

# Lawrence Berkeley National Laboratory

## LBL Publications

### Title

The *Kalanchoë* genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism

### Permalink

<https://escholarship.org/uc/item/81d1b1rh>

### Journal

Nature Communications, 8(1)

### ISSN

2041-1723

### Authors

Yang, Xiaohan

Hu, Rongbin

Yin, Hengfu

et al.

### Publication Date

2017

### DOI

10.1038/s41467-017-01491-7


Peer reviewed

ARTICLE

DOI: 10.1038/s41467-017-01491-7

OPEN

# The *Kalanchoë* genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism

Xiaohan Yang et al.<sup>#</sup> 

Crassulacean acid metabolism (CAM) is a water-use efficient adaptation of photosynthesis that has evolved independently many times in diverse lineages of flowering plants. We hypothesize that convergent evolution of protein sequence and temporal gene expression underpins the independent emergences of CAM from C<sub>3</sub> photosynthesis. To test this hypothesis, we generate a de novo genome assembly and genome-wide transcript expression data for *Kalanchoë fedtschenkoi*, an obligate CAM species within the core eudicots with a relatively small genome (~260 Mb). Our comparative analyses identify signatures of convergence in protein sequence and re-scheduling of diel transcript expression of genes involved in nocturnal CO<sub>2</sub> fixation, stomatal movement, heat tolerance, circadian clock, and carbohydrate metabolism in *K. fedtschenkoi* and other CAM species in comparison with non-CAM species. These findings provide new insights into molecular convergence and building blocks of CAM and will facilitate CAM-into-C<sub>3</sub> photosynthesis engineering to enhance water-use efficiency in crops.

---

Correspondence and requests for materials should be addressed to X.Y. (email: [yangx@ornl.gov](mailto:yangx@ornl.gov))

<sup>#</sup>A full list of authors and their affiliations appears at the end of the paper

C<sub>3</sub> rassulacean acid metabolism (CAM) is a metabolic adaptation of photosynthetic CO<sub>2</sub> fixation that enhances plant water-use efficiency (WUE) and associated drought avoidance/tolerance by reducing transpirational water loss through stomatal closure during the day, when temperatures are high, and stomatal opening during the night, when temperatures are lower<sup>1</sup>. In the face of the rapidly increasing human population and global warming predicted over the next century, the outstanding WUE of CAM plants highlights the potential of the CAM pathway for sustainable food and biomass production on semi-arid, abandoned, or marginal agricultural lands<sup>2–4</sup>.

CAM photosynthesis can be divided into two major phases: (1) nocturnal uptake of atmospheric CO<sub>2</sub> through open stomata and primary fixation of CO<sub>2</sub> by phosphoenolpyruvate carboxylase (PEPC) to oxaloacetate (OAA) and its subsequent conversion to malic acid by malate dehydrogenase; and (2) daytime decarboxylation of malate and CO<sub>2</sub> refixation via C<sub>3</sub> photosynthesis, mediated by ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO)<sup>5,6</sup>. Malic acid is stored in the vacuole of photosynthetically active cells reaching a peak at dawn and can be used as a reference point to divide the two phases. CAM is found in over 400 genera across 36 families of vascular plants<sup>4</sup> and is thought to have evolved multiple times independently from diverse ancestral C<sub>3</sub> photosynthesis lineages<sup>7</sup>. The core biochemical characteristics of the CAM cycle are similar in all the plant lineages in which CAM has evolved, with some variation in the enzymes that catalyze malate decarboxylation during the day, and in the storage carbohydrates that provide substrates for malic acid synthesis at night<sup>8,9</sup>.

We hypothesize that convergent evolution in protein sequence and/or temporal diel gene expression underpins the multiple and independent emergences of CAM from C<sub>3</sub> photosynthesis. Convergent evolution is generally defined as the appearance of similar phenotypes in distinct evolutionary lineages<sup>10</sup>. Although phenotypic convergence is widely recognized, its evolutionary mechanism has been extensively debated. Morris<sup>11</sup> argues that the evolutionary course is not random but selection-constrained, along certain pathways, to arrive at the same solution or outcome. Recently, comparative genomics analysis began to provide new insight into the molecular mechanism of convergent evolution. For example, Foote et al.<sup>12</sup> performed comparative genomic analyses of three species of marine mammals (the killer whale, walrus, and manatee) that share independently evolved phenotypic adaptations to a marine existence, and identified convergent amino-acid substitutions in genes evolving under positive selection and putatively associated with a marine phenotype. Also, Hu et al.<sup>13</sup> compared the genomes of the bamboo-eating giant and red pandas, two obligate bamboo-feeders that independently possess adaptive pseud thumbs, and identified 70 adaptively convergent genes (i.e., under positive selection in these two species), of which nine genes, featuring nonrandom convergent amino-acid substitution between giant and red pandas, are closely related to limb development and essential nutrient utilization. These two examples indicate that specific amino-acid replacements at a small number of key sites can result in highly predictable convergent outcomes, supporting the constrained selection theory of Morris<sup>11</sup>. However, such predictable protein sequence convergence was not found in the convergence of hemoglobin function in high-altitude-dwelling birds, indicating that possible adaptive solutions are perhaps contingent upon prior evolutionary history<sup>14</sup>. This finding supports the contingent adaptation theory<sup>15</sup> that evolution is contingent upon history and consequently replaying life's tape will give different outcomes. In addition to protein sequence convergence, convergent changes in gene expression were found to be associated with convergent evolution of vocal learning in the brains of humans and song-

learning birds<sup>16</sup>. Therefore, convergent changes in both protein sequence and gene expression are important aspects of the molecular basis of convergent evolution.

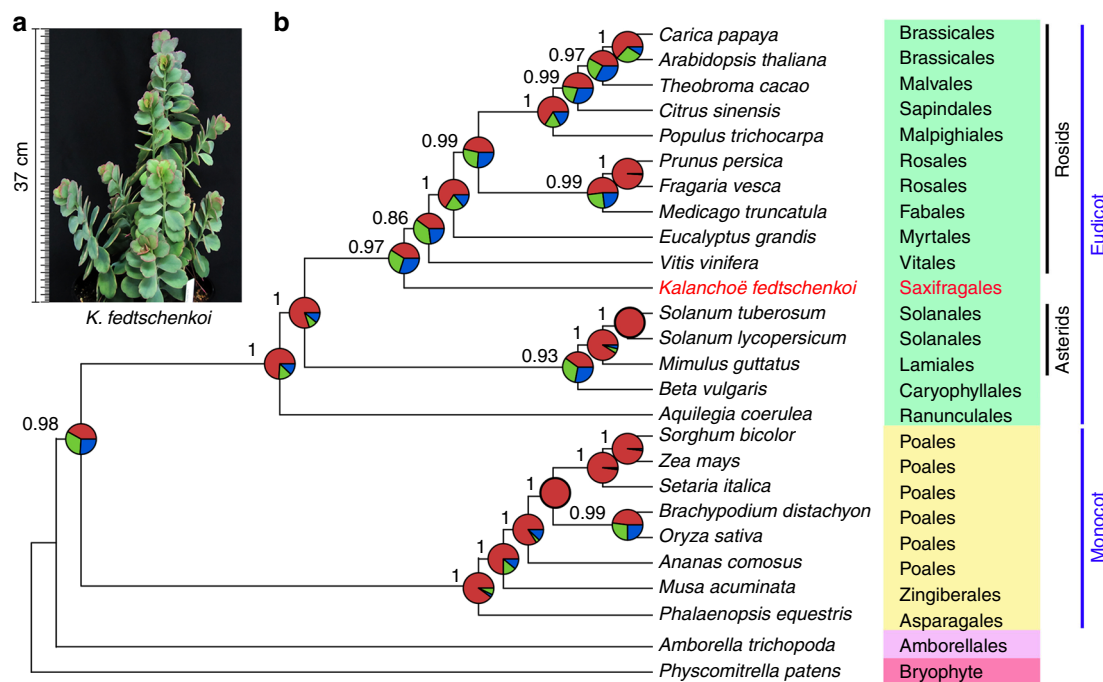
We sought to investigate whether changes in protein sequence and/or gene expression contribute to the evolutionary convergence of CAM through genome-wide screening for signatures of convergent changes in protein sequences and diel mRNA expression patterns that meet the following criteria: the signatures are (1) isomorphic in the CAM genomes of distant groups, such as eudicots and monocots, which diverged ~135 million years ago<sup>17</sup>, and (2) dimorphic in related C<sub>3</sub> photosynthesis genomes. Recently, the genome sequences of two monocot CAM species, *Ananas comosus* (L.) Merr. (pineapple)<sup>18</sup>, and *Phalaenopsis equestris* (Schauer) Rchb.f. (moth orchid)<sup>19</sup>, were published. Here we present the genome sequence of *Kalanchoë fedtschenkoi* Raym.-Hamet & H. Perrier, which is an emerging molecular genetic model for obligate CAM species in the eudicots<sup>4,6,20</sup>. Our analyses reveal the genomic signatures of convergence shared between eudicot (represented by *Kalanchoë*) and monocot (represented by pineapple and orchid) CAM species.

