nature plants

ARTICLES

PUBLISHED: 21 NOVEMBER 2016 | ARTICLE NUMBER: 16178 | DOI: 10.1038/NPLANTS.2016.178

Transcript, protein and metabolite temporal dynamics in the CAM plant *Agave*

Paul E. Abraham[1], Hengfu Yin[2], Anne M. Borland[2,3], Deborah Weighill[2,4], Sung Don Lim[5], Henrique Cestari De Paoli[2], Nancy Engle[2], Piet C. Jones[2,4], Ryan Agh[2], David J. Weston[2], Stan D. Wullschleger[6], Timothy Tschaplinski[2], Daniel Jacobson[2,4], John C. Cushman[5], Robert L. Hettich[1], Gerald A. Tuskan[2] and Xiaohan Yang[2]*

Already a proven mechanism for drought resilience, crassulacean acid metabolism (CAM) is a specialized type of photosynthesis that maximizes water-use efficiency by means of an inverse (compared to C_3 and C_4 photosynthesis) day/night pattern of stomatal closure/opening to shift CO_2 uptake to the night, when evapotranspiration rates are low. A systems-level understanding of temporal molecular and metabolic controls is needed to define the cellular behaviour underpinning CAM. Here, we report high-resolution temporal behaviours of transcript, protein and metabolite abundances across a CAM diel cycle and, where applicable, compare the observations to the well-established C_3 model plant *Arabidopsis*. A mechanistic finding that emerged is that CAM operates with a diel redox poise that is shifted relative to that in *Arabidopsis*. Moreover, we identify widespread rescheduled expression of genes associated with signal transduction mechanisms that regulate stomatal opening/closing. Controlled production and degradation of transcripts and proteins represents a timing mechanism by which to regulate cellular function, yet knowledge of how this molecular timekeeping regulates CAM is unknown. Here, we provide new insights into complex post-transcriptional and -translational hierarchies that govern CAM in *Agave*. These data sets provide a resource to inform efforts to engineer more efficient CAM traits into economically valuable C_3 crops.

The water-use efficiency of *Agave* spp. hinges on crassulacean acid metabolism (CAM), a specialized mode of photosynthesis that evolved from ancestral C_3 photosynthesis in response to water and CO_2 limitation[1] and is found in ~6.5% of higher plants. Whereas C_3 photosynthesis produces a three-carbon (3-C) molecule for carbon fixation during the day, CAM generates a four-carbon organic acid from carbon fixation at night. In CAM, this nocturnal carboxylation reaction is catalysed by phospho*enol*pyruvate carboxylase (PEPC), and the 3-C substrate phospho*enol*pyruvate (PEP) is supplied by the glycolytic breakdown of carbohydrate formed during the previous day. The nocturnally accumulated malic acid is stored overnight in a central vacuole, and during the subsequent day malate is decarboxylated to release CO_2 at an elevated concentration for Rubisco in the chloroplast. The diel separation of carboxylases in CAM is accompanied by an inverse (compared with C_3 and C_4 photosynthesis-performing species) day/night pattern of stomatal closure/opening that results in improved water-use efficiency (CO_2 fixed per unit water lost) that is up to sixfold higher than that of C_3 photosynthesis plants and up to threefold higher than that of C_4 photosynthesis plants under comparable conditions[2].

The frequent emergence of CAM from C_3 photosynthesis throughout evolutionary history implies that all of the enzymes required for CAM are homologues of ancestral forms found in C_3 species[1,3]. As such, the CAM pathway has been identified as a target for synthetic biology because it offers the potential to engineer improved water-use efficiency in C_3 crops[4–6]. However, the day/night separation of carboxylation and decarboxylation processes and the inverse night/day opening/closing of stomata that distinguish CAM from C_3 photosynthesis imply that the bioengineering of CAM will require a temporal reprogramming of metabolism in the C_3 host. Therefore, key challenges for CAM biodesign will be to establish how many genes must be reprogrammed in a diel manner to modify the behaviour of C_3 plants to perform CAM and to identify which functional or mechanistic governing principles are shared among the diel transcriptional and translational dynamics of C_3 and CAM.

Generating an integrated functional -omics dataset (transcriptomics, proteomics and metabolomics) for a CAM species is an essential first step for providing global insight into the complete set of genes controlling the metabolic steps of CAM, for revealing genes in co-occurrence networks, and for determining the functional consequences of diel co-regulation of transcription and translation. In the present study, temporal profiles of the transcriptome, proteome and metabolome of CAM-performing leaves from the obligate CAM species *Agave americana* were investigated across a 12 h/12 h light/dark diel cycle. With this experimental design, we sought to: (1) identify temporally defined clusters of co-regulated genes, (2) define diel shifts in gene expression between CAM- and C_3-specific gene networks and (3) describe the temporal dynamics between gene expression profiles and protein abundance profiles across the 24 hour light/dark CAM cycle.

Results

Metabolic reprogramming in CAM-performing leaves. CAM plants are classified according to the amount of atmospheric CO_2

[1]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. [3]School of Biology, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK. [4]The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, Tennessee 37996, USA. [5]Department of Biochemistry and Molecular Biology, University of Nevada, MS330, Reno, Nevada 89557-0330, USA. [6]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. *e-mail: yangx@ornl.gov

that is taken up at night[1,7,8]. For *A. americana*, the magnitude of nocturnal net CO_2 uptake changes according to leaf age, with a progressive shift from predominantly C_3 photosynthesis in the youngest leaf to increasing CAM activity with leaf age (Fig. 1a; Supplementary Table 1). We limited all sampling for metabolites, transcriptome and proteome to the fourth fully expanded leaf, which is a mature CAM-performing leaf under the controlled environmental conditions used here (day/night temperature 25/15 °C; 12 hour photoperiod, photon flux density 540 μmol m^{-2} s^{-1} at plant height).

We inspected gas chromatography mass spectrometry (MS) profiles of 64 abundant metabolites in the *Agave* diel cycle (Fig. 1b; Supplementary Table 2), and then compared a subset of these to those in an *Arabidopsis* C_3 leaf (Supplementary Table 3; Fig. 1c). Unlike C_3 leaves, in which malic and fumaric acid levels increase during the day and decrease during the night[9,10], *Agave* leaves accumulate malic acid at night, a defining feature of the nocturnal CO_2 fixation that occurs during CAM. *Agave* leaves also accumulated fumaric acid during the night, which is consistent with the relatively high night-time fluxes of carbon passing through the tricarboxylic acid (TCA) cycle and the carbon exchange between malate and fumarate[11] reported for CAM plants. The diel abundance profile of sucrose, which is reciprocal to that of malic acid, provides support for the hypothesis that *Agave* uses soluble sugars, mainly fructans, oligofructans, fructose, glucose and sucrose, as potential carbon sources for nocturnal malate synthesis[12–15] (Fig. 1c). Among the antioxidants found in plants, ascorbate is usually the most abundant[16] and accumulates to high concentrations in the chloroplast and vacuole following high-light stress. The levels of ascorbic acid, which is involved in metabolic crosstalk between redox related pathways, is high in *Agave* leaves (7–1,100 μg g^{-1} fresh weight (FW) sorbitol equivalents) (Fig. 1c). Interestingly, the diel pattern of daytime depletion and nocturnal accumulation of ascorbic acid in *Agave* contrasts markedly with that in *Arabidopsis* and other C_3 species[17,18].

The reprogramming of the day/night pattern of ascorbic acid turnover in *Agave* is intriguing if ascorbic acid is a key component of a redox hub that integrates metabolic information and environmental stimuli to tune responses within the cellular signalling network[19]. Recent studies have shown that many organisms, including *Arabidopsis*, have a redox rhythm that is dictated by circadian clock components and metabolic activities such as the production and scavenging of reactive oxygen species (ROS)[20–22]. The concept that redox regulation links CO_2 assimilation and related photosynthetic processes to light was established more than two decades ago. Thus, we examined the diel redox status of nicotinamide adenine dinucleotide phosphate (NADP) in *Agave* and compared it with that in *Arabidopsis* leaves that were measured under comparable environmental conditions (Fig. 1d). In *Arabidopsis*, the abundances of the coenzymes NADPH and $NADP^+$ increased in the first few hours of the photoperiod concomitantly with photosynthetic activity. However, the diel abundance patterns of these coenzymes differed in *Agave* leaves, with $NADP^+$ declining in abundance during the day but increasing overnight. In *Agave*, NAD(P)H abundance peaked around 8 h into the night, then declined over the remaining dark period as the abundance of $NADP^+$ increased. The observed pattern of NADPH turnover in *Agave* is consistent with a network scale model of the diel CAM cycle that predicts partitioning of carbohydrate into the oxidative pentose phosphate pathway (OPP) at night to produce NADPH for maintenance processes[23] (Supplementary Note 4). The contrasting diel patterns of abundance for $NADP^+$ and NADPH in *Arabidopsis* and *Agave* indicate a diel shift in the supply of and demand for reductant between C_3 and CAM. This altered diel redox poise indicates a fundamental difference between C_3 and CAM in the relative day/night fluxes through a range of central metabolic processes that include glycolysis, TCA cycle, OPP, nitrogen assimilation and respiratory electron transport.

Temporal dynamics of gene expression across a CAM diel series. Using the same leaf material as sampled for the metabolite profiles, RNA sequencing (RNA-Seq) was performed across eight time points at 3 hour intervals in biological triplicates. RNA-Seq-derived transcript profiles were obtained, and the total abundance of each transcript was assessed after normalizing the number of reads per kilobase and normalizing per million reads (RPKM). In total, 47,499 transcripts were observed in mature leaves of *Agave*. For quantitative analyses, an empirically derived threshold (maximum RPKM $\geq$ 5.02 and minimum average RPKM $\geq$ 3.483; Supplementary Note 5) was applied to remove low-abundance transcripts that had large variance across the entire transcriptomic data set[24] resulting in 37,808 transcripts (Supplementary Table 4). Examination of the data revealed that 82% (31,126) of transcripts were expressed throughout the entire 24 hour period. Pearson correlations were computed and high reproducibility was found at an average Pearson correlation coefficient of 0.91.

On the basis of paired *t*-tests, the expression patterns of 21,168 transcripts that showed at least a twofold change from their mean value with $P < 0.05$ between one or more time points across the diel cycle (Supplementary Table 5) were grouped into nine major clusters based on similarity of expression patterns identified using the *k*-means algorithm implemented in the MeV software package[25] (Supplementary Table 6). Figure 2 shows co-expression patterns across a 24 hour period, with thousands of genes showing oscillating patterns or acute, rapid changes. Interestingly, across many clusters, significantly high (cluster 3) or low (clusters 5, 6, 7) transcript abundance occurs during the middle of the night when nocturnal CO_2 fixation is at its highest, which might highlight a major metabolic transition. As Supplementary Information, we have identified over-represented gene ontology biological processes (GOBP) for each cluster (Supplementary Note 8 and Supplementary Table 7). In Fig. 2, we highlight the five most over-represented GOBP categories for each cluster.

Phase relationships of gene expression between CAM and C_3. A key challenge for CAM biodesign will be to establish the number of genes that need to be reprogrammed to modify the behaviour of C_3 plants to perform CAM. Gene expression is remarkably flexible and constantly reconfigures to respond and adapt to perturbations, and plants have evolved a scheduling mechanism to coordinate and synchronize biological processes during the day/night cycle.

To provide insight into the required degree of reorganization of diel gene expression, we compared the global gene expression profiles of the *Arabidopsis* C_3 leaf and *Agave* CAM-performing leaf during a day/night cycle. We leveraged a tractable and widely used diel gene expression dataset from the *Arabidopsis* community, which was sampled under similar environmental growth conditions to those used for *Agave*, except at 4 hour intervals[26]. To account for the different sampling intervals and numbers, the 4 hour intervals were adjusted to 3 hour intervals using cubic spline interpolation (Supplementary Fig. 2 and Supplementary Table 8). We used the reciprocal best Basic Local Alignment Search Tool (BLAST) hit to identify orthologues based on sequence similarity and then computed Pearson correlations to characterize the temporal relationships of their expression. Among genes with Pearson correlation coefficients $>|0.6|$, we identified 584 genes that had similar expression profiles and 641 genes that had opposite time-of-day expression patterns (Supplementary Table 9). From the combined set of over 1,000 *Agave* and *Arabidopsis* gene profiles, *k*-means clustering generated four clusters that capture the general relationships among orthologues (Fig. 3a), highlighting the established diel rhythms that are either in phase or reciprocal to one another.

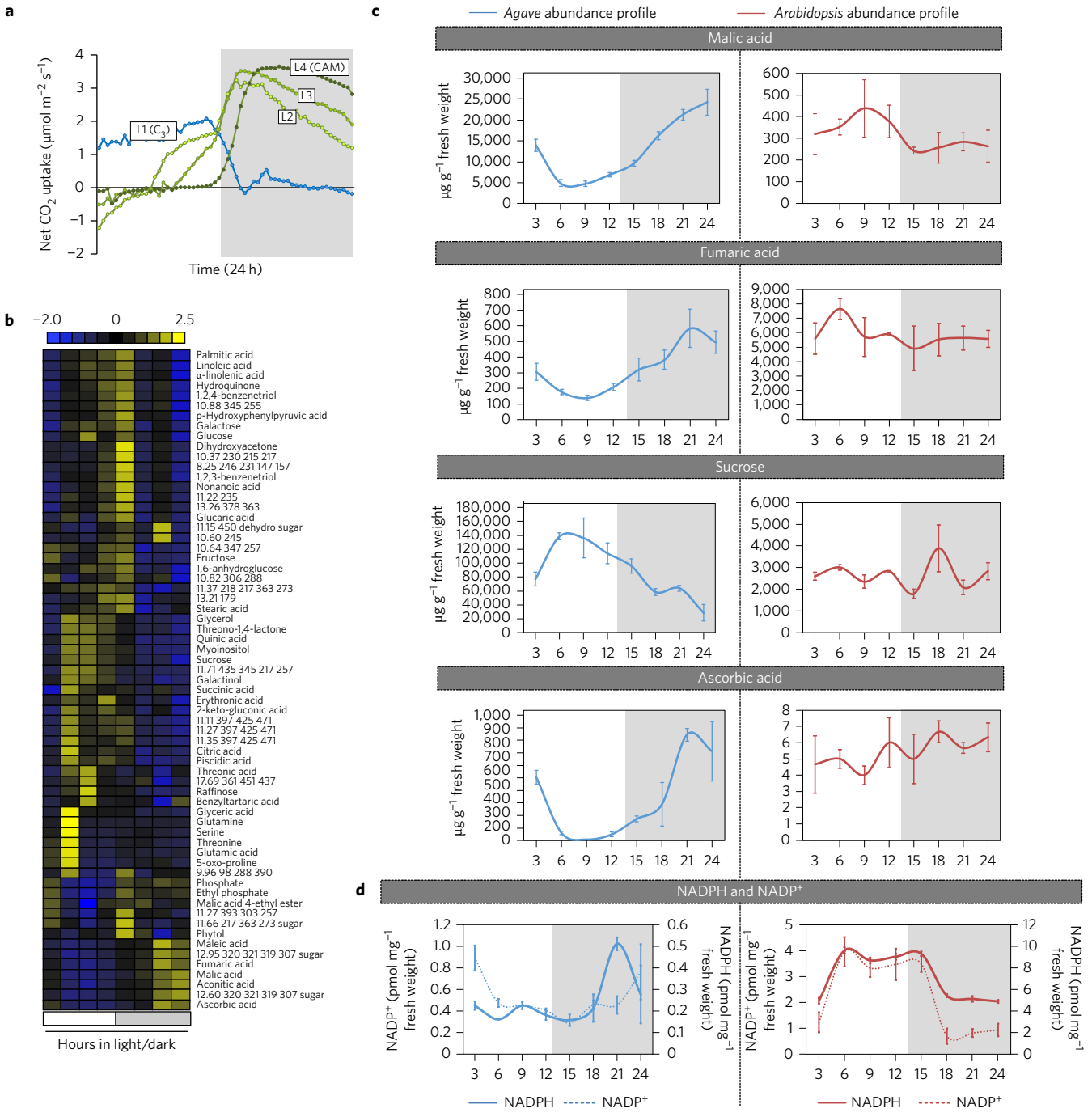


Figure 1 | CAM exhibits rescheduled central metabolism and redox homeostasis relative to C_3. a, The rate of net CO_2 uptake was measured over a light/dark cycle across four leaf developmental states in *Agave*. The leaf performing the greatest degree of CAM (L4) was selected for all experiments. **b**, A hierarchical clustering (Fast Ward method) heatmap of the standard *z*-scores ((abundance – mean)/s.d.) highlights quantified *Agave* metabolites. **c**, Abundance profiles (sorbitol equivalents) for select metabolites between *Agave* (blue) and *Arabidopsis* (red). Both *Agave* and *Arabidopsis* metabolite measurements were taken from tissue collected at diel times 3, 6, 9, 12, 15, 18, 21 and 24 h from the beginning of the light period. Error bars represent standard error for three biological replicates. **d**, Abundance of NADPH (solid line) and $NADP^+$ (dashed line) cofactors for *Agave* (blue) and *Arabidopsis* (red). Both *Agave* and *Arabidopsis* NADPH and $NADP^+$ measurements were taken from tissue collected at diel times 3, 6, 9, 12, 15, 18, 21 and 24 h from the beginning of the light period. Error bars represent the standard deviation for two *Agave* biological replicates and three *Arabidopsis* biological replicates.

Importantly, the clusters with altered diel expression (clusters 1 and 3) include, yet were not limited to, several *Arabidopsis* genes related to redox poise that further corroborate the altered diel redox poise in CAM plants (Supplementary Table 9).

