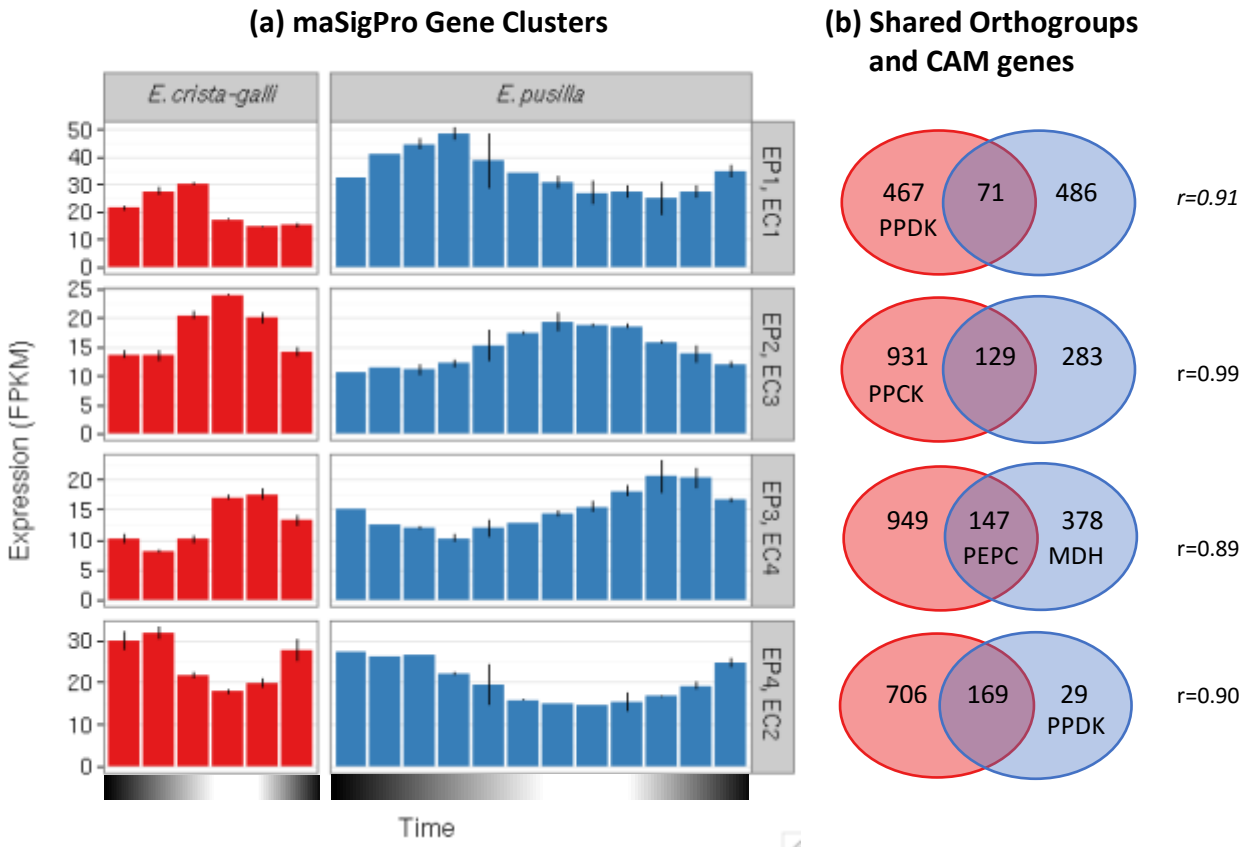


Figure 3.3 – Summary distributions and shared orthogroups in maSigPro gene clusters. A) Summary distributions of gene clusters resolved in *E. pusilla* (starting at 12AM) and *E. crista-galli* (starting at 2AM) using maSigPro calculated from the mean of median time points across genes. B) Gene family enrichment analysis using the 14 species gene set between cross-species pairs of clusters (only unique orthogroups counted).