## Results

***Kalanchoë* genome assembly and annotation.** The diploid *K. fedtschenkoi* (2n = 2x = 34 chromosomes; Supplementary Fig. 1) genome size was estimated to be ~260 Mb (Supplementary Table 1). The *K. fedtschenkoi* genome was assembled from ~70× paired-end reads and ~37× mate-pair reads generated using an Illumina MiSeq platform (Supplementary Table 2 and Supplementary Fig. 2). The genome assembly consisted of 1324 scaffolds with a total length of 256 Mb and scaffold N50 of 2.45 Mb (Supplementary Table 3), in which we predicted and annotated 30,964 protein-coding genes (Supplementary Table 4).

**The phylogenetic placement of *Kalanchoë*.** *Kalanchoë* is the first eudicot CAM lineage with a genome sequence to date and serves as an important reference for understanding the evolution of CAM. In addition, *K. fedtschenkoi* is the first sequenced species in the distinct eudicot lineage, Saxifragales. Although the monophyly of this morphologically diverse order is well supported by molecular data, its phylogenetic placement has been less clear<sup>21,22</sup>. The recent consensus view, based mainly on analyses of plastid DNA sequences, has placed the Saxifragales as a sister group to the rosids, and together they comprise the large clade of superrosids<sup>23,24</sup>. However, there have been indications of conflict between trees based on plastid genomes and nuclear genomes for this clade<sup>19,24</sup>. Additionally, the major lineages of core eudicots are thought to have diversified rapidly following their first appearance, making resolution of the relationships among these clades particularly challenging<sup>17,25</sup> and implicating incomplete lineage sorting (ILS) as a potentially important process that would result in discordance among gene histories<sup>26</sup>.

We performed phylogenetic analyses with 210 single-copy nuclear genes from 26 sequenced plant genomes using multiple phylogenetic inference strategies. The resulting species trees are congruent with each other except for the placement of *K. fedtschenkoi*, which was placed either as sister to the rosids in a phylogenetic tree reconstructed using a quartet-based coalescent species tree method (Fig. 1) or as sister to all other core eudicots as revealed by alternative phylogenetic trees reconstructed from (1) concatenated protein sequence alignment without gene partition using maximum-likelihood (Supplementary Fig. 3), (2) a partitioned analysis of multi-gene alignment using maximum-likelihood and Bayesian methods (Supplementary Fig. 4), and (3) analysis of individual gene trees using fully Bayesian multispecies coalescent method (Supplementary Fig. 5). Despite substantial



**Fig. 1** A species tree reconstructed from 210 single-copy genes using a summary method. **a** Diploid plant of *Kalanchoë fedtschenkoi*. **b** Individual maximum-likelihood gene trees were reconstructed from the CDS alignments for each of the 210 single-copy-gene ortholog groups using RAxML<sup>78</sup>, and the species tree was summarized from the gene trees using ASTRAL-II<sup>79</sup>. Pie graphs on nodes represent the proportion of gene trees that support the various quartets at every node, with red for the main topology shown in this tree, blue for the first alternative, and green for the second alternative, respectively. Quartet frequencies displayed in pie graphs and the posterior-probability at each node are calculated by ASTRAL-II<sup>79</sup>

discordance among estimated nuclear gene trees, the coalescence-based tree was consistent with the results of the plastome-based analyses, placing *Kalanchoë* as sister to the rosids (Fig. 1). Coalescent species tree estimation can account for gene tree discordance due to ILS<sup>27</sup>. At the same time, alternative placements of *Kalanchoë* as sister to the asterids, or as sister to all other core eudicots were observed in many gene trees (Fig. 1 and Supplementary Fig. 5). Gene tree discordance due to rapid diversification early in eudicot history has also been characterized by others<sup>24</sup>. Regardless of the optimal placement of the Saxifragales, including *Kalanchoë*, individual gene trees will often have alternative histories due to ILS in the face of rapid species diversification.