Inverse stomatal behaviour in CAM plants presents an attractive perspective from which to study guard cell signalling because CAM-performing leaves differ in the timing of perception and response to physiological signals related to stomata opening/closing. Whether there are any time-of-day redundancies in major signalling components between C_3 and CAM plants for regulation of stomatal behaviour is still unclear. Therefore, we compared the expression profiles of *Arabidopsis* genes previously associated with light or CO_2 responses to their reciprocal BLAST hits in *Agave* (Fig. 3b,c).

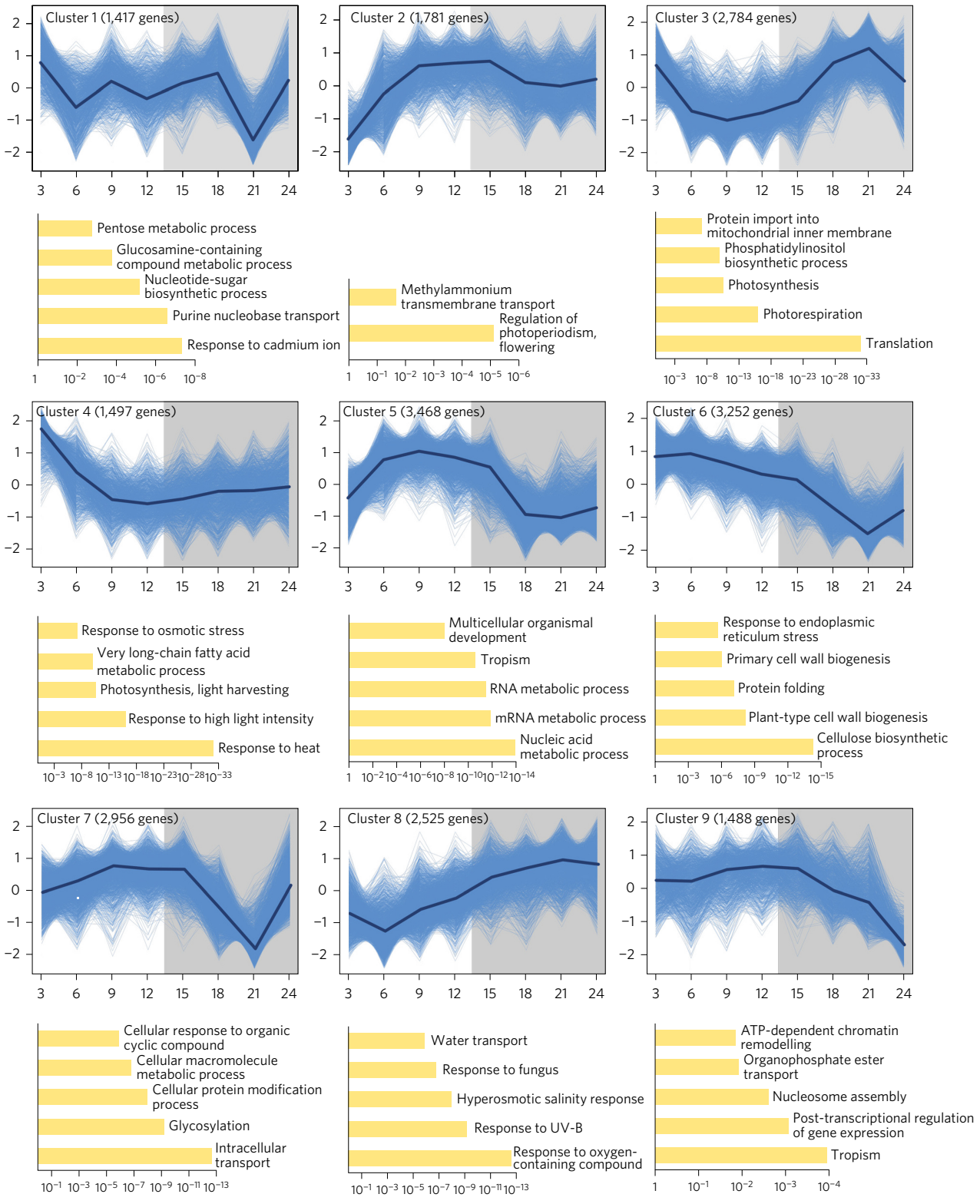


Figure 2 | Temporal changes in *Agave* gene expression across the diel cycle. *Agave* RNA-Seq measurements were taken from tissue collected at diel times 3, 6, 9, 12, 15, 18, 21 and 24 h from the beginning of the light period. Clustering of *k*-means using Pearson's correlation grouped genes into nine clusters based on the similarity of their abundance profiles. The *y*-axis represents the standard *z*-score for each gene ((expression – mean)/s.d.) and highlights prominent patterns of abundance across the diel cycle. For each cluster, each blue line represents an individual gene, the median pattern of expression is represented by a dark blue line and the number of genes belonging to each cluster is reported. Below each cluster, up to the top five most significant gene ontology groups are graphically represented according to their adjusted *P*-values.

Various stimuli can lead to stomatal closure, but stomatal opening is predominately evoked by means of wavelength-responsive mechanisms[27]. The regulatory mechanism of stomatal opening by blue light has been well studied in a number of C_3 plants[28,29], but its role is less clear in CAM plants[30]. Previous work in facultative CAM plants suggests that a blue light receptor mediates a

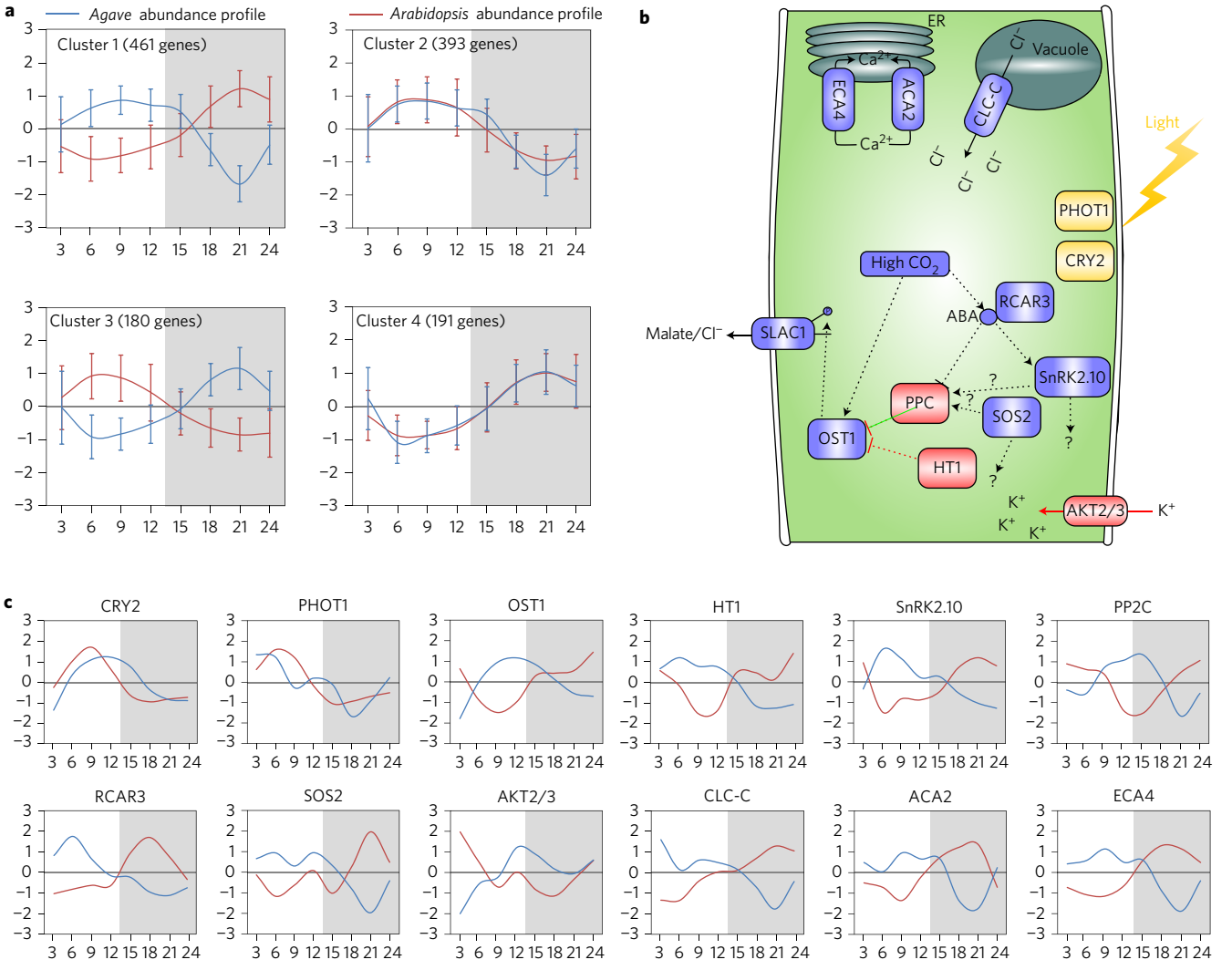


Figure 3 | Diel gene expression and the rescheduling of stomatal movement-related genes in *Agave* compared with *Arabidopsis*. a, *Agave* and *Arabidopsis* reciprocal best BLAST hit orthologues with Pearson correlation coefficients >|0.6| were clustered together using *k*-means to characterize the temporal relationship of their expression. For each cluster, the median pattern of expression is represented for *Agave* (blue) and *Arabidopsis* (red). The *y*-axis represents the standard *z*-score for each cluster profile ((median expression – mean)/s.d.) and highlights prominent correlative and anti-correlative relationships. The number of genes belonging to each cluster is reported and error bars represent the standard deviation of their expression. **b**,**c**, A subset of genes implicated in stomatal movement are illustrated (**b**) and the standard *z*-score of each gene ((expression – mean)/s.d.) for *Agave* (blue) and *Arabidopsis* (red) is shown (**c**). SLAC1, slow-anion channel-associated 1.

light-induced switch from C_3 to CAM in *Clusia minor*[31]. In the present study, we observed a light-induced gene expression profile for a blue/UV-A light-absorbing *cryptochrome 2* (*CRY2*) (Aam348626) that has been implicated in inhibition of hypocotyl elongation, regulation of flowering time and entrainment of the circadian clock[32]. Furthermore, similar expression was observed for the blue/UVA light-induced *photoreceptor 1* (*PHOT1*) (Aam086385) that has been implicated in mediating stomatal opening in response to light[32]. Because these light receptor genes have similar expression patterns in *Arabidopsis* and *Agave*, these data suggest that these particular genes may not be involved in stomatal opening in a constitutive CAM plant, for which other photoreceptors or cues, such as low CO_2, could be the predominant signal.

The perception of CO_2 by guard cells serves as a physiological signal regulating stomatal activity: stomata open at low CO_2 concentrations and close at high CO_2 concentrations in conjunction with abscisic acid (ABA) and the presence of ABA receptors[33]. Previously implicated as a central regulator of stomatal CO_2 signalling, *high leaf temperature 1* (*HT1*) negatively regulates high CO_2-induced stomatal closing. Consequently, *Arabidopsis* plants lacking HT1 activity show a constitutive high CO_2 stomatal response and do not open stomata in response to low CO_2[34]. Interestingly, expression of the *HT1* (Aam018566) gene in *Agave* was rescheduled relative to that in *Arabidopsis* (Fig. 3c). CO_2 and ABA-induced perception and signalling are interdependent and *open stomata 1* (*OST*)/*SNF-related protein kinase 2.6* (Aam349853), which is a downstream target of HT1[35] and a convergence point for ABA and CO_2 signalling pathways, also exhibited rescheduled expression in *Agave* compared to *Arabidopsis* (Fig. 3c). Two other sucrose non-fermenting (SNF)-related kinases, *salt overly sensitive 2* (*SOS2*) (Aam080324) and *SnRK2.10* (Aam332354), exhibited shifted expression patterns in *Agave* (Fig. 3c). Several classes of serine/threonine phosphatases (PP1A, PP2A, PP2B and PP2C) can all regulate aspects of guard cell signalling. PP2C protein phosphatases, in particular, contribute to the ABA perception complex with PYR1/PYL1/RCAR by inhibiting SNF-related protein kinases, such as OST1[36,37]. Transcript abundances of *PP2C family protein* (Aam012848) as well as the regulatory component of an ABA

receptor *(RCAR3)* (Aam022092) exhibited temporal shifts in abundance compared to that in *Arabidopsis*.

The opening and closing of stomata is driven by turgor and volume changes in guard cells surrounding the stomatal pores[38]. The osmotic uptake of water driven by the accumulation of ions and sugars causes the stomata to open or close. Therefore, the varying activities of different ion channels and their fluctuating spatiotemporal patterns contribute to the regulation of stomatal apertures. Several different sources of ion flux show shifted temporal profiles in *Agave* compared to *Arabidopsis* (Fig. 3c). Investigations of the osmotic changes driving guard cell behaviour have mainly focused on the role of K^+ transport across the plasma membrane of guard cells, which is a major contributor to stomatal opening and closing. The activity of inward-rectifying channels in guard cells induces swelling (opening) or shrinking (closing) of the guard cells surrounding the stomatal apertures[39]. The inward-rectifying Ca^{2+}-sensitive K^+ channels are thought to serve as a major pathway for K^+ migration into guard cells during stomatal opening[39]. The transcript abundance of the *potassium transporter 2/3* (*AKT2/3*) (Aam018832), which controls Ca^{2+}-sensitive uptake of K^+ by guard cells, showed an expression pattern reciprocal to that in *Arabidopsis*. Two endoplasmic reticulum (ER) Ca^{2+} ATPase transcripts, including *calcium ATPase 2* (*ACA2*) (Aam003442) and *endomembrane-type CA-ATPase 4* (*ECA4*) (Aam088048), also exhibited shifts in temporal expression patterns relative to those in *Arabidopsis*. These ATPases might serve as part of a tuning mechanism to regulate the magnitude or duration of a calcium flux[40]. Essential to stomatal activity, the K^+ flux within guard cells must be counterbalanced by fluxes of anions, such as Cl^-. Thus, it was particularly interesting that a member of the chloride channel family (CLC-c) (Aam081659) that is localized to the vacuole and highly expressed in guard cells in *Arabidopsis*[41] showed reciprocal expression behaviour in *Agave* compared with that in *Arabidopsis*. Collectively, the expression patterns we observed provide substantial evidence for the temporal reprogramming of particular genes essential to regulation of stomatal behaviour in an obligate CAM plant.

Detection of candidate regulators of reprogrammed metabolism. Given the observed rescheduling of gene expression in *Agave*, a comparative co-expression analysis has great potential for characterizing the evolution of biological pathways between well-studied *Arabidopsis* and relatively uncharacterized *Agave*. Because transcription factors are part of a prime mechanism that orchestrates specific control over the time of day during which biological processes operate, transcription factors that show reprogrammed expression in *Agave* relative to *Arabidopsis* could help unravel novel differences in transcriptional regulatory control between C_3 and CAM. Therefore, we sought to identify transcription factors with reciprocal expression profiles in *Agave* and *Arabidopsis*.

To predict transcription factor regulatory interactions and identify new candidate genes for CAM biodesign efforts, integrated analysis of CAM and C_3 transcriptomics data was performed by generating cross-taxa co-expression network modules (Supplementary Note 5). We defined the list of candidate regulators via the inverse pattern of their transcript expression in *Agave* relative to that in *Arabidopsis*. We also enforced strict criteria to define their target genes by the relationship of their expression and function to that of targets predicted in *Arabidopsis* (Supplementary Fig. 4). Using this approach, we identified auxin response factor 4 (ARF4) as a candidate transcription factor that could regulate inverse gene expression in *Agave* compared with *Arabidopsis* as well as several candidate target genes containing auxin response elements (AuxREs) (Supplementary Note 5). Although experimental validation is needed, this result will enable future studies into the connections between CAM regulatory mechanisms and adaptation to the environment.

Protein abundances across a CAM diel series. RNA-Seq data provide insight into gene expression, but protein abundances better reflect the functional state of a cell at a given point in time. Therefore, protein was extracted from the same *Agave* tissue from which the metabolomic and transcriptomic profiles were generated. Tryptic peptides generated from each sample were measured by two-dimensional liquid chromatography nano-electrospray tandem mass spectrometry and yielded 32,561 non-redundant distinct peptide sequences that mapped to 14,207 *A. americana* protein accessions (~20% of total predicted *Agave* protein sequences) across the entire data set (Supplementary Table 12). The data revealed that >90% of these proteins were observed throughout the entire 24 hour period. Pearson correlations show high biological reproducibility with an average correlation coefficient of 0.90. Given the incompleteness of the data (fragmented gene models) and the protein inference problem (shared peptides), we grouped proteins with 90% sequence homology to more accurately report identifications. When considering only protein groups that were uniquely identified, a total of 6,714 protein groups representing 11,337 protein accessions were observed. From this subset, total abundances of proteins were assessed by adding peptide intensities (spectral counts) obtained in the MS analysis and using the normalized spectral abundance factors (NSAF)[42]. For quantitative analysis, an empirically derived threshold (maximum NSAF > 1.5 and minimum average NSAF > 1) was used to remove low-abundance proteins with large variances across the entire proteomic data set, resulting in 4,710 protein accessions (2,434 protein groups) (Supplementary Table 13). On the basis of paired *t*-tests, the abundance patterns of 2,002 proteins (1,226 protein groups) showed at least a twofold change from their mean value with $P < 0.05$ between one or more time points across the diel cycle (Supplementary Table 14). These proteins were grouped across six major clusters based on similarity of expression patterns identified by the *k*-means algorithm implemented in the MeV software package[43] (Supplementary Table 15). Figure 4 shows the oscillating patterns or acute, rapid changes in protein abundances across a 24 hour period, similar to those observed in gene expression profiles. To detect functional specialization within the clusters, we tested for over-representation of GOBP terms and show the five most over-represented GOBP categories for each cluster (Fig. 4; Supplementary Table 16).