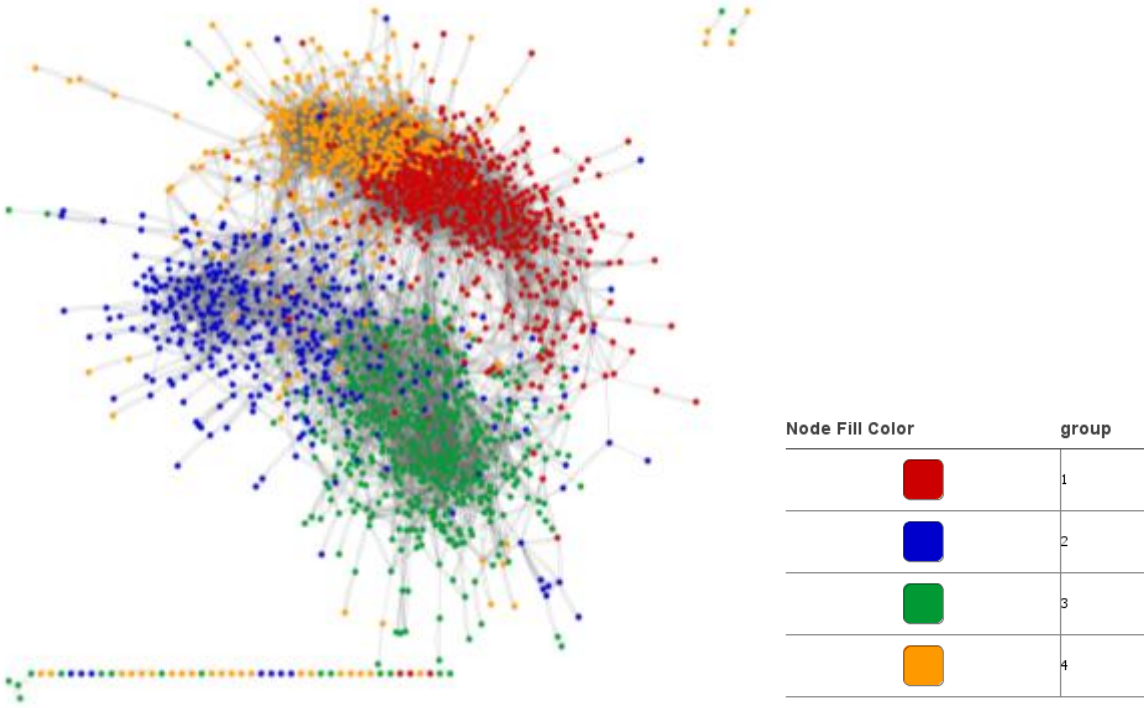


Figure. 3.4 – Network of maSigPro gene clusters generated with ARACNe-AP.

CHAPTER IV

CONCLUSION AND DISCUSSION

TRANSCRIPTOMIC METHODS FOR CAM

As RNA-Seq becomes more and more popular in the study of transcriptomics, methodologies in experimental design, assembly, and analysis must be investigated in parallel. Transcriptome assembly, in particular, is a difficult step with important decisions that many studies have glossed over. One major decision that can greatly affect the assembly is the number of input reads. In Ch1, I showed that the inclusion of too many reads may be detrimental when generating an assembly in Trinity. Further, taking time to evaluate assembly quality across filtering thresholds using different quality assessment strategies can improve resolution in downstream analyses. I also demonstrated that choice of FPKM filtering alone has a dynamic impact on assembly quality. Lastly, quality assessment is best performed with a combination of different statistics, such as read mapping rate or conserved protein representation, as different metrics can exhibit different behaviors across assemblies under varying filtering thresholds.

Best practices in gene expression analysis are vastly contingent on the research question being investigated. In this study, the research focus was to investigate the evolution and regulation of CAM photosynthesis. The dynamic expression of CAM genes across time cannot be captured using traditional differential expression analyses used in binary experiments. Using a gene expression clustering algorithm based on pattern recognition methods allowed

incorporation of temporal information to the analysis to identify distinct temporal expression patterns in *E. pusilla* and *E. crista-galli*.

CAM IN ERYCINA

Because the Orchidaceae constitutes a phenotypically diverse group of plants in terms of photosynthesis, many sub-families, like *Erycina*, are favorable candidates for a comparative analysis. Yet, CAM photosynthesis in the Orchidaceae has hardly been explored. In ChIII, I presented a first attempt to uncover the nuances of CAM in the Orchidaceae using a comparative transcriptomics-based approach between a CAM species and a C_3 species. Gene expression analyses revealed canonical CAM genes showing CAM-like expression patterns in *E. crista-galli*, a C_3 species, which raises many hypotheses as to how CAM evolved in *Erycina* and whether other C_3 ancestors in the Orchidaceae show a pre-disposition for CAM before the evolution of CAM. Additional network analyses on gene expression clusters with similar temporal patterns highlighted genes that held important roles within clusters, raising the possibility of playing regulatory roles in the CAM photosynthesis pathway. Overall, this transcriptomic study of gene expression in CAM between *E. pusilla* and *E. crista-galli* builds a gateway into further exploration of CAM evolution in the Orchidaceae and the introduction of *E. pusilla* as a model species for CAM in the Orchidaceae.