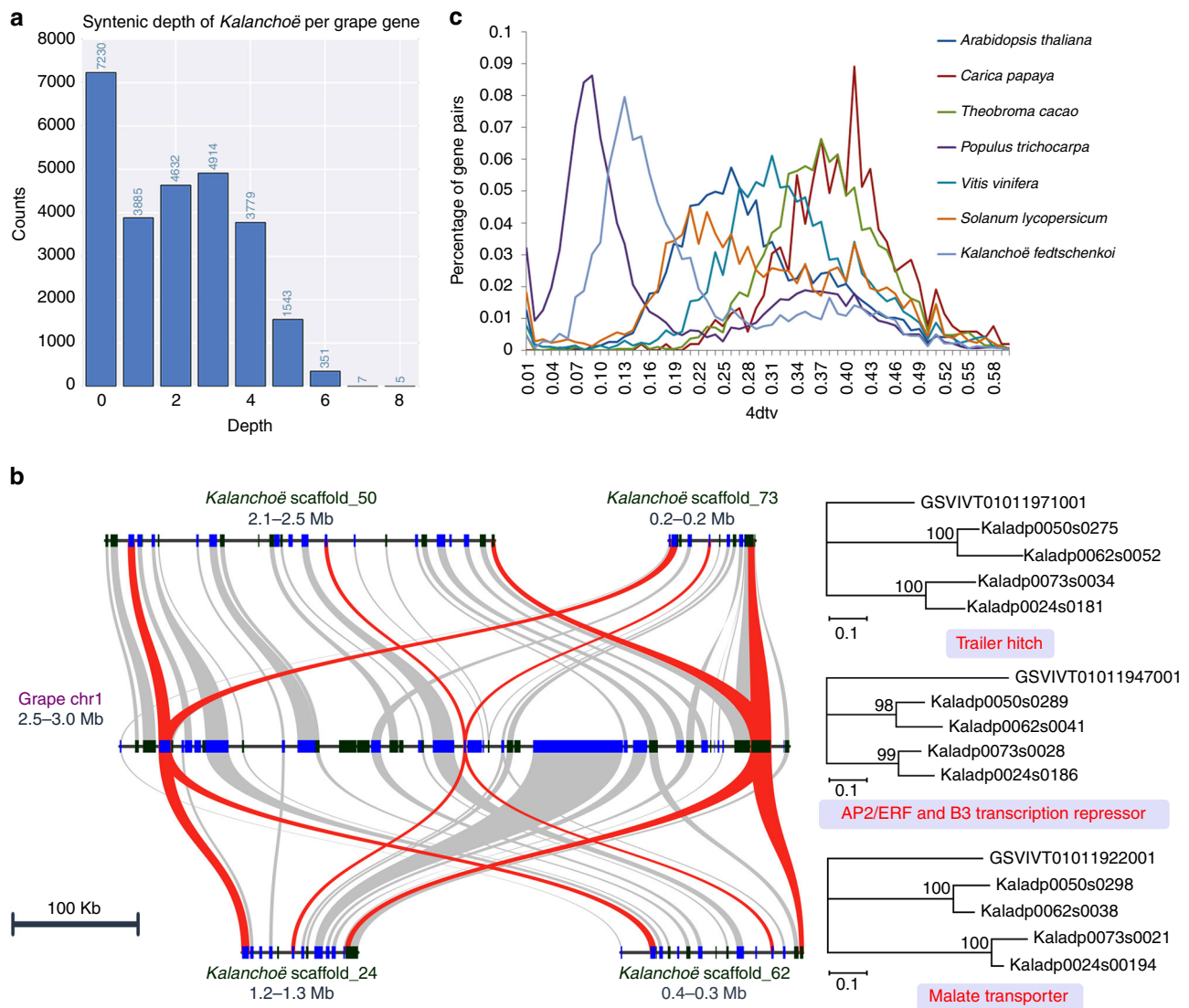
***Kalanchoë* genome duplication.** The grape genome has no additional genome duplication after the ancestral gamma hexaploidization<sup>28,29</sup> and is the best available reference for studying ancestral eudicot genome duplication events. Syntenic depth analyses<sup>30,31</sup> showed that there are multiple *K. fedtschenkoi* blocks covering each grape gene (Fig. 2a and Supplementary Fig. 6). Specifically, 65% of the grape genome had from one to four syntenic blocks in *K. fedtschenkoi*. In contrast, a sudden drop in syntenic depth occurred after a depth of 4× (Fig. 2a), indicating that each grape genome region has up to four *K. fedtschenkoi* blocks and thus providing strong evidence for two distinct whole-genome duplications (WGDs) events in *K. fedtschenkoi*. The microsynteny patterns further support two WGDs on the lineages leading to *K. fedtschenkoi*. Specifically, the microsynteny pattern reflects a 1:4 gene copy ratio between the grape genome and the diploid *K. fedtschenkoi* genome (Fig. 2b).

From the *Kalanchoë* point of view, we found that 49% of the *Kalanchoë* genome was covered by one grape-*Kalanchoë* block, 7% covered in two grape-*Kalanchoë* blocks, and 1% covered in

three grape-*Kalanchoë* blocks (Supplementary Fig. 7). This suggests that we could often find one best grape-*Kalanchoë* block out of the three gamma triplicated regions in grape. This fits the scenario that the gamma WGD predated the divergence and there has been no WGD in the grape lineage since grape-*Kalanchoë* diverged. Alternatively, if the divergence predated the gamma WGD, then from the *Kalanchoë* point of view we should instead see three matching grape regions. Hence, the grape-*Kalanchoë* genome comparisons strongly supported the gamma WGD as a shared event, and further supported the phylogenetic position of *Kalanchoë* in Fig. 1.

Despite two apparent WGDs in the *K. fedtschenkoi* lineage, synonymous substitutions per synonymous site ( $K_s$ ) between duplicate gene pairs showed only one prominent peak  $\sim 0.35$  (Supplementary Fig. 8). The unimodal distribution of  $K_s$  suggests the two WGD events occurring close in time. Similarly, two distinct peaks appear in the distribution of the four-fold transversion substitution rate (4dvt) values between the *K. fedtschenkoi* gene pairs (Fig. 2c). Grape-*Kalanchoë* gene pairs show a prominent peak around  $K_s = 1.5$  (Supplementary Fig. 8), indicating that the WGDs in the *K. fedtschenkoi* lineage occurred well after its divergence from grape early in the history of the rosid lineage.

**Gene co-expression modules and clusters in *Kalanchoë*.** To elucidate gene function in *K. fedtschenkoi*, we performed a weighted correlation network analysis of transcript expression in 16 samples including 12 mature leaf samples collected every 2 h over a 24-h period and four non-leaf samples collected 4 h after the beginning of the light period, including shoot tip (leaf pair 1 plus the apical meristem), stem (between leaf pair 3 and leaf pair 8), root, and flower. Our analysis identified 25 co-expression modules, among which one module (MEblack containing 782

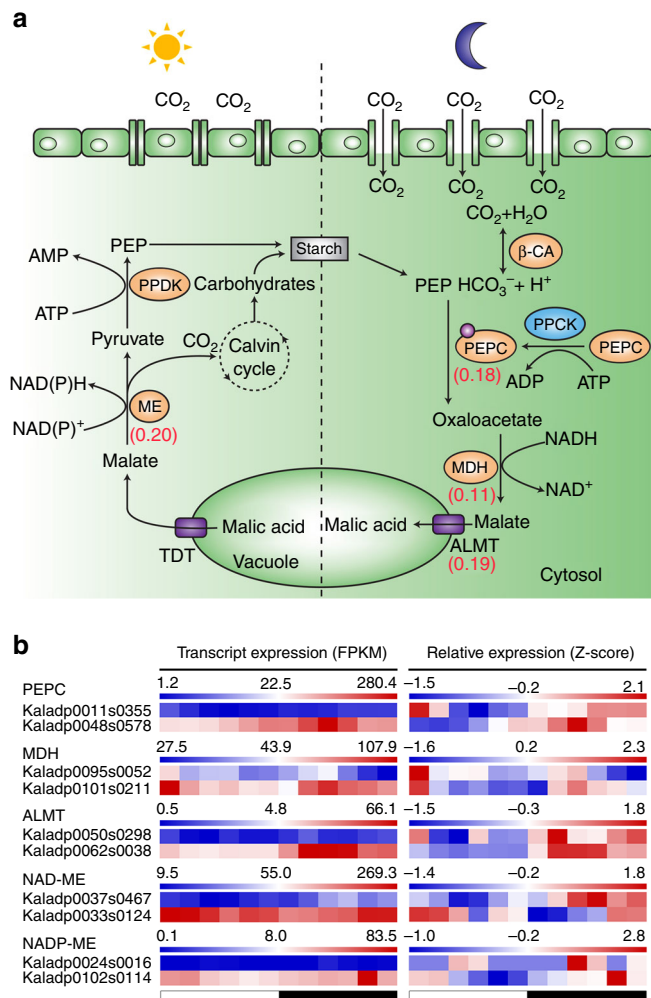


**Fig. 2** Genome duplication in *Kalanchoë fedtschenkoi*. **a** Syntenic depth of the *K. fedtschenkoi* genome for each grape gene. Syntenic depth refers to the number of times a genomic region is covered by synteny blocks against another genome. **b** Typical micro-colinearity patterns between genomic regions from grape and *K. fedtschenkoi*. Rectangles show predicted gene models with colors showing relative orientations (blue: same strand, black: opposite strand). Matching gene pairs are displayed as connecting shades. Three orthologous gene groups that were maximally retained as four copies in *K. fedtschenkoi* were highlighted with phylogenetic trees on the right suggesting two rounds of genome duplications in the *Kalanchoë* lineage. **c** Four-fold transversion substitution rate (4dtv) in *K. fedtschenkoi* and six other eudicot plant species