In addition to inverse stomatal behaviour, another major distinctive feature of CAM is the nocturnal fixation of CO_2 by *phosphoenolpyruvate carboxylase (PEPC)* and subsequent remobilization the following day to release CO_2 for the Calvin–Benson cycle plus pyruvate, which is recycled by gluconeogenesis by means of *pyruvate orthophosphate dikinase* (*PPDK*) (Supplementary Fig. 5 and Supplementary Note 6). Given their importance to CAM, we show that the transcript abundance of *PEPC1* changes substantially across the diel cycle, peaking at the end of the day. More importantly, we show for the first time that PEPC1 protein abundance follows a diel oscillation similar to that of the transcript. The expression of the PPDK transcript and protein were largely coincident with one another, peaking in the morning, which is consistent with a role in the decarboxylation of malic acid during the early morning hours in *Agave*. As anticipated, the protein responsible for downregulating the activity of PPDK, *PPDK-regulatory protein* (*RP1*) (Aam051010), reaches peak abundance at night in *Agave*, yet has an abundance profile reciprocal to that of its transcript.

Variation in temporal dynamics of transcript and protein abundance. The temporally distinct modulation of the transcript and protein abundance profiles has great potential for elucidating gene function and biological pathway regulation by revealing regulatory mechanisms that occur after transcription and beyond. Therefore, we explored the temporal relationships of the expression of each transcript and its

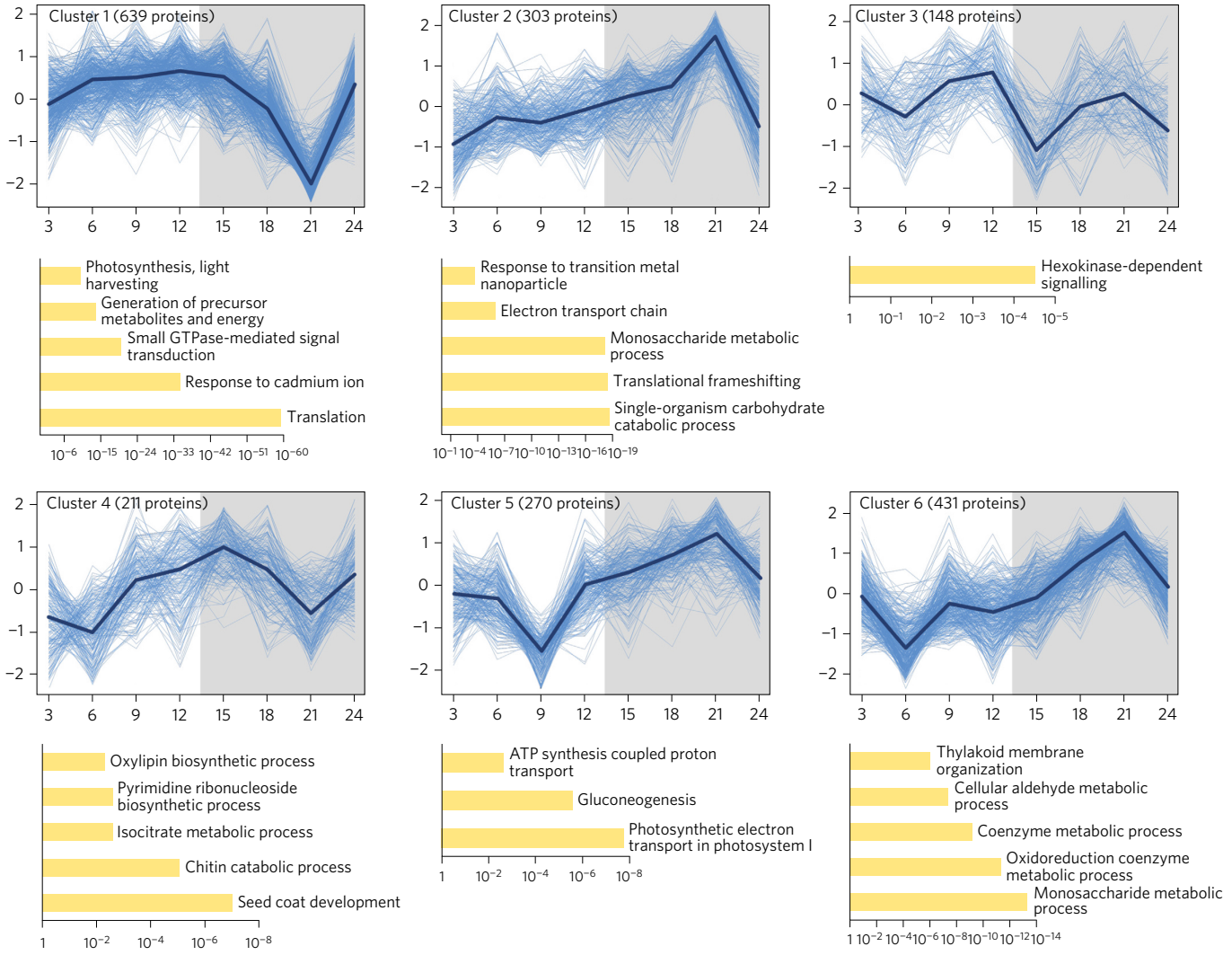


Figure 4 | Temporal changes in protein abundances in *Agave* across the diel cycle. Protein measurements were made on *Agave* leaf samples collected at diel times 3, 6, 9, 12, 15, 18, 21 and 24 h from the beginning of the light period. Clustering of *k*-means using Pearson's correlation grouped genes into six clusters based on their similarity in abundance profiles. The *y*-axis represents the standard *z*-score for each protein ((protein abundance – mean)/s.d.) to highlight prominent patterns of protein abundance across the diel cycle. For each cluster, each blue line represents an individual protein, the median pattern of abundance is represented by a dark blue line and the number of proteins belonging to each cluster is reported. Below each cluster, up to the top five most significant gene ontology groups are graphically represented according to their adjusted *P*-values.

encoded protein in *Agave*. When looking at relationships for a single time point, we observed non-linear relationships with weak correlations (average Pearson's correlation between log-transformed abundances was $r = 0.48$), which is consistent with results of previous studies[44]. The temporal dynamics between the expression of transcripts and their encoded proteins could explain this variation, which results from various rates of biosynthesis and degradation and subsequent post-transcriptional or -translational modifications. For this reason, rational bioengineering design efforts must consider the temporal relationship between the expression of a transcript and its subsequent protein product not only in native CAM species, but also in the engineered C_3 target species for particular combinations of promoter and expressed protein.

Transcript and protein abundances that occur in phase represent transcription directly linked to translation with either very little or rapid regulation. Transcript and protein abundance patterns that are out of phase, on the other hand, are likely regulated at or beyond the post-transcriptional level. To calculate the cross-correlations between transcript and protein abundance profiles, we implemented a cross-correlation function to estimate delays (0 and time delay ±0–7) between transcript and protein signals and then rearranged the data according to lags to calculate a correlation coefficient and *P*-value. Only transcripts and proteins observed at every time point and only relationships involving non-ambiguous, uniquely identified proteins exhibiting a significant change in abundance were used for this analysis. A correlation coefficient cutoff of > 0.7 and $P < 0.05$ were selected as thresholds to generate a refined subset of 336 transcript–protein relationships. Further manual annotation was used to refine the set of transcript and protein abundance relationships while considering variation across replicates. In total, 254 transcript–protein relationships were retained and their curated transcript-to-protein time delays are reported in Supplementary Table 17. We were thus able to exploit high-resolution sampling of transcript and protein abundances to evaluate the temporal dynamics of several key processes related to photosynthesis and respiration (Supplementary Fig. 7 and Supplementary Note 6).

Discussion

Metabolic profiling not only corroborated previous findings for CAM, for example, oscillations in organic acid concentrations, but

also provided novel insights into diel variations in other identifiable metabolites, which now serve as a rich data set to facilitate future investigations into CAM. By comparing *Agave* and *Arabidopsis* leaves under comparable growth conditions, we were able to examine rescheduled components of C_3- and CAM-specific gene expression controlling other processes. Comparison of sucrose abundance at different times of day lends further support to the premise that this carbon source is broken down in *Agave* at the end of the light period to release glucose and fructose, which supply the PEP for nocturnal carbon fixation, as in other *Agave* species[45]. Overall, the nocturnal increases in malic acid, fumaric acid and $NADP^+$ in *Agave* are consistent with the reportedly high mitochondrial fluxes of carbon and electron transport that occur in CAM plants at night. The elevated levels of ascorbic acid that accumulate during night-time in *Agave* are consistent with the need for anti-oxidant activity to deal with reactive oxygen species generated by high rates of respiratory electron transport that occur at night in CAM plants.

The diel patterns of gene expression in plants are likely to be meaningful indicators of the innate relative timing of different cellular and metabolic processes, particularly the manners in which gene expression is affected by environmental and endogenous signals. Interestingly, we observed significant increases and decreases in gene expression at a period during the night coinciding with maximum net CO_2 uptake. The relative contributions of unknown external or internal regulatory inputs during this period remain to be determined. These high-resolution transcriptional profiles will certainly contribute to our understanding of the diel regulation of gene expression in CAM, but here we instead focused our analysis on the similarities and differences between the temporally regulated transcriptomes of CAM and C_3 leaves.

We examined convergent and divergent timing of gene expression systems that reflect the adapted physiology of CAM species relative to C_3 plant species. Interestingly, we did not observe temporal differences in the expression of blue-light-responsive genes identified in *Agave*. Instead, we identified diel variation between CAM and C_3 orthologues implicated in CO_2- and ABA-related signalling events. We propose these rescheduled genes are among the key components of the core signalling mechanism responsible for inverse stomatal activity in CAM plants. Moreover, the expression of the transcripts for many of the genes discussed here, including *SOS2*, have also shown distinct temporal changes in response to salinity or oxidative stress in *Arabidopsis*[46] and might be appropriate candidates for improving stress tolerance or water-use efficiency as part of CAM biodesign research efforts.

We also describe the generation of the first large-scale proteomic profile for CAM-performing leaves to identify the temporal protein abundance profiles underpinning CAM. Across the diel cycle, we observed significant changes in protein abundance similar to patterns of changes observed in transcript profiles. Interestingly, some proteins exhibited substantial abundance changes during the middle of the dark period, coinciding with nocturnal CO_2 fixation and the increased abundance of the coenzyme NADPH. In addition to detailing the temporal dynamics of over-represented GOBP processes, we have illustrated protein abundance patterns for many key metabolic processes pertinent to CAM. By comparing diel patterns of transcript and protein abundance in *Agave*, we have revealed new insights that will help facilitate rational design to enhance water-use efficiency and improve drought tolerance of C_3 crops through a better understanding of the complex regulatory processes that govern the operation of CAM.

Methods

Plant materials. *Agave americana* 'Marginata' plants were obtained from Notestein's Nursery, Gainesville, FL (http://southerngardening.org). *Arabidopsis thaliana* (Col-0) seeds were obtained from TAIR (http://www.arabidopsis.org/). The *A. americana* and *A. thaliana* plants were grown in controlled environments (Supplementary Note 1).

Measurement of leaf gas exchange. Net CO_2 uptake in *A. americana* 'Marginata' was measured using a compact mini cuvette system in a Central Unit CMS-400 with BINOS-100 infrared gas analyser working in an open format (Heinz Walz GmbH) (Supplementary Note 2).

Metabolite profiling by gas chromatography mass spectrometry. For *A. americana* metabolite identification, 8 samples were collected with three biological replicates of mature leaf samples (fourth fully expanded leaf) collected at 3, 6, 9, 12, 15, 18, 21 and 24 h after the starting of the light period. Samples were frozen in liquid nitrogen and ground using a mortar and pestle and stored at −80 °C until metabolite profiling. For *A. thaliana* (Col-0), eight samples were collected with three biological replicates of fully-expanded leaf samples collected at 3, 6, 9, 12, 15, 18, 21 and 24 h after the starting of the light period. Samples were frozen in liquid nitrogen and ground using a mortar and pestle and stored at −80 °C until metabolite profiling (Supplementary Note 3).

NADPH and $NADP^+$ measurement. Fully expanded leaves of *A. americana* and *A. thaliana* (Col-0) were collected for enzymatic assays to determine total NADP and calculated NADPH from decomposed NADP using the NADP/NADPH Quantification Kit (BioVision) according to the manufacturer's instructions (Supplementary Note 4).

Chloroplast genome sequencing, assembly and annotation. Chloroplasts were isolated from the *A. americana* leaf tissue using a Chloroplast Isolation Kit (Sigma, Cat CP-ISO). DNA was extracted from enriched chloroplasts using the DNeasy DNA Extraction Kit (QIAGEN, Cat No. 69104). Paired-end sequencing libraries with an average insert size of 500 bp were constructed from the chloroplast DNA using an Illumina TruSeq DNA Sample Prep Kit v2 and sequenced on a MiSeq instrument using the MiSeq Reagent Kit v3 (600 cycle). Paired end reads (2×300 bp) were trimmed using Trimmomatic[47] with settings of MINLEN = 100 and SLIDINGWINDOW = 4:20. Trimmed overlapping paired end reads (NCBI SRA accession SRP076143) were merged into extended long reads using FLASH[48]. Merged long reads were searched against public chloroplast genome sequences available at NCBI (http://www.ncbi.nlm.nih.gov/) using BLASTN[49] with an e-value cutoff of $\times 10^{-5}$. Merged long reads with BLASTN hits in the NCBI chloroplast database and the un-merged paired end reads were used to create *de novo* genome assemblies using SOAPdenovo version r240[50] with multiple k-mer lengths from 20 to 99. Individual assemblies were merged using CAP3[51] with default settings. For filling the gaps in the genome assembly, two pairs of PCR primers (pair1: 5′-GAATTCGCGCCTACTCTGAC-3′, 5′-GGCCGATTGATCTTCCAATA-3; pair2: 5′-AATCCACTGCCTTGATCCAC-3′, 5′-ATCAACCGTGCTAACCTTGG-3′) were designed based on the *Agave* chloroplast genome sequence. Gap sequences were obtained by sequencing PCR amplified chloroplast DNA fragments using Sanger sequencing on an ABI machine. Chloroplast genome annotation was performed using CpGAVAS[52]. The chloroplast genome assembly and annotation were deposited at GenBank (accession KX519714).

Transcriptomics. For transcriptome sequencing, 15 *A. americana* samples were collected with three biological replicates, including eight samples of the mature fourth fully expanded leaf collected at 3, 6, 9, 12, 15, 18, 21 and 24 h after the beginning of the light period; three young leaf samples collected at diel time points of 6, 12 and 21 h, respectively; and four non-leaf samples (meristem, rhizome, root or stem) collected at 3 h after the beginning of the light period. Samples were frozen in liquid nitrogen, ground using a mortar and pestle and then stored frozen at −80 °C until transcriptomics analysis (Supplementary Note 5).

Proteomics. For proteome sequencing, *A. americana* leaf samples were collected with three biological replicates and included eight samples of the mature fourth fully expanded leaf collected at 3, 6, 9, 12, 15, 18, 21 and 24 h after the beginning of the light period. Samples were frozen in liquid nitrogen, ground using a mortar and pestle and then stored frozen at −80 °C until proteomics analysis (Supplementary Note 6).

Statistical analysis. For this study, we performed pair-wise comparisons of time points as our hypothesis is concerned with the change among different time-points and not the overall change in transcripts and proteins. To this end, we employed two approaches for each dataset to provide a comprehensive assessment of the statistical confidence (Supplementary Note 7 and Supplementary Fig. 8).

Temporal relationship between mRNA and protein expression. As illustrated (Supplementary Fig. 6), cross-correlations between RNA-Seq and proteomic datasets were calculated using the *crosscorr* function implemented in the Econometrics Toolbox (Matlab) to estimate time lags with the sample cross-correlation for each gene in the two datasets. Because both datasets were periodic, we rearranged the data according to lags and calculated the correlation coefficients for each gene. A correlation coefficient cutoff of > 0.7 and $P < 0.05$ were selected as thresholds to ensure a subset of high-quality relationships. Visual inspection of the relative transcript and protein abundances and standard error of the means were then used for further validation.

Data availability. Data that support the findings of this study have been deposited into public repositories. Chloroplast sequence data is deposited at GenBank with the accession code KX519714. The metabolite data is deposited at MetaboLights under the accession code MTBLS363. The transcriptomics data is deposited at GenBank (http://www.ncbi.nlm.nih.gov/genbank/) under the accession code GBHM00000000. The proteomics data has been deposited at MassIVE under the accession code MSV000079780 and ProteomeXchange with the accession code PXD004239.