FUTURE DIRECTIONS

Optimizing methodologies for RNA-Seq experiments is no easy task, but continued study on assessment methods in assembly and expression analyses will improve best practices in the future. Assembly evaluation across varying filtering threshold levels is only one approach to

measure quality. Similarly, a regression-based approach to clustering time series expression data offered an alternative to differential expression analyses, but many other variations exist, such as model-based methods or machine learning algorithms, that can be applied to analyze temporal data sets. An evaluation of these different clustering approaches would benefit future work in temporally expressed gene data sets.

Understanding the evolution of CAM in the Orchidaceae and in *Erycina* necessitates further investigation between close C_3 and CAM relatives. Here we provide support for the possibility of C_3 ancestors of CAM species undergoing a regulatory rewiring of genes in preparation for evolved CAM function. Temporal expression data from *E. pusilla* and *E. crista-galli* revealed canonical CAM genes exhibiting characteristic CAM expression patterns in both species, despite *E. crista-galli* being a C_3 plant. Since there have been multiple independent origins of CAM evolution within the Orchidaceae, evidence of CAM pre-disposition in other lineages should be explored, and a transcriptomics approach may be a strategic means to do so.

APPENDIX A

SUPPLEMENTARY FIGURES AND TABLES FROM CHAPTER II

Table S2.1 – Reference table of canonical CAM and clock genes.

	Gene Name	**Gene Symbol**
CAM	Malate dehydrogenase	MDHP
CAM	NADP-dependent malic enzyme	NADP-me, MAOP
CAM	NAD-dependent malic enzyme	NAD-me, MAO
CAM	Phosphoenolpyruvate carboxylase	PEPC, CAPP
CAM	Phosphoenolpyruvate carboxykinase	PEPCK, PCK
CAM	Phosphoenolpyruvate carboxylase kinase	PPCK, PPCK
CAM	Pyruvate phosphate dikinase	PPDK
Clock	Adagio protein 1	ADO1/ZTL
Clock	Circadian clock associated 1	CCA1
Clock	Cryptochrome-1	CRY1
Clock	De-etiolated homolog 1	DET1
Clock	EARLY FLOWERING 3	ELF3
Clock	GIGANTEA	GI
Clock	Late elongated hypocotyl	LHY
Clock	LUX ARRHYTHMO/PHYTOCLOCK1	LUX/PCL1
Clock	Phytochrome A	PHYA
Clock	Phytochrome B	PHYB
Clock	Phytochrome-interacting factor 3	PIF3
Clock	PSEUDO RESPONSE REGULATOR 3, 5, 7, 9	PRR3, PRR5, PRR7, PRR9
Clock	Two-component response regulator-like APRR1	TOC1/APRR1
Clock	SNW/SKI-interacting protein	SKIP

Note. Clock genes reference from: Bendix, C., Marshall, C. M., & Harmon, F. G. (2015). Circadian clock genes universally control key agricultural traits. *Molecular plant*, *8*(8), 1135-1152. CAM genes referenced from: Borland, A. M., Barrera Zambrano, V. A., Ceusters, J., & Shorrock, K. (2011). The photosynthetic plasticity of crassulacean acid metabolism: an evolutionary innovation for sustainable productivity in a changing world. *New Phytologist*, *191*(3), 619-633.

Table S2.2 – Parameters for software used in assembly construction and assessment.

Software	**Parameters**
BLAT v2	tileSize=11, stepSize=11, minMatch=2, minScore=30, minIdentity=90, maxGap=2, -fine, maxIntron=750000
BUSCO v2	ev=0.01, mode=trans
RSEM v1.3.0	Default settings in the Trinity pipeline manual
TransRate vXX	--retain_long_orfs, --retain_pfam_hits, retain_blastp_hits
TransDecoder v2.0.1	Default settings in the Trinity pipeline manual
Trinity v2.0.6	Default settings in the Trinity pipeline manual

Note. Protocol for Trinity pipeline manual from: Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013 Aug;8(8):1494-512. Open Access in PMC doi: 10.1038/nprot.2013.084.