genes) was significantly (Student's *t*-test,  $P < 0.001$ ) associated with the leaf samples collected during the dark period (Supplementary Fig. 9), with an increase in transcript abundance at night (Supplementary Fig. 10). Several biological processes (e.g., carboxylic acid biosynthesis, terpene biosynthesis, and lipid metabolism) were over-represented (hypergeometric enrichment test,  $P < 0.05$ ) (Supplementary Data 1), and several key genes encoding proteins involved in nocturnal CAM carboxylation and vacuolar uptake of malate such as Kaladp0018s0289 ( $\beta$ -CA), Kaladp0048s0578 (*PEPC2*), Kaladp0037s0517 (*PPCK*), Kaladp0022s0111 (*MDH*), and Kaladp0062s0038 (*ALMT6*) were present in this module (Fig. 3a, Supplementary Note 1 and Supplementary Table 5). These results suggest that genes in the co-expression module MEblack play important roles in the nighttime processes that define CAM. One alternate module (MEblue containing 1911 genes) was significantly correlated with the leaf samples collected during the day (Supplementary Fig. 9), with an increase in transcript abundance during the light period

(Supplementary Fig. 10). Several biological processes (e.g., starch biosynthesis, coenzyme biosynthetic process) were over-represented (hypergeometric enrichment test,  $P < 0.05$ ) in this module (Supplementary Data 1). One gene in the CAM decarboxylation process, Kaladp0010s0106 (*PPDK-RP*), belongs to this module (Supplementary Table 6).

We also performed cluster analysis on the CAM leaf time-course expression data for the transcripts that showed significantly (ANOVA of glm models where  $H_0$  = a flat line,  $P < 0.05$ ) time-structured diel expression patterns as determined by a polynomial regression. Clustering of transcripts with time-structured expression identified 11 clusters (Supplementary Fig. 11 and Supplementary Table 7). Networks constructed for each cluster implicated highly connected hub genes and their direct or indirect interactions with CAM-related genes (Supplementary Data 2). For example, cluster 7, which contains *PEPC1* (Kaladp0095s0055) and *PPCK2* (Kaladp0604s0001), has a zinc-finger protein *CONSTANS-like* gene as a central hub



**Fig. 3** An overview of CAM pathway in *Kalanchoë fedtschenkoi*. **a** The CAM pathway map in *K. fedtschenkoi*. Orange colors indicate the key enzymes involved in the CAM pathway. The numbers in parenthesis are the four-fold transversion substitution rate (4dtv) values. **b** Diel expression profiles of duplicated genes in CAM-related gene families. ALMT tonoplast aluminum-activated malate transporter,  $\beta$ -CA  $\beta$  type carbonic anhydrase, ME malic enzyme, MDH malate dehydrogenase, PEP phosphoenolpyruvate, PEPC PEP carboxylase, PPCK PEPC kinase, PPK pyruvate phosphate dikinase, TDT tonoplast dicarboxylate transporter. White and black bars indicate daytime (12-h) and nighttime (12-h), respectively

(Supplementary Data 2). *CONSTANS-like* genes are part of the circadian clock regulatory network<sup>32</sup>. Similarly, multiple *REVEILLE* transcripts, which encode transcription factors for genes with evening elements in their promoters<sup>33</sup>, are hubs in cluster 4 that contains *NADP-ME* genes (Kaladp0092s0166) (Supplementary Data 2).

**Overview of genes that have undergone convergent evolution.**

To determine the possibility that the diel reprogramming of metabolism that distinguishes CAM from C<sub>3</sub> photosynthesis was achieved, at least in part, by convergent shifts in diel patterns of gene expression, we performed comparative analysis of diel transcript abundance patterns in CAM and C<sub>3</sub> photosynthesis species. Specifically, we compared the diel expression patterns of 9733 ortholog groups of genes from *K. fedtschenkoi* (eudicot, CAM photosynthesis), *A. comosus* (monocot, CAM

photosynthesis), and *Arabidopsis thaliana* (eudicot, C<sub>3</sub> photosynthesis), with transcript abundances >0.01 FPKM in mature leaf samples collected at six or more diel time points. Sampling time points included dawn (22, 24, and 2 h from the start of the light period), midday (4, 6, and 8 h from the start of the light period), dusk (10, 12, and 14 h from the start of the light period), and midnight (16, 18, and 20 h from the start of the light period) (Fig. 4a). A gene from *K. fedtschenkoi* was defined as having undergone convergent evolution of gene expression if it met all of the following criteria: (1) its diel transcript expression pattern was highly correlated (Spearman’s rank correlation coefficient,  $r > 0.8$ ) with those of at least one of the orthologs in *A. comosus*, but not highly correlated ( $r < 0.5$ ) with those of any of the orthologs in *A. thaliana*; (2) it displayed a significant difference (false discovery rate <0.01) in transcript abundance either between midday and midnight (e.g., Fig. 4b), or between dawn and dusk (e.g., Fig. 4c); and (3) the time shift between *K. fedtschenkoi* and *A. comosus* transcript time-courses was less than or equal to 3 h, whereas the time shifts between CAM species (*K. fedtschenkoi* and *A. comosus*) transcripts and their *A. thaliana* ortholog transcript were equal to or greater than 6 h. Based on these criteria, 54 *K. fedtschenkoi* genes were identified as candidates for involvement in the convergent shift in diel gene expression patterns specific to the two CAM species relative to *A. thaliana* (Supplementary Note 2, Supplementary Data 3 and Supplementary Table 8).

To identify genes that had likely undergone convergent evolution in protein sequence in the CAM species, we reconstructed gene tribes based on protein sequences from the species listed in Supplementary Fig. 4. We then created phylogenetic trees for the genes from all tribes that include at least one gene from each of the 13 studied species (Supplementary Table 9). A *K. fedtschenkoi* gene was defined as having undergone convergent evolution in protein sequence if it met all of the following criteria: (1) the *K. fedtschenkoi* gene is clustered with gene(s) from at least one of the two monocot CAM species (*A. comosus* and *P. equestris*) in a phylogenetic clade containing no genes from C<sub>3</sub> or C<sub>4</sub> photosynthesis species; (2) convergent amino-acid changes were detected between the *K. fedtschenkoi* gene with gene(s) from at least one of the two monocot CAM species; and (3) the *K. fedtschenkoi* gene shared at least one amino-acid mutation with its ortholog in at least one of the two monocot CAM species, as compared with C<sub>3</sub> and C<sub>4</sub> photosynthesis species. A total of four *K. fedtschenkoi* genes showing convergent changes in protein sequences were identified (Supplementary Figs. 12–15 and Supplementary Table 10).