Received 14 April 2016; accepted 20 October 2016; published 21 November 2016

References

1. Silvera, K. *et al.* Evolution along the crassulacean acid metabolism continuum. *Funct. Plant Biol.* **37,** 995–1010 (2010).
2. Borland, A., Griffiths, H., Hartwell, J. & Smith, J. Exploiting the potential of plants with crassulacean acid metabolism for bioenergy production on marginal lands. *J. Exp. Bot.* **60,** 2879–2896 (2009).
3. West-Eberhard, M. J., Smith, J. A. C. & Winter, K. Photosynthesis, reorganized. *Science* **332,** 311–312 (2011).
4. Borland, A. M. *et al.* Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci.* **30,** 327–338 (2014).
5. Borland, A. M. & Yang, X. Informing the improvement and biodesign of crassulacean acid metabolism via system dynamics modelling. *New Phytol.* **200,** 946–949 (2013).
6. DePaoli, H. C., Borland, A. M., Tuskan, G. A., Cushman, J. C. & Yang, X. Synthetic biology as it relates to CAM photosynthesis: challenges and opportunities. *J. Exp. Bot.* **65,** 3381–3393 (2014).
7. Winter, K., Garcia, M. & Holtum, J. On the nature of facultative and constitutive CAM: environmental and developmental control of CAM expression during early growth of *Clusia*, *Kalanchoë*, and *Opuntia*. *J. Exp. Bot.* **59,** 1829–1840 (2008).
8. Winter, K. & Holtum, J. Facultative crassulacean acid metabolism (CAM) plants: powerful tools for unravelling the functional elements of CAM photosynthesis. *J. Exp. Bot.* **65,** 3425–3441 (2014).
9. Chia, D. W., Yoder, T. J., Reiter, W. D. & Gibson, S. I. Fumaric acid: an overlooked form of fixed carbon in *Arabidopsis* and other plant species. *Planta* **211,** 743–751 (2000).
10. Fahnenstich, H. *et al.* Alteration of organic acid metabolism in *Arabidopsis* overexpressing the maize C(4)NADP-malic enzyme causes accelerated senescence during extended darkness. *Plant Physiol.* **145,** 640–652 (2007).
11. Osmond, C. B. *et al.* Regulation of malic-acid metabolism in crassulacean-acid-metabolism plants in the dark and light: in-vivo evidence from (13)C-labeling patterns after (13)CO 2 fixation. *Planta* **175,** 184–192 (1988).
12. Arrizon, J., Morel, S., Gschaedler, A. & Monsan, P. Comparison of the water-soluble carbohydrate composition and fructan structures of *Agave tequilana* plants of different ages. *Food Chem.* **122,** 123–130 (2010).
13. Mancilla-Margalli, N. & López, M. Water-soluble carbohydrates and fructan structure patterns from *Agave* and *Dasylirion* species. *J. Agric. Food Chem.* **54,** 7832–7839 (2006).
14. Raveh, E., Wang, N. & Nobel, P. Gas exchange and metabolite fluctuations in green and yellow bands of variegated leaves of the monocotyledonous CAM species *Agave americana*. *Physiol. Plant.* **103,** 99–106 (1998).
15. Wang, N. & Nobel, P. Phloem transport of fructans in the crassulacean acid metabolism species *Agave deserti*. *Plant Physiol.* **116,** 709–714 (1998).
16. Horemans, N., Foyer, C., Potters, G. & Asard, H. Ascorbate function and associated transport systems in plants. *Plant Physiol. Biochem.* **38,** 531–540 (2000).
17. Bartoli, C. G. *et al.* Inter-relationships between light and respiration in the control of ascorbic acid synthesis and accumulation in *Arabidopsis thaliana* leaves. *J. Exp. Bot.* **57,** 1621–1631 (2006).
18. Dutilleul, C. *et al.* Leaf mitochondria modulate whole cell redox homeostasis, set antioxidant capacity, and determine stress resistance through altered signaling and diurnal regulation. *Plant Cell* **15,** 1212–1226 (2003).
19. Foyer, C. H. & Noctor, G. Ascorbate and glutathione: the heart of the redox hub. *Plant Physiol.* **155,** 2–18 (2011).
20. Lai, A. *et al.* CIRCADIAN CLOCK-ASSOCIATED 1 regulates ROS homeostasis and oxidative stress responses. *Proc. Natl Acad. Sci. USA* **109,** 17129–17134 (2012).
21. Stangherlin, A. & Reddy, A. Regulation of circadian clocks by redox homeostasis. *J. Biol. Chem.* **288,** 26505–26511 (2013).
22. Zhou, M. *et al.* Redox rhythm reinforces the circadian clock to gate immune response. *Nature* **523,** 472–476 (2015).
23. Cheung, C. Y., Poolman, M. G., Fell, D. A., Ratcliffe, R. G. & Sweetlove, L. J. A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C3 and crassulacean acid metabolism leaves. *Plant Physiol.* **165,** 917–929 (2014).
24. Lochner, A. *et al.* Label-free quantitative proteomics for the extremely thermophilic bacterium *Caldicellulosiruptor obsidiansis* reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. *J. Proteome Res.* **10,** 5302–5314 (2011).
25. Saeed, A. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34,** 374–378 (2003).
26. Mockler, T. C. *et al.* The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harb. Symp. Quant. Biol.* **72,** 353–363 (2007).
27. Zhou, X. *et al.* CYCLIN h;1 regulates drought stress responses and blue light-induced stomatal opening by inhibiting reactive oxygen species accumulation in *Arabidopsis*. *Plant Physiol.* **162,** 1030–1041 (2013).
28. Inada, S., Ohgishi, M., Mayama, T., Okada, K. & Sakai, T. RPT2 is a signal transducer involved in phototropic response and stomatal opening by association with phototropin 1 in *Arabidopsis thaliana*. *Plant Cell* **16,** 887–896 (2004).
29. Tseng, T.-S. & Briggs, W. R. The *Arabidopsis rcn1-1* mutation impairs dephosphorylation of Phot2, resulting in enhanced blue light responses. *Plant Cell* **22,** 392–402 (2010).
30. Ceusters, J. *et al.* Light quality modulates metabolic synchronization over the diel phases of crassulacean acid metabolism. *J. Exp. Bot.* **65,** 3705–3714 (2014).
31. Grams, T. & Thiel, S. High light-induced switch from C3-photosynthesis to crassulacean acid metabolism is mediated by UV-A/blue light. *J. Exp. Bot.* **53,** 1475–1483 (2002).
32. Banerjee, R. & Batschauer, A. Plant blue-light receptors. *Planta* **220,** 498–502 (2005).
33. Chater, C. *et al.* Elevated CO_2-induced responses in stomata require ABA and ABA signaling. *Curr. Biol.* **25,** 2709–2716 (2015).
34. Hashimoto, M. *et al. Arabidopsis* HT1 kinase controls stomatal movements in response to CO_2. *Nat. Cell Biol.* **8,** 391–397 (2006).
35. Tian, W. *et al.* A molecular pathway for CO(2) response in *Arabidopsis* guard cells. *Nat. Commun.* **6,** 6057 (2015).
36. Umezawa, T. *et al.* Type 2C protein phosphatases directly regulate abscisic acid-activated protein kinases in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **106,** 17588–17593 (2009).
37. Xie, T. *et al.* Molecular mechanism for inhibition of a critical component in the *Arabidopsis thaliana* abscisic acid signal transduction pathways, SnRK2.6, by protein phosphatase ABI1. *J. Biol. Chem.* **287,** 794–802 (2012).
38. Azoulay-Shemer, T. *et al.* Guard cell photosynthesis is critical for stomatal turgor production, yet does not directly mediate CO_2-and ABA-induced stomatal closing. *Plant J.* **83,** 567–581 (2015).
39. Wang, Y., Chen, Z., Zhang, B., Hills, A. & Blatt, M. PYR/PYL/RCAR abscisic acid receptors regulate K^+ and Cl^- channels through reactive oxygen species-mediated activation of Ca^{2+} channels at the plasma membrane of intact *Arabidopsis* guard cells. *Plant Physiol.* **163,** 566–577 (2013).
40. Sze, H., Liang, F., Hwang, I., Curran, A. & Harper, J. Diversity and regulation of plant Ca^{2+} pumps: insights from expression in yeast. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **51,** 433–462 (2000).
41. Jossier, M. *et al.* The *Arabidopsis* vacuolar anion transporter, AtCLCc, is involved in the regulation of stomatal movements and contributes to salt tolerance. *Plant J.* **64,** 563–576 (2010).
42. Zybailov, B., Florens, L. & Washburn, M. Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. *Mol. Biosyst.* **3,** 354–360 (2007).
43. Howe, E., Sinha, R., Schlauch, D. & Quackenbush, J. RNA-Seq analysis in MeV. *Bioinformatics* **27,** 3209–3210 (2011).
44. Walley, J. W. *et al.* Reconstruction of protein networks from an atlas of maize seed proteotypes. *Proc. Natl Acad. Sci. USA* **110,** E4808–E4817 (2013).
45. Christopher, J. & Holtum, J. Patterns of carbohydrate partitioning in the leaves of crassulacean acid metabolism species during deacidification. *Plant Physiol.* **112,** 393–399 (1996).
46. Yang, Q. *et al.* Overexpression of *SOS (Salt Overly Sensitive)* genes increases salt tolerance in transgenic *Arabidopsis*. *Mol. Plant* **2,** 22–31 (2009).
47. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
48. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27,** 2957–2963 (2011).
49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
50. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience* **1,** 18 (2012).
51. Huang, X. & Madan, A. CAP3: a DNA sequence assembly program. *Genome Res.* **9,** 868–877 (1999).
52. Liu, C. *et al.* CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. *BMC Genomics* **13,** 715 (2012).

Acknowledgements

This material is based on work supported by the Department of Energy Office of Science Genomic Science Program under award number DE-SC0008834. The authors would like to

thank R. Giannone and M.A. Cushman for critical review and clarifying comments on the manuscript. This research used resources of the Compute and Data Environment for Science (CADES) and the Oak Ridge Leadership Computing Facility (OLCF) at the Oak Ridge National Laboratory. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the US Department of Energy (under contract number DE-AC05-00OR22725).

Author contributions

X.Y., G.A.T., P.E.A. and R.L.H. contributed to conception and design of the experiment; P.A., H.Y., A.M.B., S.D.L., H.C.D.P., N.E., R.A. and T.T. contributed to the acquisition of data; and P.A., H.Y., A.B., D.J.W., P.C.J., D.J., T.T. and J.C.C. contributed to data analysis and interpretation; P.A., X.Y., G.T. and A.B. drafted the manuscript and all authors critically revised and approved the final version of the manuscript for publication.

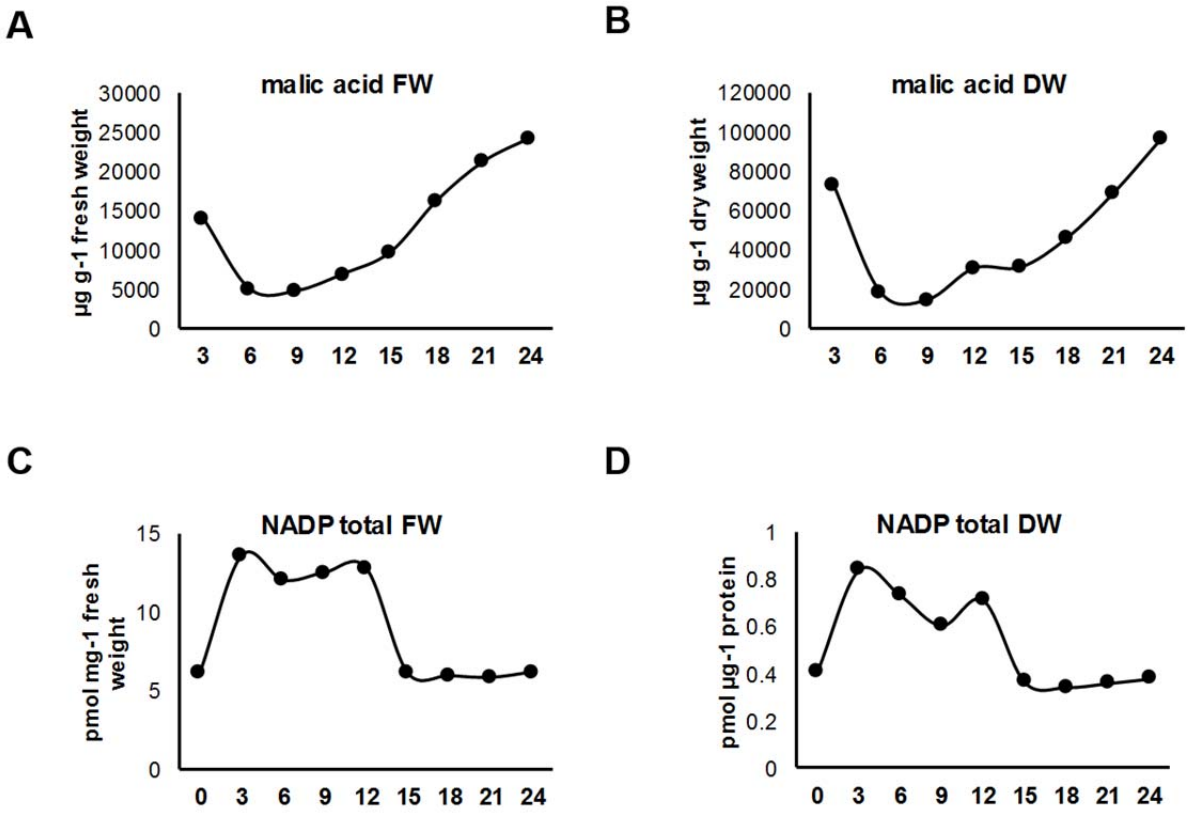
Additional information

Supplementary information is available for this paper. Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to X.Y.

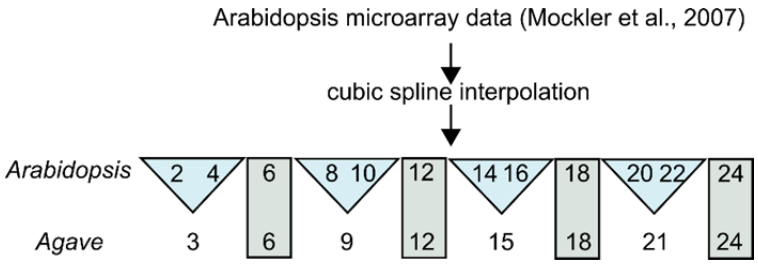
Competing interests

The authors declare no competing financial interests.

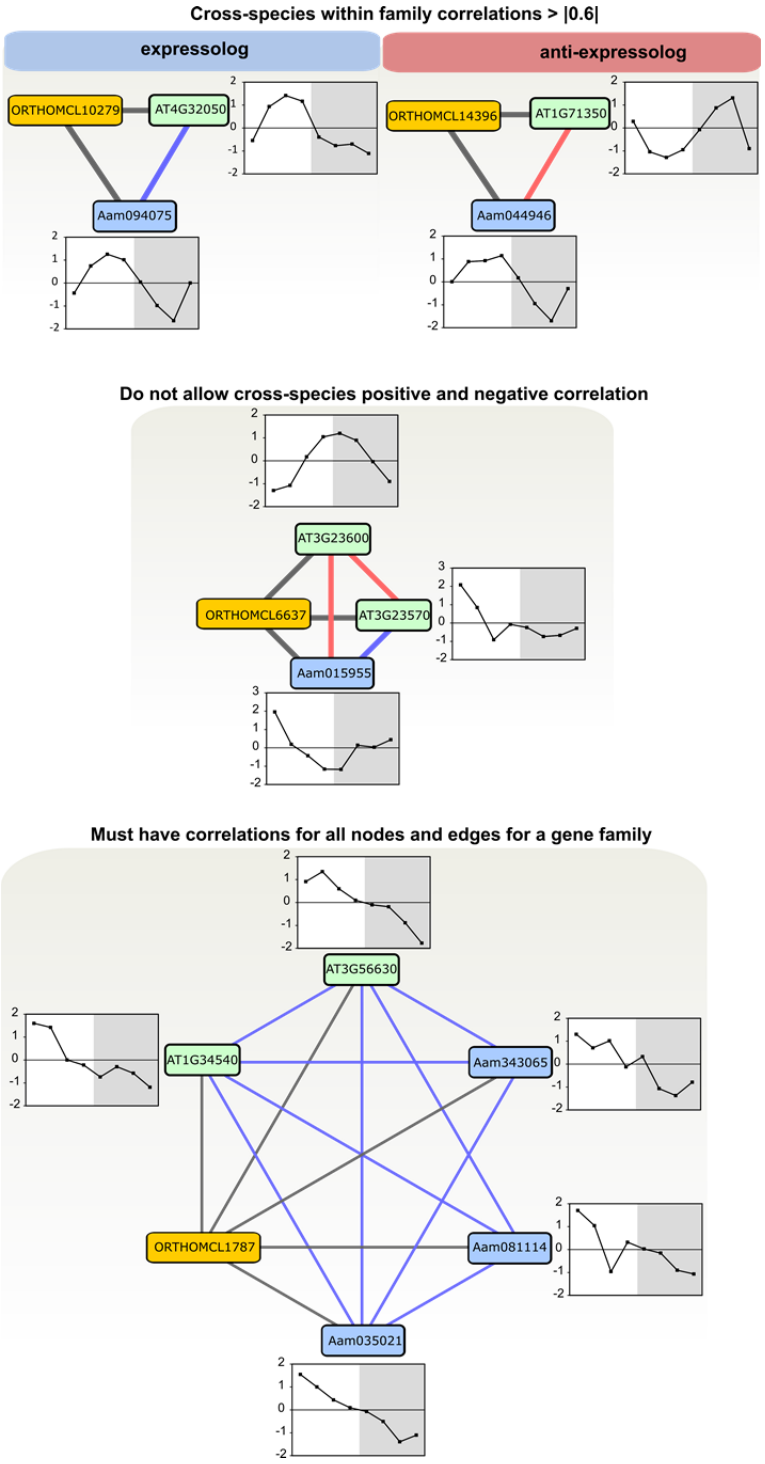
Supplemental Figures



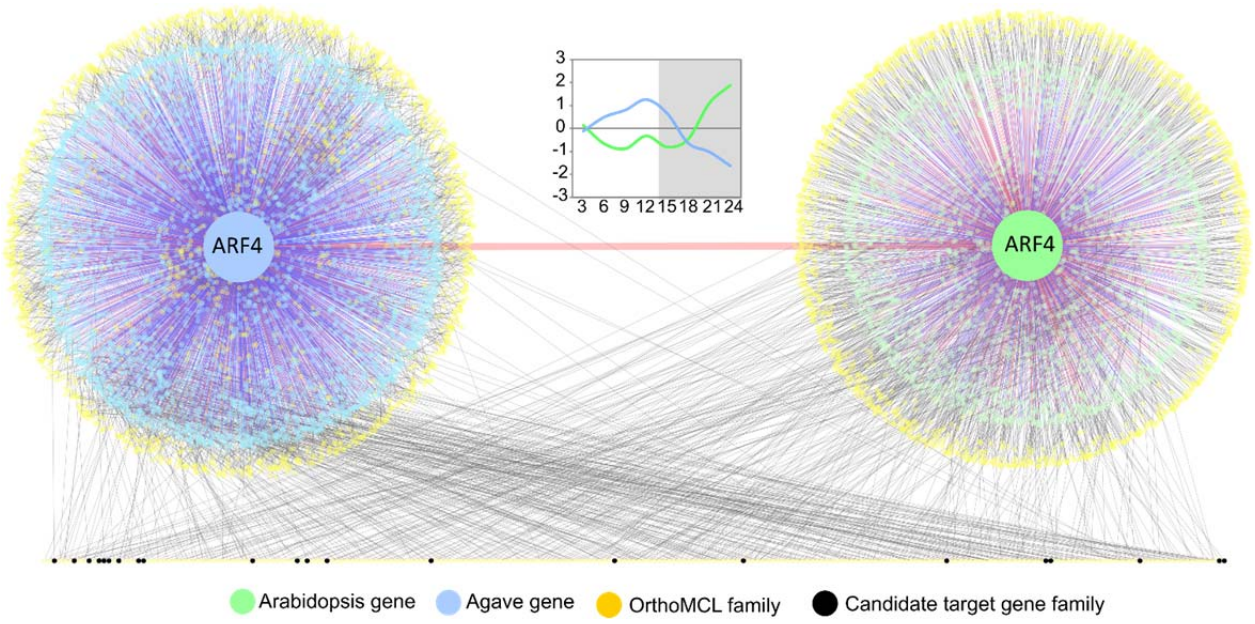
Supplemental Figure 1. Comparison of metabolite relative abundances using fresh weight or dry weight. The relative abundance of malic acid in *Agave* was normalized by the leaf biomass (a) fresh weight (FW) and (B) dry weight measurements. The relative abundance of the metabolic cofactor NADP was normalized by the leaf biomass (c) fresh weight measurements and (d) dry weight using protein biomass measurements.