APPENDIX B

SUPPLEMENTARY FIGURES AND TABLES FROM CHAPTER III

Table S3.1 – Titratable acidity measurements in *E. pusilla*. (Mean = 28.56, median = 25.915, stdev = 11.83)

Sample	Time	TA
36	12AM	38.46
3	12AM	16.28
28	2AM	2.81
7	2AM	30.27
25	4AM	29.27
4	4AM	42.86
15	6AM	57.14
5	6AM	50
29	8AM	37.41
32	8AM	29.78
31	10AM	20.47
6	10AM	41.88
18	12PM	31.33
27	12PM	16.13
30	2PM	25.47
33	2PM	23.42
8	4PM	24.15
26	4PM	38.46
34	6PM	19.39
13	6PM	22.77
35	8PM	19.91
11	8PM	26.36
1	10PM	19.9
9	10PM	21.47

Table S3.2 – Canonical clock genes found in maSigPro gene clusters.

Gene Name	Gene Symbol	Cluster
Adagio protein 1	ADO1	EP3, EC3
Circadian clock associated 1	CCA1	EP4, EC1
Cryptochrome-1	CRY1	EP4, EC1
Early flowering 3	ELF3	EP3, EC4
Gigantea protein	GI	EP3, EC3
Late elongated hypocotyl	LHY	EP4, EC1, EC4
Psuedo-response regulator	PRR	EC3, EC4
Phytochrome interacting factor 3	PIF3	EC1
Phytochrome A	PHYA	EC1
Phytochrome B	PHYB	EC2
SNW/Ski-interacting protein	SKIP	EP1, EC2
Timing of CAB expression 1	TOC1	EP1

Table S3.3 – Between-species cluster correlations.

	EC1	**EC2**	**EC3**	**EC4**
EP1	0.63 (0.18)	0.91 (0.011)	-0.78 (0.62)	-0.92 (0.0084)
EP2	0.008 (0.99)	-0.93 (0.0070)	0.99 (2.6e-4)	0.55 (0.26)
EP3	-0.82 (0.042)	-0.62 (0.19)	0.41 (0.42)	0.89 (0.017)
EP4	-0.046 (0.93)	0.90 (0.016)	-0.94 (0.0048)	-0.53 (0.28)

Table S3.4 – Within-species *E. pusilla* cluster correlations.

	EP1	**EP2**	**EP3**	**EP4**
EP1	1	-0.60 (0.037)	-0.86 (3.3e-4)	0.63 (0.0297)
EP2	-0.60 (0.038)	1	0.25 (0.43)	-0.96 (6.09e-7)
EP3	-0.86 (3.3e-4)	0.25 (0.43)	1	-0.35 (0.27)
EP4	0.63 (0.030)	-0.96 (6.09e-7)	-0.35 (0.27)	1

Table S3.5 – Within-species *E. crista-galli* cluster correlations.

	EC1	**EC2**	**EC3**	**EC4**
EC1	1	0.32 (0.54)	-0.14 (0.79)	-0.82 (0.05)
EC2	0.32 (0.54)	1	-0.97 (0.0017)	-0.79 (0.063)
EC3	-0.14 (0.80)	-0.97 (0.0017)	1	0.67 (0.15)
EC4	-0.82 (0.046)	-0.79 (0.063)	0.67 (0.15)	1

Table S3.6 – Top ranked genes in N-CAM-1-3 network. GO Term annotations derived from UniProt-GOA database. Ranks are based on genes within the cluster only. Unannotated genes are not listed.

Gene Symbol	GO Terms	Cluster	Degree Rank	Stress Rank	Radiality Rank
PTA12	Response to blue light, positive regulation of red light, chloroplast	3	2	1	36
PP215	RNA binding, RNA modification	3	31	2	2
AAED1	Antioxidant activity, oxidation-reduction process	3	32	3	5
CBP1	Carboxypeptidase activity, secondary metabolic process	3	49	6	12
PPR75	RNA binding, RNA modification	3	1	28	35
PPR70	RNA binding, RNA modification	3	3	11	11
PP285	Zinc ion binding, chloroplast mRNA processing, RNA modification	3	4	<50	<50
PP363	Zinc ion binding, RNA binding, embryo development ending in seed dormancy, RNA modification	3	6	12	6
PIX7	Transmembrane receptor protein serine/threonine kinase activity, cell surface receptor signaling pathway	1	1	1	1
TAF15	RNA binding, zinc ion binding, regulation of transcription	1	2	4	2
GBF1	Regulation of hydrogen peroxide metabolic process, transcription regulatory region DNA binding	1	6	41	36
MTPA2	Zinc ion transmembrane transporter activity, regulation of sequestering of zinc ion	1	7	32	48
PEP	Negative regulation of photoperiodism, RNA binding, shoot system development	1	9	9	6
ERF1Z	Translation release factor activity, regulation of growth, translational termination	1	38	2	3
TIF3A1	Translation initiation factor activity, formation of cytoplasmic translation initiation complex	1	<50	3	4
Y1015		1	50	5	14
STIP1	U1 snRNP binding, spliceosomal complex disassembly	1	22	8	5

Table S3.7 – Top ranked genes in N-PEPC-not-shared network. GO Term annotations derived from UniProt-GOA database. Unannotated genes are not listed.