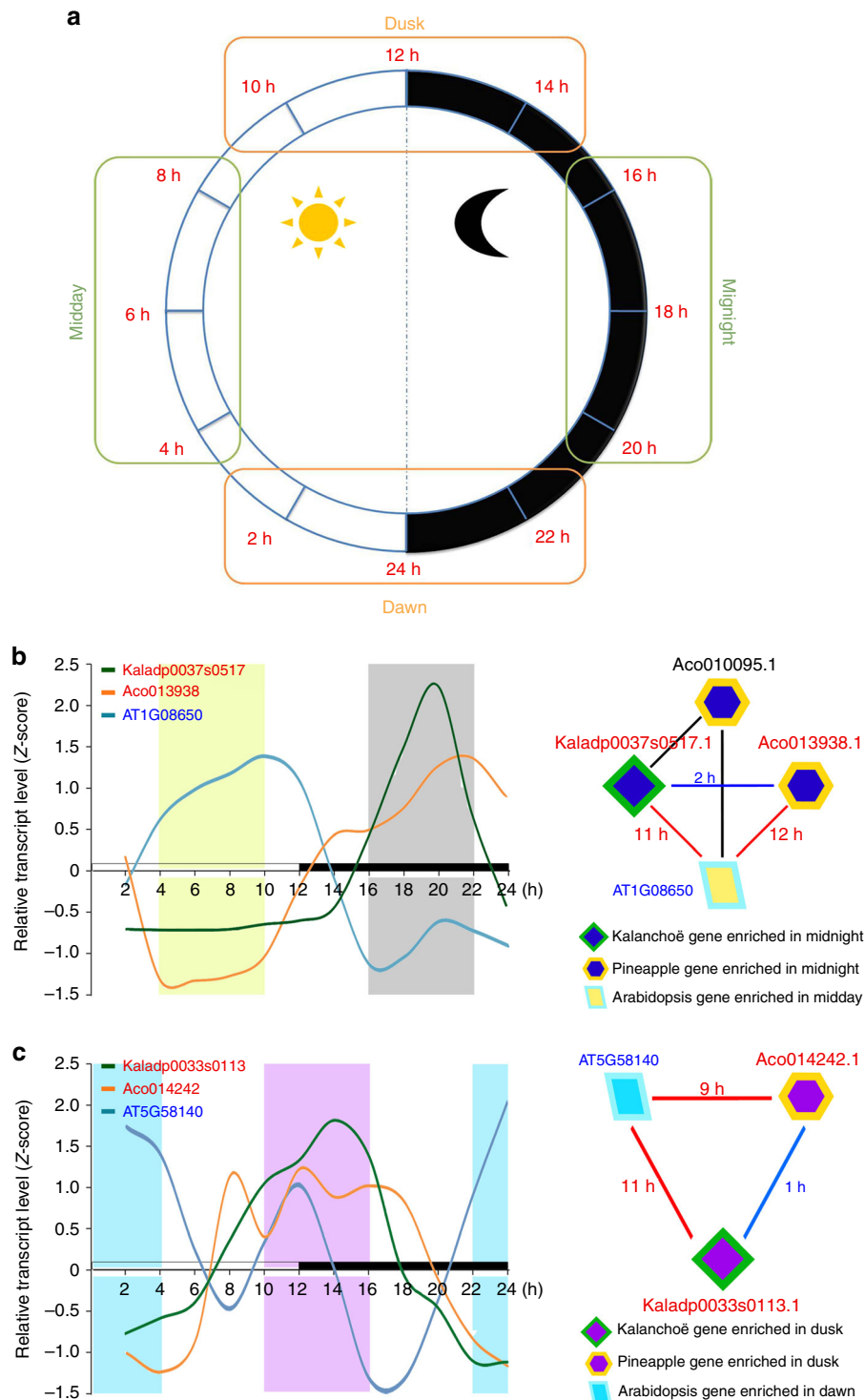
We also performed genome-wide positive selection analysis in each of the three CAM species (i.e., *A. comosus*, *P. equestris*, and *K. fedtschenkoi*) in comparison with 21 non-CAM species (Supplementary Method 1) and identified two genes that were under positive selection in the dicot CAM species *K. fedtschenkoi* and one of the monocot CAM species (Supplementary Figs. 16–17).

**Convergent evolution of genes involved in CO<sub>2</sub> fixation.**

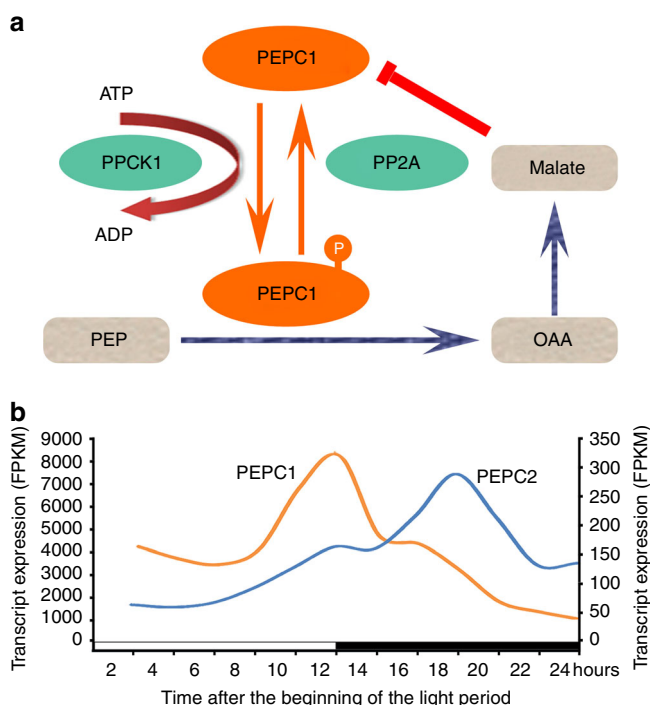
PEPC is a key enzyme for nocturnal CO<sub>2</sub> fixation and PPCK is a pivotal protein kinase that regulates PEPC in response to the circadian clock in CAM plants<sup>4,6,34</sup>. PPCK phosphorylates PEPC in the dark (Fig. 5a) and thereby reduces malate inhibition of PEPC activity, promoting nocturnal CO<sub>2</sub> uptake<sup>35,36</sup>. Multiple PPCK genes were identified in the *K. fedtschenkoi* genome, among which two genes (Kaladp0037s0517 and Kaladp0604s0001) showed higher transcript abundance than the others in CAM leaves (Supplementary Table 5). The diel expression patterns of the most abundant PPCK transcripts in *K. fedtschenkoi*

(Kaladp0037s0517.1) and *A. comosus* (Aco013938.1) were highly correlated, with only a 1.5-hour time shift between them, whereas both showed an ~11-hour time shift relative to their best matched

ortholog in *Arabidopsis* (AT1G08650) (Fig. 4b and Supplementary Table 8). Peak PPCK transcript abundance was shifted from daytime in *C<sub>3</sub>* photosynthesis species (*Arabidopsis*) to nighttime



**Fig. 4** Examples of convergent change in diel transcript expression pattern in CAM species. **a** The four time-windows for comparative analysis of temporal changes in transcript expression, which were represented by 12 time points: 2, 4, ..., 24 h after the beginning of the light period. **b** and **c** Comparison of diel transcript expression pattern of phosphoenolpyruvate carboxylase kinase 1 (PPCK1) and phototropin 2 (PHOT2), respectively, between CAM species (*Kalanchoë fedtschenkoi* and pineapple) and *C<sub>3</sub>* species (*Arabidopsis*). Left panels show the diel transcript expression profiles. Right panels show enrichment triangle networks, in which a *K. fedtschenkoi* gene and a pineapple ortholog had significantly enriched expression in the same time-window, whereas an *Arabidopsis* ortholog had significantly enriched expression in the opposite time-window. The numbers are the time shifts in diel transcript expression pattern between genes connected by each edge. White and black bars indicate daytime (12-h) and nighttime (12-h), respectively. X-axis represents the time after the beginning of the light period