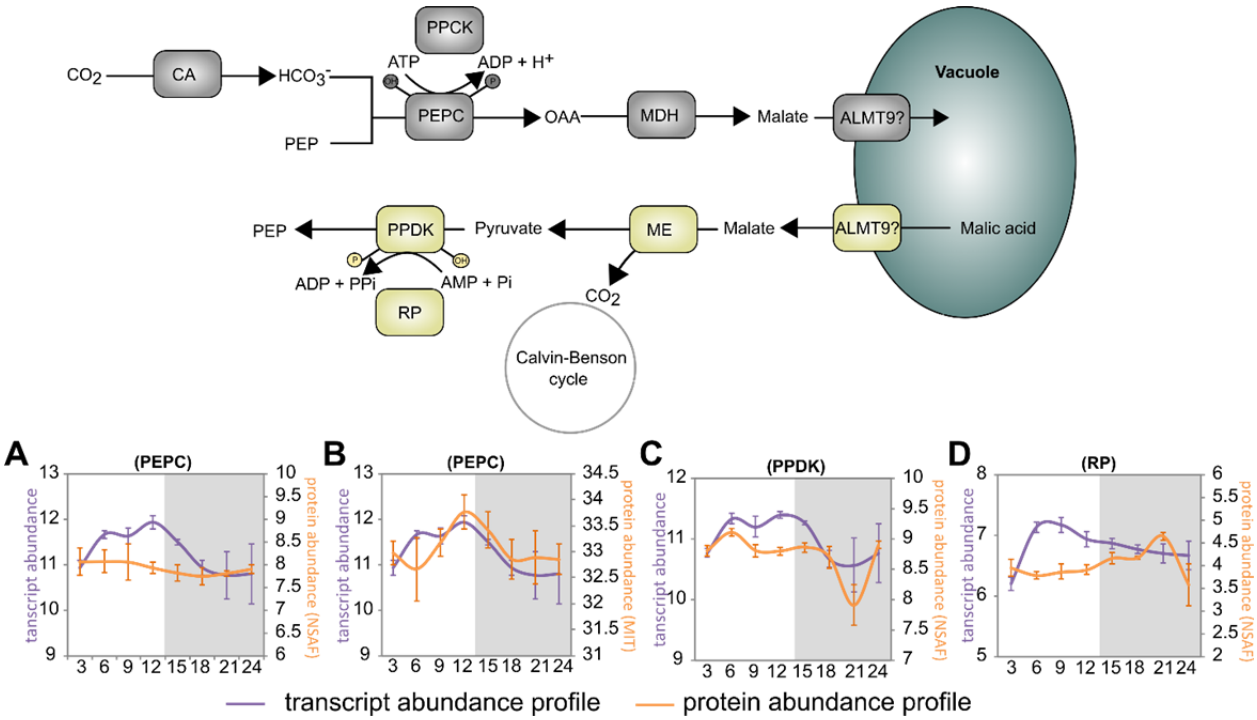
Supplemental Figure 2. *Arabidopsis* gene expression measurements were adjusted using cubic spline interpolation and then time points in blue triangles were averaged to get an expression value for the *Agave* time point between them.



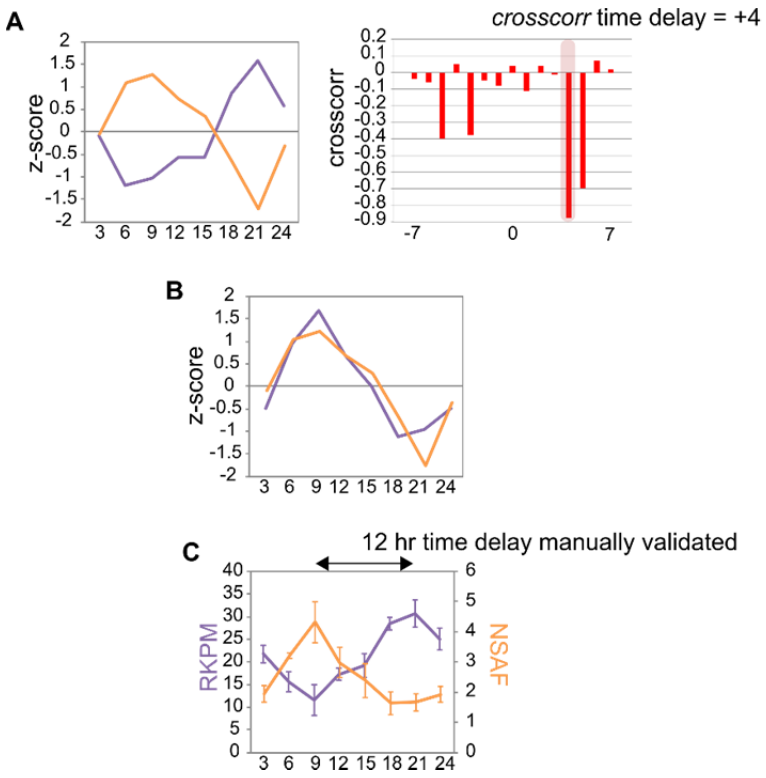
Supplemental Figure 3. Illustration of criteria applied to reduce ambiguity and improve confidence in comparison of ORTHOMCL orthologs between *Agave* and *Arabidopsis* co-expression networks to identify candidate regulators of rescheduled CAM processes.



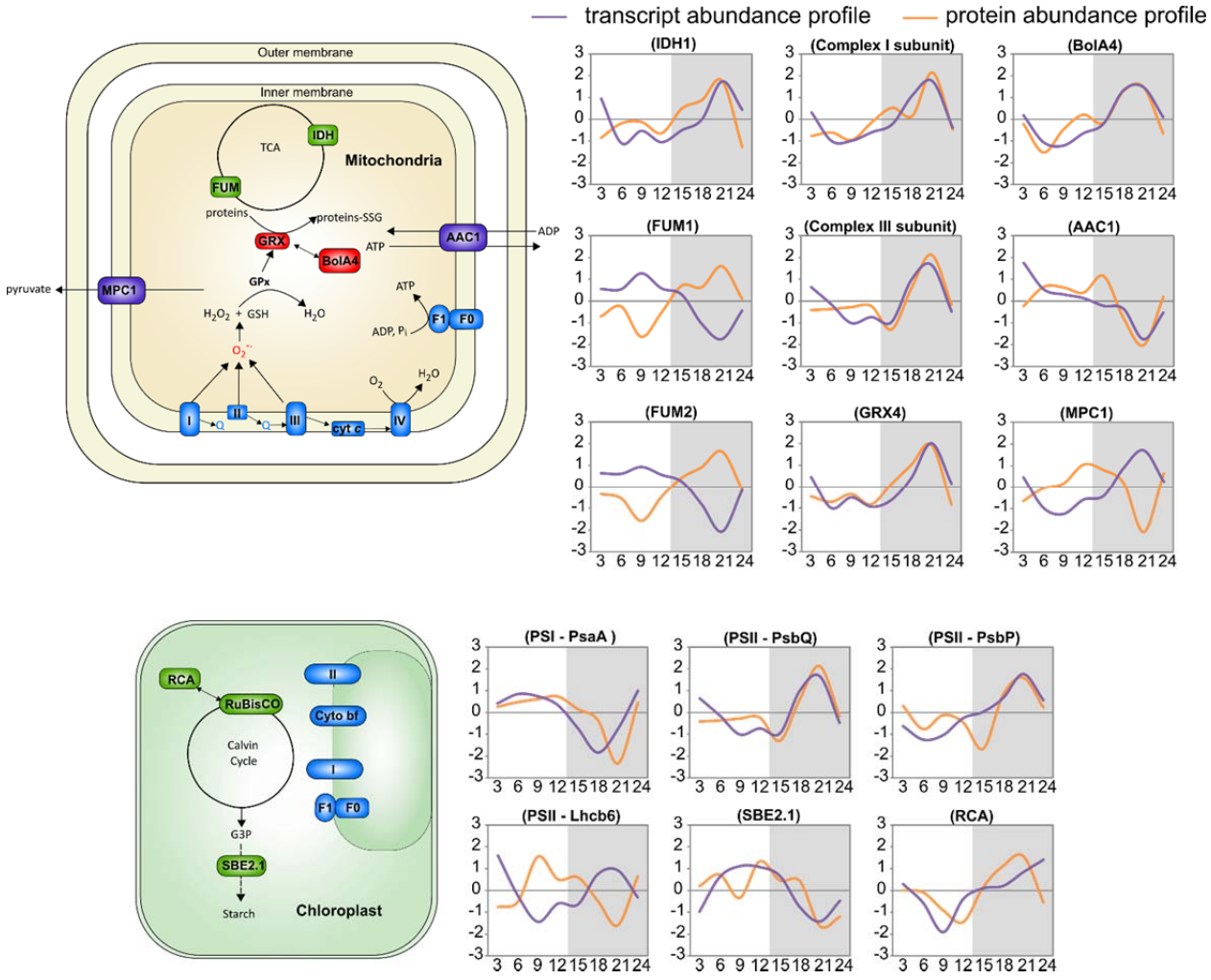
Supplemental Figure 4. Comparative cross-species co-expression network for an auxin response factor (ARF) to identify candidate target genes. Graphical network representation of rescheduled gene expression within the gene family OrthoMCL11025, which represents auxin response factor 4 (ARF4). The y-axis represents the standard deviation z-score [(expression - mean)/SD] and shows expression peaking at the end-of-night in *Arabidopsis* (green) and end-of-day in *Agave* (blue). For the cross-species co-expression network, relationships between genes are represented in networks, where nodes represent genes passing the correlation threshold (>= |0.8|) and edges represent the correlation relationship to the species TF and also link to an OrthoMCL gene family. Blue edges represent a positive correlation and red edges represent a negative correlation.



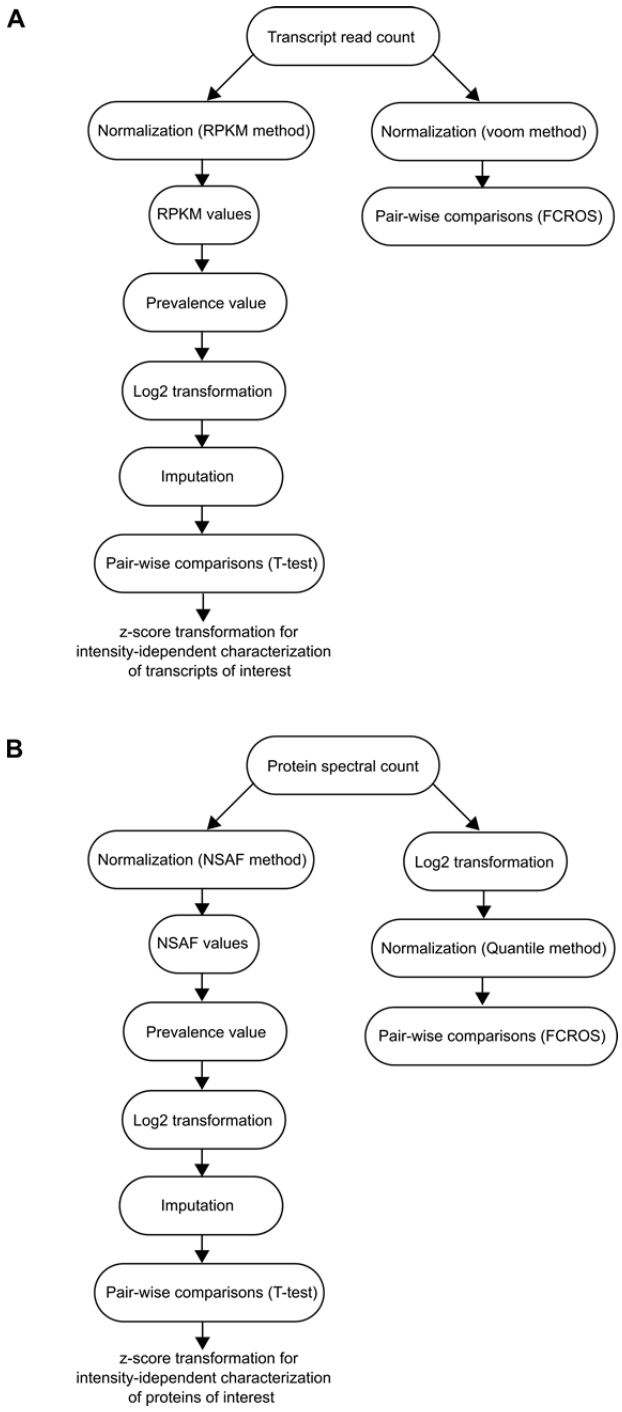
Supplemental Figure 5. Diel expression and protein abundance patterns for genes implicated in carboxylation and decarboxylation reactions in CAM. Transcript abundance (Log2 transformed RPKPM values; purple) and protein abundance (NSAF or MIT; orange) profiles. Error bars represent the standard error for three biological replicates. (a) The transcript and protein abundance profile of *Agave* phosphoenolpyruvate carboxylase (PEPC) (Aam080248) using Log2 transformed NSAF values or (b) peak intensity information obtained from matched ion intensities (MIT) (see Supplementary Note 6), as a function of time (hours from the beginning of the light period on x-axis). (c) The expression and protein abundance profiles of *Agave* pyruvate orthophosphate dikinase (PPDK) (Aam010102) and (d) PPDK-regulatory protein (RP) (Aam051010) are shown.



Supplemental Figure 6. Illustration of criteria applied to report high-quality transcript and protein abundance relationships. (a) the *crosscorr* function from Econometrics Toolbox (Matlab) was first applied to first estimate temporal differences in transcripts and proteins having similar abundance profiles. (b) After enforcing *crosscorr* threshold of 0.7, the time delay was then used rearrange the data to calculate a Pearson correlation coefficient and *p*-value. (c) The relative abundance and standard error from the mean (SEM) was then used for manual validation.



Supplemental Figure 7. Temporal dynamics of transcript and protein abundances for mitochondria- and chloroplast-related genes. The y-axis represents the standard deviation z-score for each gene or protein [(abundance - mean)/ SD].



Supplemental Figure 8. Test for differential abundance workflow. This illustration provides an overview of the data processing workflow for (A) transcript and (B) protein data to test for differential abundances between time points. For both datasets, two different statistical tests were performed. Note, for the protein data, in order to address the stochastic sampling of low-

abundant proteins, only proteins above the limit of quantification were assessed by the FCROS method.

Supplementary Tables

Supplemental Table 1
Net gas exchange and photon flux density measurements across the *Agave* diel cycle.

Supplemental Table 2
Fresh weight (FW) concentration (µg/g) of *Agave* metabolites normalized to sorbitol for diel time (DT) 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period.

Supplemental Table 3
Fresh weight (FW) concentration (µg/g) of *Arabidopsis thaliana* (Col-0) metabolites normalized to sorbitol for diel time (DT) 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period.

Supplemental Table 4
List of *Agave* transcripts that passed the prevalence value threshold. Values represent Log2 transformed RPKMs. Missing values were replaced by values imputed by drawing random numbers from a normal distribution to simulate signals from low abundant transcripts.

Supplemental Table 5
Differential abundance analyses for Log2 transformed RPKM values in *Agave* for diel time (DT) 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period.

Supplemental Table 6
K-means clusters for differentially abundant *Agave* transcripts for diel time (DT) 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period.

Supplemental Table 7
List of over-represented biological process ontologies for each transcript K-means cluster. Each value represents a corrected p-value from a right-sided hypergeometric enrichment test using the Bonferroni step down method. Terms that did not pass a significance threshold are marked as "NaN".

Supplemental Table 8
Arabidopsis microarray abundance values (Mockler et al., 2007).