Gene Symbol	GO Terms	Cluster	Degree Rank	Stress Rank	Radiality Rank
PTA12	Response to blue light, positive regulation of red light, chloroplast	3	1	3	2
PP363	Zinc ion binding, RNA binding, embryo development ending in seed dormancy, RNA modification	3	2	2	1
PAP4	Chloroplast thylakoid membrane, platsoglobule	3	3	11	6
PP313	Zinc ion binding, RNA binding, RNA modification	3	4	4	3
PAO	Pheophorbide a oxygenase activity, chlorophyll catabolic process	3	5	17	7
NFD4	Karyogamy, polar nucleus fusion, response to salt stress, transport	3	6	7	5
SDG40	Peptidyl-lysine monomethylation	3	7	14	11
PPH	Pheophytinase activity, chlorophyll catabolic process	3	12	48	50
PEX6	Transmembrane receptor protein serine/threonine kinase activity, cell surface receptor signaling pathway	3	19	10	14

Table S3.8 – Top ranked genes in N-PEPC-shared network. GO Term annotations derived from UniProt-GOA database. Unannotated genes are not listed.

Gene Symbol	GO Terms	Cluster	Degree Rank	Stress Rank	Radiality Rank
GPAT3	G3P O-acyltransferase activity, CDP-diacylglyercerol biosynthetic process	3	1	1	1
ADO3	Circadian rhythm, flower development, response to blue light, photoreceptor activity	3	24	17	18
ELI1	Chloroplast modification, zinc ion binding	3	11	18	9
PHT4	L-ascorbic acid transporter activity, xanthophyll cycle	3	14	21	15
ALMT9	Malate transmembrane transporter activity, malate transport	3	21	28	17
PPR Proteins	RNA binding, RNA modification	3	2 – 10, 12	2 - 16	2 – 9

Table S3.9 – Top ranked genes in N-CAM-02-04 network. GO Term annotations derived from UniProt-GOA database. Unannotated genes are not listed.

Gene Symbol	GO Terms	Cluster	Degree Rank	Stress Rank	Radiality Rank
HS25P	Response to high light intensity, chloroplast	2	1	4	5
MD37C	Ubiquitin protein ligase binding, protein ubiquitination, response to high light intensity	2	3	2	1
RPT3	Signal transducer activity, phototropism, protein ubiquitination	2	12	22	13
CLPB1	ATP binding, protein unfolding, response to high light intensity	2	15	14	25
Y4791	Transferase activity, GPI anchor biosynthetic process	2	36	1	14
CIA2	Protein targeting to chloroplast, regulation of transcription	2	<50	12	7
PUB44	Ubiquitin-protein transferase activity, leaf senescence, regulation of chlorophyll catabolic process	4	1	7	12
MOS2	RNA binding, defense response signaling pathway	4	4	25	35
FIMB5	Actin filament binding, actin cytoskeleton organization	4	5	14	6
RDO2	Translation elongation factor activity, negative regulation of flower development, response to gibberlin, seed germination	4	6	22	38
NERD	Histone binding, gene silencing by RNA, regulation of chromatin silencing by small RNA	4	9	38	46
MCCB	Cobalt ion binding, methylcrotonoyl-CoA carboxylase activity, leucine catabolic process	4	<50	2	2
AI5L5	Transcription factor activity, abscisic acid-activated signaling pathway	4	35	6	3
MIP1	Ubiquitin-protein transferase activity, regulation of signal transduction	4	11	9	8
ACR9	Amino acid binding, metabolic process	4	<50	21	9