**Fig. 5** Two phosphoenolpyruvate carboxylase (PEPC) genes with relative high transcript abundance in *Kalanchoë fedtschenkoi*. **a** Regulation of PEPC1 activity. **b** Diel expression profiles of *PEPC1* (Kaladp0095s0055.1) and *PEPC2* (Kaladp0048s0578.1) transcripts in *K. fedtschenkoi*, shown in the left and right Y-axis, respectively. OAA Oxaloacetate, PEP phosphoenolpyruvate, PEPC PEP carboxylase, PPCK PEPC kinase, PP2A protein phosphatase 2 A. White and black bars indicate daytime (12-h) and nighttime (12-h), respectively

in the two CAM species (Fig. 4b), which suggests convergence and is consistent with PPCK activation of PEPC-mediated nocturnal CO<sub>2</sub> fixation. Among the PEPC genes identified in *K. fedtschenkoi*, Kaladp0095s0055 and Kaladp0048s0578 showed higher transcript abundance than the others (Supplementary Table 5). Kaladp0095s0055 (named *PEPC1* herein) was an abundant transcript throughout both the light and the dark period, with its peak transcript level phased to dusk. The second most abundant PEPC transcript (Kaladp0048s0578, named *PEPC2* herein) showed a much higher transcript level during the dark period than during the light period (Fig. 5b). We found that a duplicated pair of *K. fedtschenkoi* *PEPC2* genes (Kaladp0048s0578 and Kaladp0011s0355) clustered together with a PEPC gene (PEQU\_07008) from *P. equestris* (Supplementary Fig. 12). PEQU\_07008 was recently reported as the CAM-type PEPC in *P. equestris*, and, like Kaladp0048s0578, this orchid *PEPC* gene also showed higher transcript abundance during the dark period than during the light period<sup>37</sup>.

Convergent changes in *PEPC2* protein sequence were found between *K. fedtschenkoi* and *P. equestris* (Fig. 6a, b). Specifically, multiple protein sequence alignment revealed that an aspartic acid residue (D509) in Kaladp0048s0578 is conserved in PEQU\_07008 and Kaladp0011s0355 (a duplicated copy of Kaladp0048s0578), but there was an arginine (R), lysine (K), or histidine (H) in the corresponding sites of the PEPC protein sequences of other tested species (Fig. 6c and Supplementary Fig. 12). The structural model of the Kaladp0048s0578 protein indicates that this single amino-acid substitution (from a basic amino-acid R/K/H to an acidic amino-acid D) is located in an  $\alpha$ -helix adjacent to the active site in a  $\beta$ -barrel (Fig. 7a). We hypothesize that an activator binds to the active site of one

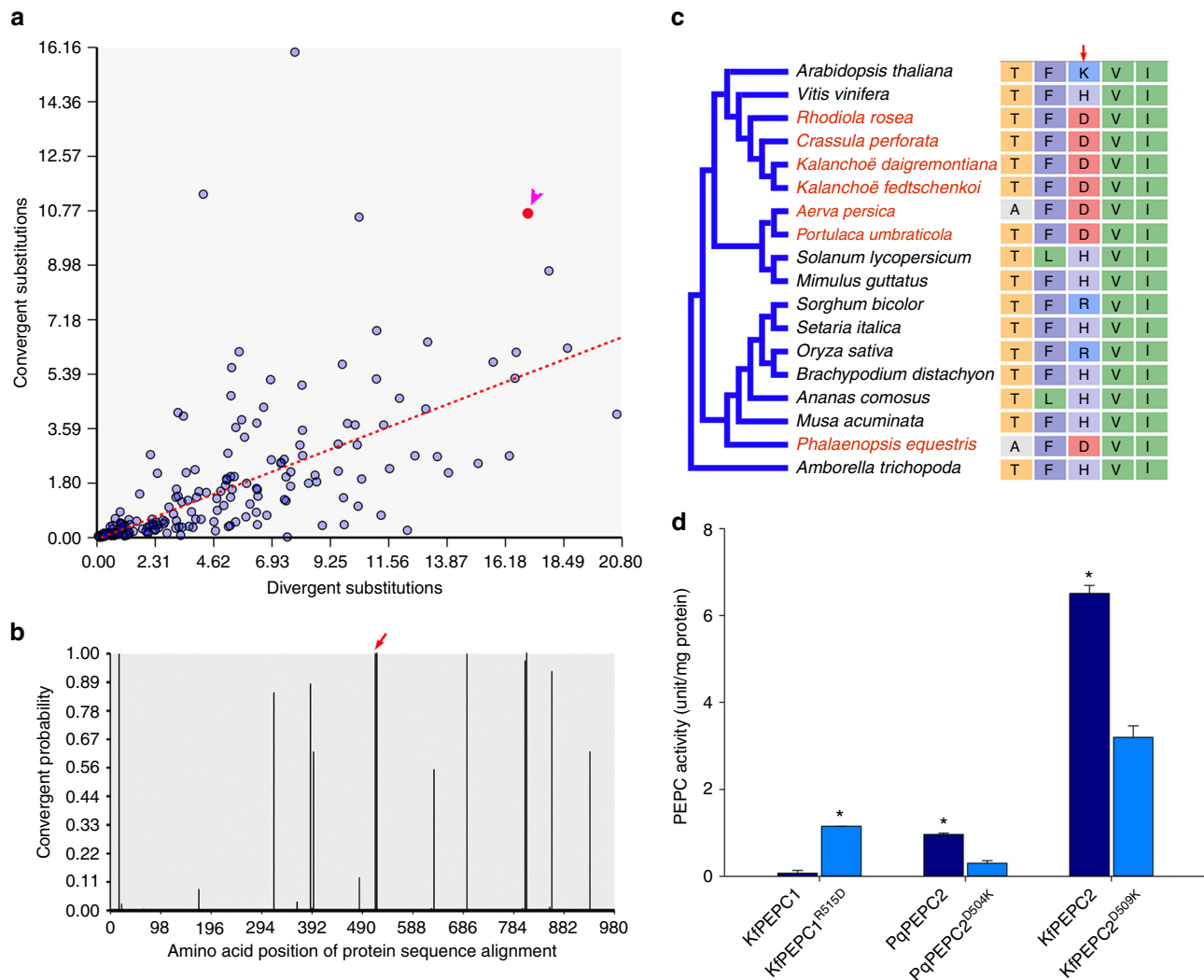
subunit of the tetrameric complex of *PEPC2*, leading to allosteric conformational changes that subsequently activate another subunit of the tetramer (Fig. 7b). This model was supported by a recent crystallography structure of the *Flavaria trinervia* (a C<sub>4</sub> photosynthesis plant) PEPC with an activator glucose-6-phosphate (G6P) bound at the  $\beta$ -barrel active center<sup>38</sup>. Based on this model, because D509 of *PEPC2* (Kaladp0048s0578) is also negatively charged as G6P, the observed substitution may play a similar role as the activator by triggering allosteric conformational changes that lead to activation of the other subunits of PEPC tetramer. Nimmo<sup>39</sup> reported that PEPC is subject to posttranslational regulation in the dark via phosphorylation by PPCK. *In vitro* analysis of the activities of different heterologously expressed PEPC isoforms showed that without phosphorylation by PPCK, *PEPC1* from *K. fedtschenkoi* had a much lower activity than *PEPC2* from either *K. fedtschenkoi* or *P. equestris* (Fig. 6d). Further, the R515D mutation significantly (Student's *t*-test,  $P < 0.01$ ) increased the activity of *K. fedtschenkoi* *PEPC1*, whereas the D509K and D504K mutations significantly (Student's *t*-test,  $P < 0.01$ ) reduced the activities of *K. fedtschenkoi* *PEPC2* and *P. equestris* *PEPC2*, respectively (Fig. 6d). These results indicate that a single amino-acid mutation could significantly modify PEPC activity.