Supplemental Table 9
Reciprocal best BLAST hit gene expression relationships having identified Pearson correlation coefficients > |0.6|

Supplemental Table 10
List of ORTHOMCL gene families between *Agave* and *Arabidopsis*

Supplemental Table 11
List of candidate target genes of *Agave* ARF4 (Aam004755)

Supplemental Table 12
An inclusive list of all Agave and chloroplast protein accessions that passed the FDR threshold. These values represent raw spectral counts that have not been normalized.

Supplemental Table 13
An inclusive list of protein accessions that passed the prevalence value threshold. Only those protein accessions that belonged to proteins groups unambiguously identified (i.e., unique) were used for quantitative analyses. Values represent Log2 transformed NSAF values. Missing values were replaced by values imputed by drawing random numbers from a normal distribution to simulate signals from low abundant proteins.

Supplemental Table 14
Differential abundance analyses for Log2 transformed NSAF values in *Agave* for diel time (DT) 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period.

Supplemental Table 15
K-means clusters for differentially abundant *Agave* proteins for diel time (DT) 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period. ** signifies differential abundances determined to be statistically significant by the ranked fold change method (Dembélé and Kastner, 2014).

Supplemental Table 16
List of over-represented biological process ontologies for each protein K-means cluster. Each value represents a corrected p-value from a right-sided hypergeometric enrichment test using the Bonferroni step down method. Terms that did not pass a significance threshold are marked as "NaN".

Supplemental Table 17

List of high-quality *Agave* mRNA and protein temporal abundance relationships.

Supplementary Notes

1. Plant material—*Agave americana* "marginata" plants were obtained from Notestein's Nursery, Gainesville, FL (http://southerngardening.org). Two-year-old plants with an average of eight leaves per plant were placed in 3.8-gallon (14.4-liter) plastic pots filled with Metro-Mix PX3 soil (Sun Gro Horticulture, Agawam, MA, USA). The plants were maintained in a Conviron CMP6050 Control system (Pembina, ND, USA) with day/night temperatures of 25/15^{o}C and day/night relative humidity of 45/75%. The photoperiod was 12-hr with a photosynthetic photon flux (PPF, 400 to 700 nm) of 540 µmol m^{-2} s^{-1} on the upper surface of the leaves examined, as determined with a quantum meter (Model MQ-100, APOGEE, USA). The plants were watered once weekly until drainage from the pots occurred.

Arabidopsis thaliana (Col-0) plants used for gas chromatography mass spectrometry (GC-MS) measurements were obtained from TAIR (http://www.arabidopsis.org/). Seeds were incubated with desiccant for 48 hours, sterilized in 75% ethanol and 100% ethanol for 8 minutes each, respectively, and air dried on sterile paper filter inside the hood. Seeds were plated on 0.5X MS media (M5524, Sigma, USA), stratified for 2-5 days at 4ºC in the dark, and placed in growing room at 23ºC. Photosynthetic photon flux (PPF, 400 to 700 nm), as determined with a quantum meter (Model MQ-100, APOGEE, USA), was 70 µmol m^{-2} s^{-1} for day period (14-hr) and zero for night period (10-hr). After 8 days, seedlings were transplanted into plastic trays of 10"x20" (6 plants per pot, 12 pots per tray) in soil 3B MIX (Conrad Fafard INC. MA, USA) and maintained under similar conditions (except for PPF = 100-120 µmol m^{-2} s^{-}) in a Conviron CMP6050 Control system (Pembina, ND, USA) at 50% humidity. Plants were watered twice a week (~1000ml) and fertilized twice a month (Miracle-Gro All Purpose Dry Plant Food). Adult plants (3 weeks on soil) had 3 leaves from the rosette harvested from different plants for each time point trialed. Each plant was harvested only once. Biological replicates were sampled from same batch of plants grown under similar conditions.

2. Leaf Gas Exchange—Net CO_2 uptake in *A. americana* was measured using a compact mini cuvette system, Central Unit CMS-400 with BINOS-100 infra-red gas analyzer, working in an open format (Heinz Walz GmbH, Germany). A single leaf, ranging in age from the youngest still-expanding to the 4th fully expanded, was clamped in the cuvette, ensuring it received full

light (i.e., 400 μmol m^{-2} s^{-1}) within the growth chamber. Environmental conditions used for gas exchange analyses tried to mimic as closely as possible those experienced by plants sampled for metabolomics, mRNA and protein sequencing (i.e., the oldest leaf used for gas exchange analyses was leaf #4). Temperature of the cuvette was set to track environmental conditions within the growth room (i.e., 27°C day/19°C night, 60/80% day/night relative humidity, 12-hr photoperiod). Data for net CO_2 uptake were collected every 15 minutes and gas flow through the cuvette was maintained between 400 and 500 mL min^{-1} to avoid water condensation inside the cuvette. Each leaf was maintained inside the cuvette for at least 48 hours to get a complete 24-hr gas exchange profile. Data were analyzed using DIAGAS software (Heinz Walz GmbH, Germany) based on the area of leaf inside the cuvette. Each leaf gas exchange curve presented is representative of that obtained from 3 biological replicates (Supplemental Table 1).

3. Gas chromatography mass spectrometry metabolite profiling— For *A. americana* metabolite identification, a total of 8 samples were collected with three biological replicates for samples (4^{th} fully expanded leaf) collected every at 3, 6, 9, 12, 15, 18, 21, 24 hours after the starting of the light period. The plant samples were frozen in liquid nitrogen and ground using a mortar and pestle and frozen at −80°C until further use. For *Arabidopsis thaliana* (Col-0) metabolite identification, a total of 8 samples were collected with three biological replicates of fully-expanded leaf samples collected at 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period. The plant samples were frozen in liquid nitrogen and ground using a mortar and pestle and stored at −80°C until further use.

Fast-frozen tissues were ground with liquid nitrogen in a chilled mortar and pestle with ~50 (19-60) mg FW Agave leaf subsequently twice extracted with 2.5 mL 80% ethanol overnight and then combined prior to drying a 1.0 ml aliquot in a nitrogen stream. A 50 μL aliquot was also dried for analysis of high concentration metabolites. For Arabidopsis, ~120 (52-161) mg FW of fast-frozen plant tissue was twice extracted with 2.0 mL 80% ethanol overnight and then combined prior to drying a 0.5 ml aliquot in a nitrogen stream. Sorbitol (75 μL of a 1 mg/mL aqueous solution) was added to the first 80% ethanol extraction volume into which the frozen tissue was directly weighed for extraction as an internal standard to correct for differences in extraction efficiency, subsequent differences in derivatization efficiency and changes in sample volume during heating. Dried extracts were dissolved in 500 μL of silylation–grade

acetonitrile followed by the addition of 500 µL N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (Thermo Scientific, Bellefonte, PA), and samples were then heated for 1-hr at 70°C to generate trimethylsilyl (TMS) derivatives (Li et al., 2012, Tschaplinski et al., 2012). After 2 days, 1-µL or 0.1-µL aliquots were injected into an Agilent Technologies Inc. (Santa Clara, CA) 5975C inert XL gas chromatograph-mass spectrometer, fitted with an Rtx-5MS with Integra-guard (5% diphenyl/95% dimethyl polysiloxane) 30 m x 250 µm x 0.25 µm film thickness capillary column. The standard quadrupole GC-MS was operated in the electron impact (70 eV) ionization mode, targeting 2.5 full-spectrum (50-650 Da) scans per second, as described previously (Tschaplinski et al., 2012). TCA cycle organic acids, sugars, and abundant secondary metabolites known or thought to be under diurnal regulation were the focus of this study. Metabolite peaks were extracted using a key selected ion, characteristic *m/z* fragment, rather than the total ion chromatogram, to minimize integrating co-eluting metabolites. The extracted peaks of known metabolites were scaled back up to the total ion current using predetermined scaling factors. Peaks were quantified by area integration and concentrations normalized to the quantity of the internal standard (sorbitol) recovered, amount of sample extracted, derivatized, and injected. A large user-created database (>2300 spectra) of mass spectral electron impact ionization (EI) fragmentation patterns of TMS-derivatized compounds, as well as the Wiley Registry 10th Edition combined with NIST 14 mass spectral database, were used to identify the metabolites of interest to be quantified. Unidentified metabolites were denoted by their retention time as well as mass-to-charge (*m/z*) ratios. The *Agave* data (Supplemental Table 2) and *Arabidopsis* data (Supplemental Table 3) have been provided. Because we highlight the relative abundances of metabolites for both succulent leaves (i.e., *Agave*) and standard leaves (i.e., *Arabidopsis*), we compared the relative abundances of metabolites using both fresh weight and dry weight as normalizing measures. For both plants, we did not observe a substantial difference between the two methods (Supplemental Figure 1). Research has shown that changes in water content between day and night periods of well-watered plants is small (Castro-Camus et al., 2013).

4. NADPH and NADP+ measurement—For *A. americana* measurements, 4th fully expanded leaves were collected with biological duplicates at 3-hr intervals for a 24-hr period. For *A. thaliana* (Col-0), the wild-type plants were grown in 3.5" plastic pots containing 0.6 liter

Sunshine 781 soil mix (custom blend, 45-50% peat moss) (Scotts Sierra Horticultural Products, Marysville, OH) at 23/21°C (day/night) in a Percival® model AR-77L2 growth chamber under 12-hr photoperiod (light, 135 μmol m^{-2} s^{-1}) conditions for four weeks. Fully expanded 5th leaves were collected with biological triplicate at 3-hr intervals for a 24-hr period. Each of ground samples (~210 mg for *Arabidopsis* and ~ 580 mg for *Agave*) was deproteinized and neutralized using the Deproteinizing Sample Preparation Kit (BioVision, Mountain View, CA) according to manufacturer's instructions. For all samples, enzymatic assays were performed to determine total NADP and calculated NADPH from decomposed NADP using the NADP/NADPH Quantification Kit (BioVision, Mountain View, CA) according to manufacturer's instructions. Cofactor values were normalized to the amount of material and were reported as picomoles of cofactor per milligram of fresh weight.

Theoretical energetics of the CAM cycle in Agave

The assimilation of 1 mole CO_2 and accumulation of 1 mole of malic acid in the vacuole at night requires 1 mole of ATP. In soluble sugar storing PEPCK-type CAM plants like *Agave*, it is predicted that this ATP is produced by mitochondrial oxidative phosphorylation with the most plausible respiratory substrate being malate (Winter and Smith, 1996). Complete oxidation of 1 mole malate to 4 moles CO_2 can yield 14.75 moles ATP (Winter and Smith, 1996). These respired 4 CO_2 must be conserved at night by re-fixation via PEPC into malate, but 3 of the 4 malate produced can be removed to the vacuole as malic acid, at a cost of 1 ATP per malic acid accumulated. During the subsequent day, the CO_2 released by decarboxylation of these 3 malates will be converted *via* photosynthetic carbon reduction back to storage carbohydrate which will be retained to provide PEP for nocturnal carboxylation. To sustain steady-state operation of the day-night CAM cycle, any respiratory CO_2 produced from malate at night must be quantitatively re-assimilated back to storage carbohydrate during the day. Net energy requirements for day-time decarboxylation in a PEPCK–type CAM plant have been calculated as 3.8 ATP: 2.6 NADPH per CO_2 (Winter and Smith, 1996). Thus over a 'typical' 24-hr cycle in a PEPCK-type CAM plant, the theoretical net energy requirement is 4.8 ATP: 3.2 NADPH per CO_2 assimilated (see Winter and Smith, 1996 for a detailed description of theoretical energetics of different CAM sub-types). Modelling of the diel CAM cycle at a network scale *via* flux balance analysis provides additional insight into the energetics of CAM and has shown that photon use in a

mature CAM leaf is similar to that in a C_3 leaf, being ±10% of C_3 photosynthesis depending on the CAM subtype (Cheung et al., 2014). Thus, there appear to be no overall energetic advantage to CAM compared to C_3, despite the potential for suppression of photorespiration through CO_2 concentrating mechanism in CAM.

5. Transcriptomics—For transcriptome sequencing, a total of 15 *A. americana* samples were collected with three biological replicates, including eight mature leaf (4^{th} fully expanded leaf) samples collected at 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period; three young leaf samples collected at diel time points 6, 12, and 21-hr, respectively; and four non-leaf samples (i.e., meristem, rhizome, root, stem) collected at time point 3-hr after the starting of the light period. The plant samples were frozen in liquid nitrogen and ground using a mortar and pestle and frozen at −80°C until further use.

RNA isolation

RNA was extracted from *A. americana* samples using a Spectrum™ Plant Total RNA isolation kit (Sigma, St. Louis, MO, USA) according to the protocol provided. The increased binding buffer option was used due to the high water content of the tissues. The optional on-column DNase treatment was included during RNA isolation to rid the samples of potential genomic DNA contamination. Total RNA quantity was determined using a NanoDrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA), and RNA quality was determined using an Experion RNA StdSens Analysis kit (Bio-Rad Laboratories, Hercules, CA, USA). Only non-degraded samples with an acceptable A260: A280 ratio (≥1.8) and RNA Quality Indicator (RQI) ≥ 7 were used. mRNA purification was done using a Dynabeads® mRNA purification kit (Invitrogen, USA).

Transcriptome sequencing using Roche/454 platform

Purified mRNA was used to synthesize a double-stranded cDNA library using the cDNA Rapid Library Preparation protocol provided by Roche 454 (Branford, CT, USA). After the library was synthesized, library quantity was determined using a 96-well plate Fluoroskan Ascent (Labsystems). Library quality was determined using an Agilent Bioanalyzer High Sensitivity DNA chip. The libraries were then diluted according to the protocol provided by

Roche. Fragmentation of the cDNA library was done according to the protocol provided by Roche and resulted in an average fragment size of 1000bp. A GS FLX Titanium emPCR Lib-L SV kit (Roche) was used to do an emulsion titration on the fragmented library to determine the amount of library to add to the large volume emulsions. After following the Roche protocol to determine the proper amount, a GS FLX Titanium emPCR Lib-L LV kit was used to populate beads with the cDNA library. The resulting beads were sequenced using a GS Titanium sequencing kit XLR70 on a Genome Sequencer FLX Instrument (Roche).

Transcriptome sequencing using Illumina HiSeq platform

The mRNA was prepared into sequencing libraries as described previously (Wang et al., 2011). Each library was assayed by an Agilent High Sensitivity Chip (Agilent, Cat. No. 5067-4627) and measured using the dsDNA-HS protocol on the Qubit Fluorometer (Life Technologies). Equal quantities of libraries (~5 ng per sample) with different indices were mixed and stored in −80°C freezer before sequencing. Sequencing was performed in a v3 flowcell on an Illumina HiSeq 2000 sequencer, using the TruSeq Paired-End Cluster Kit v3 (Illumina PE-401-3001) and the TruSeq SBS HS Kit v3 200 cycles (Illumina FC-401-3001), generating 2 × 100 bp reads. Image analysis and base calling was done using the HiSeq Control Software v1.4 and the Off-Line Base Caller v1.9.

Transcriptome assembly and transcripts expression estimates in Agave americana

~231 million high quality RNA-Seq reads (96bp with quality score >20 for each base) pooled from Illumina sequencing of each of the 45 samples (three biological replicates of 15 tissues) (NCBI SRA accession SRS631988) were assembled into contigs using Trinity (Release2012-04-27) (Grabherr et al., 2011), with the k-mer set as 25. Then Roche/454 sequencing reads (NCBI SRA accession SRS632003) that were not redundant with the contigs, as obtained by comparison using CD-HIT-EST-2D (Fu et al., 2012, Li and Godzik, 2006) with a sequence identity threshold of 0.95, were combined with the contigs and assembled into unigenes using CAP3 (Huang and Madan, 1999), with an overlap length cutoff of 40 and an overlap identity of 95%. The RNA-Seq reads from 15 biological samples (NCBI SRA accessions SRS631987 and SRS631989 - SRS632002; three biological replicates per sample) were aligned to the unigenes using bowtie (Langmead et al., 2009) and abundance was estimated using RSEM

(Li and Dewey, 2011) implemented in the Trinity (Release2012-05-18) (Grabherr et al., 2011), with abundance defined as Reads Per Kilobase of transcript per Million mapped reads (RPKM). A total of 91,702 unigenes with an average abundance (i.e., the average of three biological replicates) of 5 RPKM or higher in at least one of the 15 biological samples were retained as the final transcriptome assembly. The stranded Illumina RNA-Seq reads were mapped to the unigenes using blastn (Altschul et al., 1990) to identify the strand of unigenes. In the 91,702 unigene set, 61,634 transcripts are 200bp or longer. Among these 61,634 transcripts, 85 contaminant sequences (e.g., Non-viridiplantae, rRNA, Vector) and one duplicated sequence were identified automatically through the NCBI Transcriptome Shotgun Assembly (TSA) submission process. After removing these contaminant/duplicated sequences, 61,548 transcripts of 200bp or longer in length were deposited at GenBank under the accession GBHM00000000.

For the *Agave* gene expression analyses, only samples pertaining to the mature leaf tissue (4^{th} fully expanded leaf) were used. To assess reliable quantitative differences across the mature leaf sample set, only those transcripts with substantive abundance values, as determined by prevalence value (PV) (Lochner et al., 2011), were carried on to subsequent quantitative analyses. Rather than choosing an arbitrary threshold, each identified transcript was given a PV, which is determined by averaging the RPKM values across all samples. Next, PVs were plotted as a histogram to graphically capture the distribution of abundances, such that one could assess the cumulative abundances assigned at varying PV cutoffs. An inflection point was identified, where transcripts with a minimum average RPKM $\geq$ 3.483 were considered to be highly representative and reproducible.