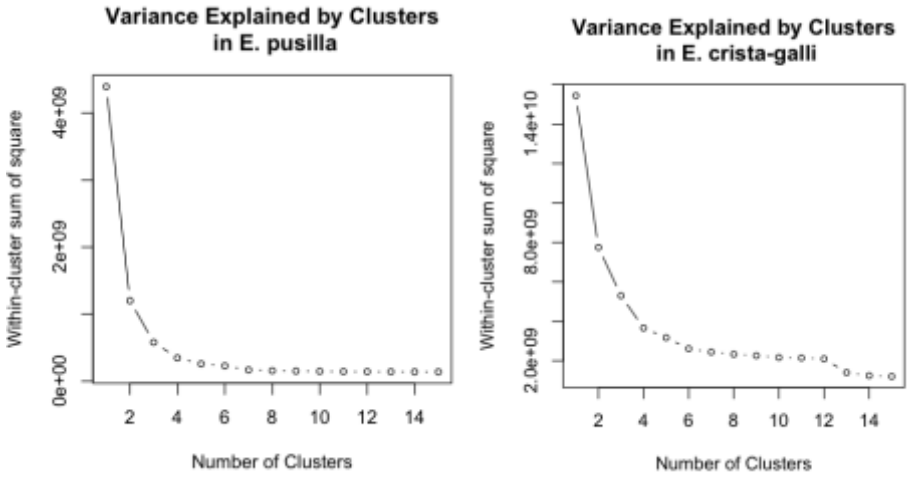


Figure S3.1 – Within-group variance against number of clusters in *E. pusilla* and *E. crista-galli*.

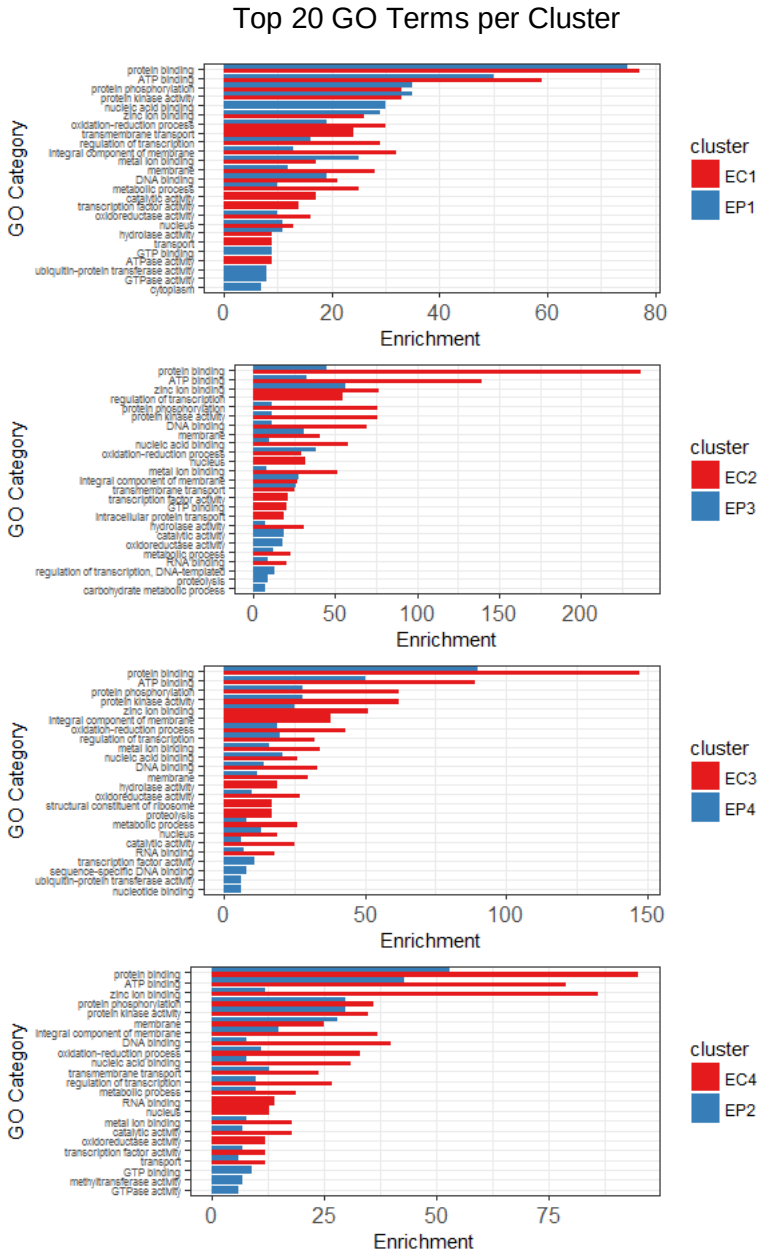


Figure S3.2 – Top 20 GO terms enriched in each gene cluster by species.