Our evolutionary analyses did not detect convergent evolution in either protein sequence or diel transcription patterns for the various decarboxylation genes that are expressed in *Kalanchoë* and *A. comosus*. In *Kalanchoë*, NAD(P)-ME genes were highly expressed, whereas the expression of the PEPCK gene was very low (Supplementary Fig. 18), consistent with the known high extractable activities of NAD-ME and NADP-ME in CAM leaves of *Kalanchoë*<sup>40,41</sup>. By contrast, in *A. comosus* the transcript abundance of PEPCK was much higher than that of malic enzyme (ME) (Supplementary Fig. 18), supporting the model that malate decarboxylation in *Kalanchoë* is mediated by ME, which was recently substantiated using a transgenic RNAi approach<sup>20,40</sup>, whereas in pineapple a combination of MDH, working in the OAA-forming direction, coupled with PEPCK, converting OAA to PEP and CO<sub>2</sub>, are the candidate decarboxylation enzymes<sup>18</sup>, consistent with previous enzyme activity studies<sup>8</sup>.

### Convergent evolution of genes involved in stomatal movement.

A unique feature of CAM physiology is the inverted light/dark pattern of stomatal movement relative to C<sub>3</sub> photosynthesis, with stomata opening during the night in CAM and during the day in C<sub>3</sub> photosynthesis plants<sup>6</sup>. Blue light is a key environmental signal that controls stomatal opening and phototropin 2 (PHOT2; AT5G58140), a blue light photoreceptor, mediates blue light regulation of stomatal opening in *Arabidopsis*<sup>42</sup>. Twenty genes that could potentially be involved in stomatal movement in *K. fedtschenkoi* were predicted based on homology to *Arabidopsis* genes involved in the regulation of stomatal movement (Supplementary Table 11). One of these genes, Kaladp0033s0113, which encodes PHOT2, showed only a 1-h time shift in transcript abundance pattern relative to its *A. comosus* ortholog (Aco014242) (Supplementary Table 8), possibly indicating a convergent change in the diel pattern of its transcript abundance pattern in the two CAM species. In support of a convergent evolution hypothesis, the transcript abundance patterns of the two *PHOT2* genes in the CAM species showed 11- (*Kalanchoë*) and 9- (pineapple) hour phase shifts, respectively, relative to that of the *PHOT2* gene (AT5G58140) in the C<sub>3</sub> photosynthesis species *Arabidopsis* (Fig. 4c). The timing of peak transcript abundance shifted from dawn in *Arabidopsis* to dusk in the two CAM species (Fig. 4c). This convergent change in diel transcript abundance pattern suggests that PHOT2 might contribute to the





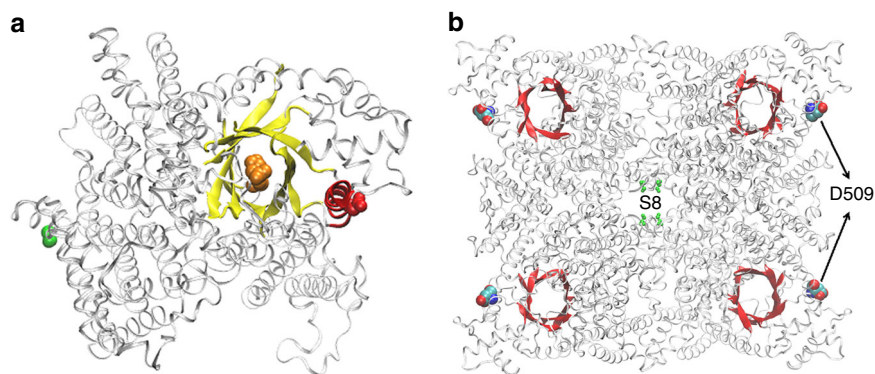
**Fig. 6** A convergent change in phosphoenolpyruvate carboxylase (PEPC) protein sequences in CAM species. **a** convergent- vs. divergent-substitutions in PEPC2 protein sequences between species listed in Supplementary Table 9. The arrow head indicates the comparison of *K. fedtschenkoi* vs. *P. equestris*. **b** Probability of convergent changes in PEPC2 protein sequence between *K. fedtschenkoi* and orchid. Red arrow indicates the protein sequence alignment site of convergent change (highlighted in red font at the alignment in panel c). **c** A convergent amino-acid change (from R/K/H to D) in PEPC2 shared by diverse species (highlighted in red font) at the alignment position indicated by the red arrow. **d** *In vitro* activity of PEPC isoforms in the absence of phosphorylation by PPCK. KfPEPC1: Kaladp0095s0055; KfPEPC1<sup>R515D</sup>: KfPEPC1 with mutation at residue 515 from arginine (R) to aspartic acid (D); KfPEPC2: Kaladp0048s0578.1; KfPEPC2<sup>D509K</sup>: KfPEPC2 with mutation at residue 509 from D to lysine (K); PqPEPC2: *P. equestris* PEPC gene PEQU07008; PqPEPC2<sup>D504K</sup>: PqPEPC2 with mutation at residue 504 from D to K. “\*” indicates significant difference between wild-type and mutant of PEPC1 or PEPC2 (Student’s *t*-test; *P* < 0.01). The error bars indicate standard deviation (SD) calculated from three replicates

inverted day/night pattern of stomatal closure and opening in CAM species such that PHOT2 might function as a switch mediating the blue-light signal to open the stomata at dusk and the stomata could then remain open during the dark period.

**Convergent evolution of genes involved in heat tolerance.** The stomata of mature CAM leaves of *K. fedtschenkoi* close for the majority of the light period<sup>40</sup>, which may exacerbate the internal heat load on the leaves<sup>43</sup>. Photosynthesis is sensitive to heat stress and can be inhibited long before other symptoms of heat stress are detected<sup>44</sup>. Numerous studies have shown that the inhibition of photosynthesis by moderate heat stress is a consequence of RuBisCO deactivation, caused, in part, by the thermal instability of RuBisCO activase<sup>45</sup>. Heat-shock proteins can play a critical role in the stabilization of proteins under heat stress conditions<sup>46</sup>. Wang et al.<sup>47</sup> reported that HSP40 (SICDJ2) contributed to the maintenance of CO<sub>2</sub> assimilation capacity mainly by protecting

RuBisCO activity under heat stress and that HSP70 (cpHsp70) acted as a binding partner for SICDJ2 in tomato. HSP70 can also function as nano-compartments in which single RbcL/RbcS subunits can fold in isolation, unimpaired by aggregation<sup>48</sup>, as illustrated in Fig. 8a. Among the HSP70 genes predicted in *K. fedtschenkoi*, Kaladp0060s0296 displayed peak transcript abundance in the morning, with only a 1-h shift in diel transcript abundance pattern relative to its *A. comosus* ortholog Aco031458, whereas these two HSP70 genes in the CAM species showed ~10-h shifts in diel transcript abundance pattern relative to their best-matched *A. thaliana* ortholog, AT5G02490 (Fig. 8b and Supplementary Table 8), suggesting that HSP70 has undergone convergent changes in diel transcript expression patterns during the evolution of CAM.