As supplemental information, we identified over-represented gene ontology biological processes (GOBP) for each cluster (Supplemental Table 7). In the main text, Figure 2 highlights the five most over-represented GOBP categories for each cluster. In general, the functional analysis reveals strong enrichment of photosynthesis-related categories, in which associated genes predominately accumulated at the beginning of the day (Cluster 4) or during the night (Cluster 3). In Cluster 4, transcripts associated with the over-represented GOBP category ***photosynthesis, light harvesting*** have coordinated expression patterns with genes associated with GOBPs ***response to heat, response to high light intensity,*** and ***response to osmotic stress***. The peak morning expression of these genes are consistent with the light-induced processes of photosynthetic electron transport and with the metabolic consequences of day-time stomatal

356 closure, which could potentially elevate heat load on the leaf and impact leaf osmotic relations.
357 In Cluster 3, transcripts with predominately higher abundance at night were associated with the
358 over-represented GOBP category ***photosynthesis*** include photosystem I subunits ***PSAO***
359 (Aam015317), ***PSAN*** (Aam303305), ***PSI-P*** (Aam047661), ***PSAH-1*** (Aam011059), ***LHCA2***
360 (Aam049417), and photosystem II subunits, for example ***PSBY*** (Aam311217), ***PSBR***
361 (Aam339724), ***PSBX*** (Aam006871), ***PSBW*** (Aam016138). Coordinated expression of genes that
362 are essential components of both photosystems and of photorespiration, together with an
363 enrichment of genes implicated in translation, can be envisaged as a means of accommodating
364 pre-dawn assembly of proteins that are critical for the effective harvesting of light and preventing
365 over-reduction of the electron transport chain.
366
367 ***Protein sequence prediction from Agave americana transcript sequences***
368 The open reading frames (ORFs) were annotated using six-frame translation based on
369 standard genetic code with a length range of 10 - 10,000 amino acids. The best ORF for each
370 transcript was chosen on the + strand of the transcript with the following criteria (1) having the
371 highest score in blastp (Altschul et al., 1990) search, with default setting, against the UniRef90
372 database (http://www.uniprot.org/) if there were blastp hits or (2) the longest ORF if there were
373 no blastp hits. In total, 70,257 representative protein sequences in *A. americana* were identified
374 and used as a reference database for proteomics.
375
376 ***Arabidopsis gene expression data***
377 The diurnal expression data for *Arabidopsis thaliana* were obtained from (Mockler et al.,
378 2007). The *Arabidopsis* expression data were collected at 4, 8, 12, 16, 20, and 24 hours whereas
379 the *Agave* data were collected at 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light
380 period. Since the *Arabidopsis* gene expression data was measured at 4-hr intervals and the *Agave*
381 gene expression data was measured at 3-hr intervals, the *Arabidopsis* expression data was
382 adjusted (Supplemental Figure 2) to arrive at expression profiles for all *Arabidopsis* and *Agave*
383 genes on the same time scale. Here, the cubic interpolation algorithm
384 (http://www.SRS1Software.com) was used to simulate the *Arabidopsis* gene expression levels at
385 additional time points and specific time points were averaged so that both time-course data sets

consisted of the same time intervals: 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period.

Ortholog analysis

The ortholog groups (OGs) were constructed using OrthoMCL (Li et al., 2003). An all-vs-all BLASTP was performed to calculate the amino acid sequence similarity between all pairs of *Agave* and *Arabidopsis* genes. This was performed using the standalone BLAST tool, version 2.2.26 (Altschul et al., 1990). An e-value threshold of 10^{-5} was applied. The FastOrtho implementation of OrthoMCL (Wattam et al., 2014), http://enews.patricbrc.org/fastortho/) was then used to determine orthologous protein families from the resulting pairwise sequence similarities. The final step in the OrthoMCL algorithm involved the use of the clustering algorithm MCL (Van Dongen, 2001), http://micans.org/mcl) to cluster the genes into their respective orthologous groups. An inflation value of 1.5 was used.

Cross-species co-expression networks for the detection of candidate regulators of reprogrammed metabolism

Gene co-expression networks were constructed within *Arabidopsis* and *Agave* species and across both species. The Pearson correlation coefficient was then calculated between the expression profiles of all pairs of genes (within and across species) using the mcxarray software in the MCL-edge package (Van Dongen, 2001), http://micans.org/mcl). Co-expression networks were then visualized in Cytoscape (Shannon et al., 2003).

Because *Arabidopsis* and *Agave* have both undergone multiple whole-genome duplication (WGD) events, the elucidation of functional orthologs is complicated by the different evolutionary trajectories that duplicated genes may follow (*e.g.*, neo-functionalization, sub-functionalization, retention of function, or loss of function). This evolutionary history further complicates the process of comparing orthologs across species, leading to many-to-many orthologous relationships. To identify transcription factor gene expression patterns that have been preserved across the evolutionary history of plants (*i.e.*, expressologs; homologous genes with similar expression patterns) and to identify those that have a rescheduled diel pattern of abundance (*i.e.*, anti-expressologs; homologous genes with opposite expression patterns), we used the OrthoMCL algorithm to identify gene families (Supplemental Table 10). For the subset

of orthologous genes that were differentially expressed in the *Agave* dataset, we computed Pearson's correlation coefficients between expression profiles of homologs between species. To select a correlation threshold, we graphed Pearson correlation distributions and selected the value near the observable inflection point of the distribution (i.e., |0.6|). The Pearson correlation levels are provided in the Supplemental Table 9. Cross-species identification of expressologs or anti-expressologs can be challenging when a set of genes within a species for a particular family show varying expression profiles (Supplemental Figure 3). For example, the occurrence of two paralogs with opposing profiles introduces ambiguity in the identification of cross-species expressologs or anti-expressologs. Because the presence of both a positive and negative cross-species correlation within a gene family introduces ambiguity into the interpretation of preserved or rescheduled expression in a functional genetic unit, we limited our analysis to only those gene families displaying a single expression profile within a species. Moreover, because there is not yet a well-curated genome for *Agave*, we avoided misinterpretation by requiring gene families to have acceptable correlations values for all nodes and edges.

For the detection of candidate regulators of reprogrammed metabolism, we aimed to identify which genes annotated as transcription factors exhibited an anti-correlative relationship between *Agave* and *Arabidopsis*. In total, 10 transcription factors with reciprocal expression patterns were confidently identified in *Agave*. Given the considerable amount of information available on the regulation of auxin response factors (ARFs), we performed a cross-species co-expression analysis of **ARF4** (Aam004755), which has shifted its expression from end-of-night in *Arabidopsis* to a predominately end-of-day in *Agave* (Supplemental Figure 4). Because these TFs can either activate or repress gene expression, the co-expression networks developed for *Arabidopsis* and *Agave* ARF4 permitted relationships with Pearson correlation coefficient >=|0.8|. A comparative cross-species co-expression network representing the many-to-many relationships between *Agave* and *Arabidopsis* ARF gene families has been provided (Supplemental Figure 4). Here, co-expression relationships among genes are represented in networks, in which nodes represent gene correlations that exceed the correlation threshold and edges represent the correlation to the species ARF4 and also link to an OrthoMCL gene family. In total, we identified 239 OrthoMCL gene families present in both co-expression networks. These gene families were limited to those that passed the ambiguity criteria outlined above, resulting in 25 gene families (Supplemental Table 11). With tractable, rescheduled gene

expression patterns in *Agave*, these gene families represent candidate genes or functional processes that could be regulated by this particular TF. Interestingly, over half of the candidate gene families encode proteins located in the nucleus and several have been related to growth and development. Because ARF transcription factors target genes containing auxin response elements (AuxRE), we asked the question whether any of the candidate gene families contained predicted AuxREs in their regulatory regions. Although ChIP-seq data for the observed ARF is not yet available, we leveraged data from a recent computational analysis that identified genes containing AuxREs in the *Arabidopsis* genome (Mironova et al., 2014) to search for genes containing AuxRE motifs. In the 25 gene families, six contained an *Arabidopsis* gene with at least one AuxRE, and therefore these targets are more likely to be activated or repressed by ARF4.

6. Proteomics—For proteome sequencing, a total of 8 *A. americana* samples were collected with three biological replicates, including eight mature leaf (4^{th} fully expanded leaf) samples collected every at 3, 6, 9, 12, 15, 18, 21, and 24 hours after the starting of the light period. The plant samples were frozen in liquid nitrogen and ground using a mortar and pestle and frozen at −80°C until further use.

Protein extraction and digestion

For all samples, ~2−4 g of ground *A. americana* tissue was suspended in SDS lysis buffer (4% SDS in 100 mM of Tris-HCl), boiled for 5 min, sonically disrupted (30% amplitude, 10 s pulse with 10 s rest, 2 min total pulse time) and boiled for an additional 5 min. Crude protein extract was pre-cleared via centrifugation, and quantified by BCA assay (Pierce Biotechnology). Three milligrams of crude protein extract were precipitated by trichloroacetic acid (TCA), pelleted by centrifugation and washed with ice-cold acetone to remove excess SDS. As previously described (Abraham et al., 2013), pelleted proteins were resuspended in 250 µL of 8 M urea, 100 mM Tris-HCl, pH 8.0 using sonic disruption to fully solubilize the protein pellet and incubated at room temperature for 30 min. Denatured proteins were then reduced with DTT (10 mM) and cysteines were blocked with iodoacetamide (20 mM) to prevent reformation of disulfide bonds. Proteins were digested via two aliquots of sequencing-grade trypsin (Promega, 1:75 [w:w]) at two different sample dilutions, 4 M urea (overnight) and subsequent 2 M urea (3-

hr). Following digestion, samples were adjusted to 200 mM NaCl, 0.1% formic acid and filtered through a 10 kDa cutoff spin column filter (Vivaspin 2, GE Health) to remove under-digested proteins. The peptide-enriched flow through was then quantified by BCA assay, aliquoted and stored at −80°C.

Two-dimensional liquid chromatography tandem mass spectrometry

~ 25 µg of each peptide mixture were bomb-loaded onto a biphasic MudPIT back column packed with ~3 cm strong cation exchange (SCX) resin followed by ~3 cm C18 reversed phase (RP) (Luna and Aqua respectively, Phenomenex). Peptide-loaded columns were first washed off-line to remove residual urea and NaCl and then placed in-line with an in-house pulled nano-electrospray emitter (100-micron ID) packed with 15 cm of C18 RP material and analyzed via 24-hr MudPIT 2D-LC-MS/MS as previously described (Abraham et al., 2012). Peptide sequencing analysis was performed with an LTQ-Orbitrap-Velos-Pro mass spectrometer (ThermoScientific). Data acquisition was managed by XCalibur version 2.1. Mass spectra were acquired in a data-dependent "top 20" mode: each survey scan (30,000 at m/z 400) was followed by MS/MS spectra of the ten most abundant precursor ions (3 *m/z* isolation window). For peptide fragmentation, charge state rejection of +1's was enforced for precursor selection and normalized collision energy of 35% was used for collision-induced dissociation (CID). Each fragmented precursor ion was dynamically excluded from targeting for 60 seconds. A dynamic exclusion repeat of 1 and an exclusion mass width of 0.20 were applied to maximize peptide sequencing.

Peptide identification

Experimental MS/MS spectra were searched against the transcriptome sequencing-derived (RNA-seq) proteome database (see Supplementary Note 5). In addition to the 70,257 representative protein sequences predicted in *A. americana*, the protein database was supplemented with proteins predicted in the *Agave* chloroplast genome (GenBank accession KX519714), and common contaminant proteins (i.e. trypsin and human keratin). A decoy database, consisting of the reversed sequences of the target database, was appended in order to discern the false-discovery rate (FDR) at the spectral level. For standard database searching, the peptide fragmentation spectra (MS/MS) were searched with MyriMatch algorithm v2.1 (Tabb et al., 2007). MyriMatch was configured to derive fully-tryptic peptides with the following

parameters: unlimited missed cleavages, max peptide length 75, minimum peptide length of 5 amino acids, maximum peptide mass of 10,000 Da, maximum number of charge states of 4, a precursor mass tolerance of 10 parts per million (ppm), a fragment mass tolerance of 0.5 m/z units, a static modification on cysteines (iodoacetamide; +57.0214 Da), dynamic modifications on the n-terminus (carbamylation; +43.0058) and methionine (oxidation; 15.9949). The raw spectrum files, peak list files, and result files have been made available through the mass spectrometry interactive virtual environment (MassIVE) as public resource. These data can be obtained using the following accessions: MassIVE accession MSV000079780 and ProteomeXchange accession PXD004239.

Protein inference and relative quantitation

Resulting peptide spectrum matches were imported, filtered and organized into protein identifications using IDPicker v.3.0 (Ma et al., 2009). Given the incompleteness of the database, proteins were only required to have a minimum of one distinct peptide match. To obtain an average FDR of 5% at the protein level for each measurement, we required a maximum FDR of 1% at the peptide spectrum match level and each peptide must have a minimum of two observations, rather than the traditional criteria of 1 spectra count per peptide.

To deal with the sequence redundancy associated with the *A. americana* protein database, all identified proteins were consolidated into groups by sequence similarity as previously described (Abraham et al., 2012). In brief, proteins in the FASTA database were grouped by sequence similarity (≥ 90%) using the UCLUST component of the USEARCH v. 5.0 software platform (Edgar, 2010). The uniqueness of each peptide was then classified as follows: (i) shared; (ii) database unique (i.e., peptides whose sequence matched only one protein); and (iii) protein-group unique (i.e., peptides whose sequence matched to multiple proteins, but only to a single protein group). These consolidated reports were instrumental in classifying the ambiguity of every identified protein during data analysis. A verbose listing of the protein groups, proteins, and peptides identified and their respective uniqueness to the reference database are deposited at MassIVE under the accession MSV000079780 and ProteomeXchange accession PXD004239.

For label-free quantification using spectra counts, summed protein spectral counts were converted to normalized spectral abundance factors (NSAF) (Zybailov et al., 2007). NSAF values were then multiplied by a value (*i.e.*, 100,000) for ease of data interpretation. To assess

reliable quantitative differences across the sample set, only those proteins with substantive abundance values, as determined by prevalence value (PV) (Lochner et al., 2011), were carried on to subsequent quantitative analyses. Similar to what was performed for transcript quantitation, each identified protein was given a PV, which was determined by averaging the adjusted NSAF values across all samples. Next, PVs were plotted as a histogram to graphically capture the distribution of abundances, such that one could assess the cumulative abundances assigned at varying PV cutoffs. An inflection point was identified, where proteins with a minimum average adjusted NSAF > 1.5 were considered to be highly representative and reproducible. The quantitative values were then log2-transformed.

Transcript and protein abundances related to the carboxylation and decarboxylation phase of CAM in Agave

A number of studies thus far have shown that CAM-specific *PEPC1* transcripts in a facultative CAM plant show higher abundance at night (Cushman et al., 2008), whereas in C_3 and C_4 plants, these transcripts are more abundant during the day (Chollet et al., 1996). Previous findings suggest that the PEPC activity is increased at night; however, its protein abundance pattern varies over the diel cycle in facultative CAM species (Häusler et al., 2000). In *Agave*, measured transcript and protein abundances of the reciprocal blast hit for *A. thaliana PEPC**1*** (Aam080248) show increased relative abundance during the day (Supplemental Figure 5A-B). However, a closer look at the NSAF measured abundance of PEPC1 shows that the relative abundance is nearly as high as RuBisCO large subunit and could therefore be outside the linear range of quantification for the label-free quantitative approach used here. Because measurements of mass spectral peak intensities can help quantify relative changes in protein abundances with high spectral counts, we evaluated the overall spectral ion intensity for PEPC1. Because proteins with a large number of spectral counts can fall outside the linear range of quantitation, summed fragment ion intensities were calculated for peptides belonging to PEPC1 (Aam080248). The matched fragment ion intensities (MIT) for each peptide spectrum match were collected directly from their corresponding mzML files and summed together to calculate each peptide's MIT. A normalized quantitative value was then calculated as follows: LOESS regression normalization was applied across replicates and median absolute deviation regression (MAD) (Callister et al., 2006) and central tendency (mean) was applied across the sample set. A final protein abundance

was determined using the Qrollup approach in InfernoRDN software by using the top 33% normalized peptide intensities for each protein (Polpitiya et al., 2008). As shown in Supplemental Figure 5B, when using MIT, we observed fluctuation in the abundance of PEPC1 across the 24-hr period, with this protein predominately abundant at the end of the photoperiod.

Malic acid accumulates as a consequence of nocturnal carboxylation and can be subsequently remobilized the following day to release CO_2 for the Calvin-Benson cycle plus pyruvate, which is recycled by gluconeogenesis *via* ***pyruvate orthophosphate dikinase* (*PPDK*)** (Aam010102). Therefore, we sought to analyze proteins related to this crucial transition in diel carbon metabolism in *Agave*. We observed that the expression of the *PPDK* transcript and PPDK protein were largely coincident with one another while peaking in the morning, consistent with a role in the decarboxylation of malic acid during the early morning hours in *Agave* (Supplemental Figure 5C). As anticipated, the protein responsible for down-regulating the activity of PPDK, ***PPDK-regulatory protein* (*RP1*)** (Aam051010), reaches peak abundance at night in *Agave*, yet has an abundance profile reciprocal to that of its transcript (Supplemental Figure 5D).

Chloroplast- and Mitochondrial-related transcript and protein abundance relationships— Because respiration and photosynthesis are intimately linked, manipulation of one must be undertaken with consideration of the effects on the relative activity of the other. Therefore, we highlighted temporal abundance relationships of transcripts and proteins for several key genes in the mitochondria and chloroplast (Supplemental Figure 6). In plant tissues, particularly leaves, mitochondria exhibit extensive flexibility for modulating cellular redox and carbohydrate homeostasis. In CAM plants, as alluded to above, the mitochondria accommodate high C fluxes and electron transport at night. The core elements of the TCA are present in plant mitochondria. Although the protein abundances of the enzymes controlling two key TCA reactions (*i.e.*, ***isocitrate dehydrogenase* (*IDH*)** (Aam010083) and ***fumarase***) are in phase in *Agave*, these enzymes differ in their transcript and protein abundance relationships. For example, the abundances of two unambiguously identified fumarase proteins encoded by ***FUM1*** (Aam085348) and ***FUM2*** (Aam045332) are both in opposite phase to those of their transcripts. Interestingly, reducing the activity of the TCA cycle enzyme fumarase via anti-sense technology had more dramatic and detrimental effects on photosynthesis (Nunes-Nesi et al., 2007) than did reducing malate dehydrogenase activity in the same manner (Nunes-Nesi et al., 2005). Therefore,

603 fumarase along with several other TCA enzymes are highly regulated and play key roles in
604 modulating respiratory carbon flux (Araujo et al., 2012). Fluctuating in parallel with the
605 abundance profiles of the TCA cycle enzymes are those of proteins involved in respiration,
606 which generates ATP by using the reducing equivalents derived from the operation of the TCA
607 cycle. The respiratory pathway is extremely sensitive to redox changes and we observed parallel
608 increases in the abundances of major intracellular antioxidant enzymes in *Agave*, such as
609 ***monothiol glutaredoxin* (*GRX4*)** (Aam313181), which decreases the concentration of
610 detrimental reactive oxygen species (ROS) (Cheng, 2008). Given its important protein-protective
611 role (Herrero and de la Torre-Ruiz, 2007), the aligned transcript and protein phases for GRX4
612 suggest a rapid response mechanism. Interestingly, in the present study, the abundance of the
613 GRX4 protein was found to be temporally associated with the protein abundance of ***BolA4***
614 (Aam075423), which physically interacts with monothiol GRXs in C_3 plants and is regulated in a
615 redox-controlled manner (Couturier et al., 2014).
616 The substantial reciprocal relationship between the daily transfer of carbon between acids
617 and carbohydrates that defines CAM involves extensive and regulated transport of metabolites
618 between chloroplasts, vacuoles, the cytosol and mitochondria. We observed similar protein
619 abundance phases for the major mitochondrial trafficking proteins ***ADP/ATP carrier 1* (*AAC1*)**
620 (Aam043344) and ***mitochondrial pyruvate carrier* (*MPC1*)** (Aam302119) with greater protein
621 abundance of both during the light or early evening period in *Agave*. Pyruvate occupies a pivotal
622 role in the regulation of CAM and the abundance of MPC1 appears to be regulated at the post-
623 transcriptional level or beyond.
624 Chloroplasts carry out photosynthesis, as well as a multitude of other functions. The
625 primary light-driven reactions of photosynthesis occur in the thylakoid membranes and are
626 mediated by the multi-component protein complexes, photosystem II (PSII) and photosystem I
627 (PSI). PSI utilizes light for electron transport through a series of redox centers to reduce
628 ferredoxin, and provides electrons in a variety of chloroplast reactions, whereas PSII harvests
629 and transfers light energy while concomitantly converting water to molecular oxygen. Given
630 their importance, we investigated the temporal dynamics of the transcripts and proteins
631 associated with PSI and PSII. The accumulation of PSII and PSI light-driven subunits coincides
632 with the light period. During the light and early dark period, we observed peak abundance of
633 transcripts encoding a ***light harvesting complex photosystem ii subunit 6* (*Lhcb6*)** (Aam047736)

and a gene encoding the protein reaction center of PSI **(*PsaA*)** (Aam004267). Most interestingly, we observed nocturnal phase abundance increases for two extrinsic proteins related to the oxygen evolving complex of PSII: one gene encoding a member of the ***photosystem ii reaction center family*** **(*PsbP*)** (Aam075610) and the other gene encoding part of the oxygen-evolving complex, ***photosystem ii subunit q*** **(*PsbQ*)** (Aam038462). In *Arabidopsis*, these extrinsic proteins are categorized into one of the three groups: the oxygen evolving complex group for water splitting, the group involved in cyclic electron transport around PSI, and a stress-responsive group. Based on their protein abundance profiles, we suspect that these two highlighted proteins could belong to the latter two groups.

The ATP and NADPH subsequently generated from these light-driven reactions are consumed by the Calvin-Benson cycle in a series of enzyme-driven reactions to transform CO_2 into organic compounds that are compatible with the needs of the cell. In all plants, the first step of the Calvin-Benson cycle is catalyzed by RuBisCO. In Supplemental Figure 6, the abundance phases of transcripts and protein in *Agave* for RuBisCO ***activase*** **(*RCA*)** (Aam041100), which helps convert RuBisCO from its inactive to its active conformational state by reducing RuBisCO's binding affinity for sugar phosphates, is shown. Rather than being transcribed at dawn as in *Arabidopsis* (Pilgrim and McClung, 1993), the transcript and protein abundances for RuBisCO activase start to increase during the middle of the light period, and peak during the dark period. The offset abundance of RuBisCO activase abundance relative to *Arabidopsis* could be attributed to CAM-specific phase-dependent changes in the intracellular CO_2 concentration as reported elsewhere (Maxwell et al., 1999).

7. Tests for differential transcript and protein abundance

Given the number of post-processing steps for identified transcripts and proteins prior to testing for differential abundance, we provide a brief summary and discussion of the rational for each step. After identifying transcript and proteins, both datasets experience several post-processing steps prior to testing for differential abundances (Supplemental Figure 8; this workflow covers post-raw data processing for the transcript and protein identifications, which were covered in Supplemental Notes 5 and 6, respectively). Normalization is a well-known, regularly utilized and critical step prior to differential analysis to mitigate the potential bias that may confound the results. For in-depth quantification, the transcript and protein abundances were

normalized to regularly utilized RPKM and NSAF values, respectively. Next, care was taken to remove transcripts and proteins that are below a limit of quantification (i.e., prevalence value), which was calculated by averaging the normalized values across all samples, plotting a histogram to graphically capture the distribution of abundances, and assessing the cumulative abundances assigned at varying prevalence value cutoffs. An inflection point was identified for both datasets, as is noted in the above sections. Given the dynamic range of both datasets, quantitative values were then log2-transformed. For the purposes of the tests for differential abundance, missing values were imputed by drawing random numbers from a normal distribution to simulate signals from low-abundance transcripts or proteins, using the freely available software Perseus (http://www.perseus-framework.org). The width parameter of this normal distribution was chosen as 0.3 of the standard deviation of all measured values and the center was shifted towards low abundance by 1.8 times this standard deviation.

For this study, we performed pair-wise comparisons of time points as our hypothesis is concerned with the among different time-points and not the overall change in transcripts and proteins. To this end, we employed two approaches for each dataset. First, a paired t-test was utilized to identify differences in quantitative abundances between time points using JMP Genomics software v. 6.0 (SAS Institute). Moreover, to assess statistical confidence based on the 28 pair-wise comparisons, the Benjamini-Hochberg method was applied to provided adjusted p-values (Supplemental Table 5 and Supplemental Table 14). To provide a more robust statistical assessment and another perspective of the statistical confidence, the differential analysis of the transcripts and proteins above the limits of quantification was also assessed using the ranked fold change method as described in (Dembele and Kastner, 2014), and implemented in R. Again, fold changes were determined by pairwise comparisons among all time points. For the transcript data, instead of using RPKM normalized values, the voom method (Law et al., 2014) was used to account for the mean-variance relationship of the transcript count data, while requiring at least 3 non-zero counts per transcript. For the protein data, instead of using NSAF values, the protein data was quantile normalized prior to calculating the fold change rank estimates, requiring at least 3 non-zero counts per protein. For both datasets, the family wise error rate was controlled with the use of the Bonferroni correction in order to adjust for multiple hypothesis bias across the intervals compared ($\alpha = 0.001$) and the statistical confidences have been noted in Supplemental Table 5 and Supplemental Table 14.

To identify and illustrate intensity-independent patterns, quantitative values across biological replicates were averaged and then transformed to a z-score: standard deviations from the mean expression [(abundance - mean)/ SD] were calculated for each transcript and protein. To capture general patterns without considering absolute expression levels, z-scores were then loaded into Multi Experiment Viewer software (MeV v. 4.9) (Saeed et al., 2003) and the Figure of Merit (FOM) algorithm was used to estimate an appropriate number of clusters (Yeung et al., 2001)for the transcript and protein of interest. K-means support using Pearson's correlation was then used to separate groups of co-abundant transcripts/proteins.

8. Gene Ontology enrichment

Whole genome gene ontology (GO) term annotation was performed using Blast2GO (Conesa et al., 2005) with a blastp E-value hit filter of 1 × 10-6, an annotation cutoff value of 55 and a GO weight of 5. Using ClueGO (Bindea et al., 2009), observed GO biological process were subjected to the right-sided hypergeometric enrichment test at medium network specificity selection and *p*-value correction was performed using the Holm-Bonferroni step-down method (Holm, 1979). There were a minimum of 3 and a maximum of 8 selected GO tree levels, while each cluster was set to include a minimum of between 3% and 4% of genes associated with each term. GO term fusion and grouping settings were selected to minimize GO term redundancy and the term enriched at the highest level of significance was used as the representative term for each functional cluster. The GO terms with *p*-values less than or equal to 0.05 were considered significantly enriched.

References:

1. Abraham, P., Adams, R., Giannone, R.J., Kalluri, U., Ranjan, P., Erickson, B., . . . Hettich, R.L. (2012) Defining the boundaries and characterizing the landscape of functional genome expression in vascular tissues of Populus using shotgun proteomics. *J Proteome Res* 11, 449-460.
2. Abraham, P., Giannone, R.J., Adams, R.M., Kalluri, U., Tuskan, G.A., and Hettich, R.L. (2013) Putting the pieces together: high-performance LC-MS/MS provides network-, pathway-, and protein-level perspectives in Populus. *Mol Cell Proteomics* 12, 106-119.
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* 215, 403-410.
4. Araujo, W., Nunes-Nesi, A., Nikoloski, Z., Sweetlove, L., and Fernie, A. (2012) Metabolic control and regulation of the tricarboxylic acid cycle in photosynthetic and heterotrophic plant tissues. *Plant Cell and Environment* 35, 1-21.
5. Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., . . . Galon, J. (2009) ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* 25, 1091-1093.
6. Callister, S.J., Barry, R.C., Adkins, J.N., Johnson, E.T., Qian, W.J., Webb-Robertson, B.J., . . . Lipton, M.S. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5, 277-286.
7. Castro-Camus, E., Palomar, M., and Covarrubias, A.A. (2013) Leaf water dynamics of Arabidopsis thaliana monitored in-vivo using terahertz time-domain spectroscopy. *Sci Rep* 3, 2910.
8. Cheng, N.H. (2008) AtGRX4, an Arabidopsis chloroplastic monothiol glutaredoxin, is able to suppress yeast grx5 mutant phenotypes and respond to oxidative stress. *Febs Lett* 582, 848-854.
9. Cheung, C.Y., Poolman, M.G., Fell, D.A., Ratcliffe, R.G., and Sweetlove, L.J. (2014) A Diel Flux Balance Model Captures Interactions between Light and Dark Metabolism during Day-Night Cycles in C3 and Crassulacean Acid Metabolism Leaves. *Plant Physiol* 165, 917-929.
10. Chollet, R., Vidal, J., and O'Leary, M. (1996) Phosphoenolpyruvate carboxylase: a ubiquitous, highly regulated enzyme in plants. *Annual Review of Plant Physiology and Plant Molecular Biology* 47, 273–298.
11. Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674-3676.
12. Couturier, J., Wu, H., Dhalleine, T., Pégeot, H., Sudre, D., Gualberto, J., . . . Rouhier, N. (2014) Monothiol glutaredoxin–BolA interactions: redox control of *Arabidopsis thaliana* BolA2 and SufE1. *Molecular Plant* 7, 187–205.
13. Cushman, J., Tillett, R., Wood, J., Branco, J., and Schlauch, K. (2008) Large-scale mRNA expression profiling in the common ice plant, *Mesembryanthemum crystallinum*, performing C_3 photosynthesis and Crassulacean acid metabolism (CAM). *J. Exp. Bot.* 59, 1875-1894.

14. Dembele, D. and Kastner, P. (2014) Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *Bmc Bioinformatics* 15.
15. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26, 2460-2461.
16. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28, 3150-3152.
17. 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. *Nat Biotechnol* 29, 644-652.
18. Häusler, R., Baur, B., Scharte, J., Teichmann, T., Eicks, M., Fischer, K., . . . Fischer, K. (2000) Plastidic metabolite transporters and their physiological functions in the inducible crassulacean acid metabolism plant *Mesembryanthemum crystallinum*. *Plant J.* 24, 285-296.
19. Herrero, E. and de la Torre-Ruiz, M.A. (2007) Monothiol glutaredoxins: a common domain for multiple functions. *Cell Mol Life Sci* 64, 1518-1530.
20. Holm, S. (1979) A Simple Sequentially Rejective Multiple Test Procedure. *Scand J Stat* 6, 65-70.
21. Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res* 9, 868-877.
22. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10, R25.
23. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014) voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 15, R29.
24. Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *Bmc Bioinformatics* 12.
25. Li, L., Stoeckert, C.J., Jr., and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13, 2178-2189.
26. Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658-1659.
27. Li, Y., Tschaplinski, T.J., Engle, N.L., Hamilton, C.Y., Rodriguez, M., Jr., Liao, J.C., . . . Graham, D.E. (2012) Combined inactivation of the Clostridium cellulolyticum lactate and malate dehydrogenase genes substantially increases ethanol yield from cellulose and switchgrass fermentations. *Biotechnol Biofuels* 5, 2.
28. Lochner, A., Giannone, R.J., Keller, M., Antranikian, G., Graham, D.E., and Hettich, R.L. (2011) Label-free quantitative proteomics for the extremely thermophilic bacterium Caldicellulosiruptor obsidiansis reveal distinct abundance patterns upon growth on cellobiose, crystalline cellulose, and switchgrass. *J Proteome Res* 10, 5302-5314.
29. Ma, Z.Q., Dasari, S., Chambers, M.C., Litton, M.D., Sobecki, S.M., Zimmerman, L.J., . . . Tabb, D.L. (2009) IDPicker 2.0: Improved protein assembly with high discrimination peptide identification filtering. *J Proteome Res* 8, 3872-3881.
30. Maxwell, K., Borland, A.M., Haslam, R.P., Helliker, B.R., Roberts, A., and Griffiths, H. (1999) Modulation of Rubisco Activity during the Diurnal Phases of the Crassulacean Acid Metabolism Plant Kalanchoe daigremontiana. *Plant Physiol* 121, 849-856.

31. Mironova, V., Omelyanchuk, N., Wiebe, D., and Levitsky, V. (2014) Computational analysis of auxin responsive elements in the *Arabidopsis thaliana* L. genome. *BMC genomics* 15, S4.
32. Mockler, T.C., Michael, T.P., Priest, H.D., Shen, R., Sullivan, C.M., Givan, S.A., . . . Chory, J. (2007) The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. *Cold Spring Harb Symp Quant Biol* 72, 353-363.
33. Nunes-Nesi, A., Carrari, F., Gibon, Y., Sulpice, R., Lytovchenko, A., Fisahn, J., . . . Fernie, A.R. (2007) Deficiency of mitochondrial fumarase activity in tomato plants impairs photosynthesis via an effect on stomatal function. *Plant J* 50, 1093-1106.
34. Nunes-Nesi, A., Carrari, F., Lytovchenko, A., Smith, A.M.O., Loureiro, M.E., Ratcliffe, R.G., . . . Fernie, A.R. (2005) Enhanced photosynthetic performance and growth as a consequence of decreasing mitochondrial malate dehydrogenase activity in transgenic tomato plants. *Plant Physiology* 137, 611-622.
35. Pilgrim, M. and McClung, C. (1993) Differential involvement of the circadian clock in the expression of genes required for ribulose-1, 5-bisphosphate carboxylase/oxygenase synthesis, assembly, and activation in *Arabidopsis thaliana*. *Plant Physiology* 103, 553–564.
36. Polpitiya, A.D., Qian, W.J., Jaitly, N., Petyuk, V.A., Adkins, J.N., Camp, D.G., 2nd, . . . Smith, R.D. (2008) DAnTE: a statistical tool for quantitative analysis of -omics data. *Bioinformatics* 24, 1556-1558.
37. Saeed, A., Sharov, V., White, J., Li, J., Liang, W., Bhagabati, N., . . . Quackenbush, J. (2003) TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374–378.
38. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., . . . Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13, 2498-2504.
39. Tabb, D.L., Fernando, C.G., and Chambers, M.C. (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J Proteome Res* 6, 654-661.
40. Tschaplinski, T.J., Standaert, R.F., Engle, N.L., Martin, M.Z., Sangha, A.K., Parks, J.M., . . . Mielenz, J.R. (2012) Down-regulation of the caffeic acid O-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnol Biofuels* 5, 71.
41. Van Dongen, S.M. (2001) Graph clustering by flow simulation.
42. Wang, L., Si, Y., Dedow, L.K., Shao, Y., Liu, P., and Brutnell, T.P. (2011) A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq. *PLoS One* 6, e26426.
43. Wattam, A.R., Abraham, D., Dalay, O., Disz, T.L., Driscoll, T., Gabbard, J.L., . . . Sobral, B.W. (2014) PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42, D581-591.
44. Winter, K. and Smith, J.A.C. (1996) Crassulacean acid metabolism: Current status and perspectives. *Ecol Stu An* 114, 389-426.
45. Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., and Ruzzo, W.L. (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics* 17, 977-987.

853 46. Zybailov, B.L., Florens, L., and Washburn, M.P. (2007) Quantitative shotgun
854 proteomics using a protease with broad specificity and normalized spectral
855 abundance factors. *Molecular Biosystems* 3, 354-360.
856