**Convergent evolution of genes in the circadian clock.** Key physiological and biochemical features of CAM including net CO<sub>2</sub>



**Fig. 7** Protein structure model of phosphoenolpyruvate carboxylase 2 (PEPC2) in *Kalanchoë fedtschenkoi*. **a** PEPC2 (Kaladp0048s0578.1) structural model with a glucose-6-phosphate (G6P) substrate (orange spheres) bound at the  $\beta$ -barrel active site (yellow). D509 (red spheres) is located at an  $\alpha$ -helix (red) in adjacent to the  $\beta$ -barrel and far from the hallmark serine residue (S8, green spheres) that is the phosphorylation target of PPCK1. **b** PEPC tetramer structure. The phosphorylation site (S8, green) is located at the interphase of the tetramer and D509 (spheres) is located at the peripheral of the tetramer. The  $\beta$ -barrel active site is shown in red, and no G6P activator may be required for activation of the PEPC activity following the competitive activating model of PEPC

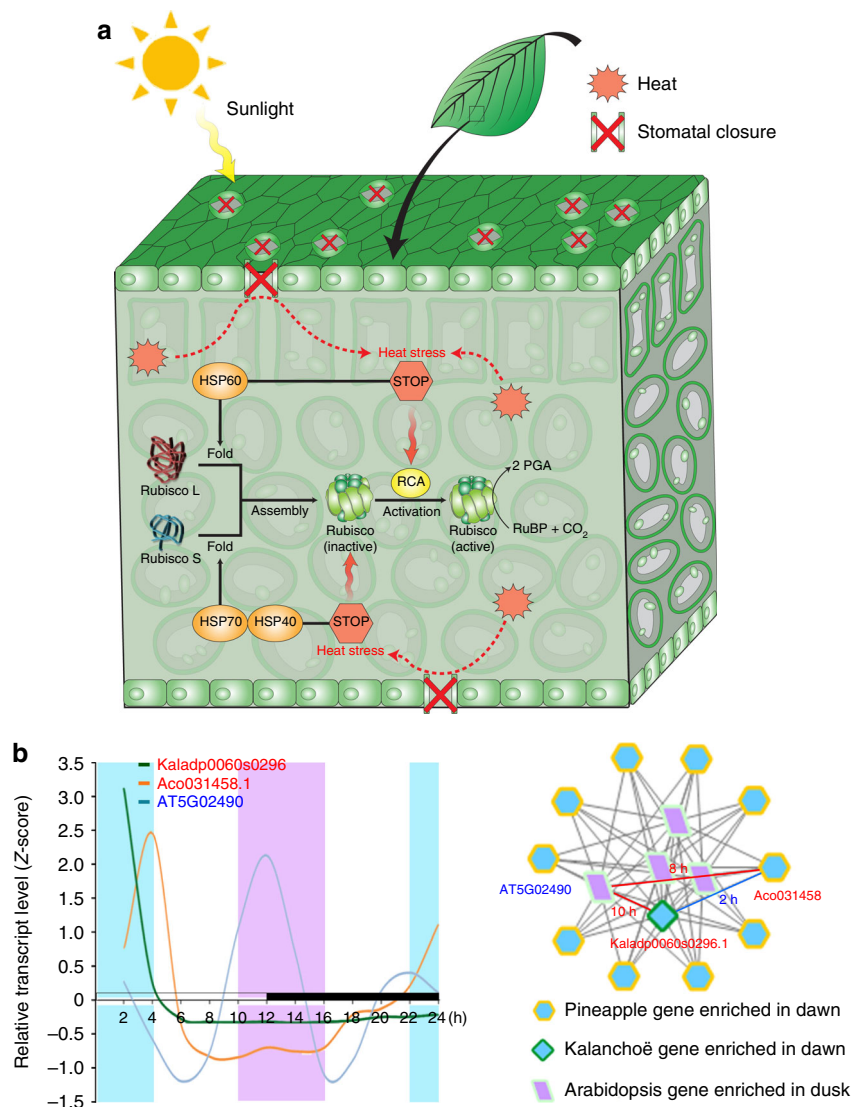
exchange and PEPC phosphorylation are well established as outputs of the circadian clock, displaying robust oscillation under free-running constant conditions<sup>20,40</sup>. Thus, the circadian clock could be a key regulator of the diel reprogramming of metabolism and stomatal function that defines CAM. The molecular basis of circadian rhythms has been studied extensively in non-CAM species<sup>33</sup>. Based on homology to *Arabidopsis* genes that have been shown to play important roles as molecular components of the circadian clock, 35 *K. fedtschenkoi* genes were predicted to be involved in circadian rhythms (Supplementary Table 12). None of these *K. fedtschenkoi* genes are among the list of genes showing convergent changes in diel expression pattern (Supplementary Data 3), suggesting that CAM evolution did not involve major changes in the diel expression pattern of these known circadian rhythm genes shared between *Arabidopsis* and *K. fedtschenkoi*. However, we cannot rule out the possibility of convergent evolution in unknown circadian rhythm genes between these two species. Also, it is possible that genes that are not involved in circadian rhythms in *Arabidopsis* could have taken on this function in *K. fedtschenkoi*. On the other hand, Kaladp0060s0460, which encodes ELONGATED HYPOCOTYL5 (HY5), showed a convergent change in protein sequences between *K. fedtschenkoi* and *P. equestris* (Supplementary Table 10). HY5 is a bZIP family transcription factor in the blue light signaling pathway that acts as an input to entrain the circadian clock<sup>33</sup> (Fig. 9a). A single amino-acid mutation (E-to-R) occurred in the C-terminal bZIP domains of the proteins encoded by Kaladp0060s0460 and its *P. equestris* ortholog PEQU\_13446 as compared with HY5 from  $C_3$  or  $C_4$  photosynthesis species (Fig. 9b and Supplementary Fig. 14). The bZIP domain determines the DNA-binding ability of HY5 as a transcription factor<sup>49</sup>, mediating the interaction between HY5 and G-BOX BINDING FACTOR 1<sup>50</sup>. HY5 has been shown to move from shoot to root to coordinate aboveground plant carbon uptake in the leaf and belowground nitrogen acquisition in the root<sup>51</sup>. Therefore, the potential roles of HY5, Kaladp0060s0460, in circadian rhythmicity and shoot-to-root communication in *K. fedtschenkoi* needs to be investigated using experimental approaches such as loss-of-function mutagenesis<sup>52</sup>.

### Convergent evolution of genes in carbohydrate metabolism.

Nocturnal production of phosphoenolpyruvate (PEP) as a substrate for dark  $CO_2$  uptake represents a substantial sink for carbohydrates in CAM plants, which has to be balanced with the provision of carbohydrates for growth and maintenance<sup>53</sup>. Carbohydrate active enzymes (CAZymes) play critical roles in

regulating carbohydrate synthesis, metabolism, and transport in living organisms. There are six CAZyme classes: glycoside hydrolases (GHs), glycosyltransferases (GTs), polysaccharide lyases, carbohydrate esterases, auxiliary activities, and carbohydrate-binding modules. Each of these classes contains from a dozen to over one hundred different protein families based on sequence similarity<sup>54</sup>. The six classes of CAZymes have different functions. For example, GH enzymes catalyze the hydrolysis of glycosidic bonds, while GT enzymes catalyze the formation of glycosidic bonds. Using CAZyme domain-specific hidden Markov models, defined in the dbCAN database<sup>55</sup>, we identified 100 CAZyme families, including 1093 genes in the *K. fedtschenkoi* genome, comparable to the total number (1149) of CAZyme genes in *A. thaliana* (Supplementary Data 4 and 5). Among these CAZyme genes, four ortholog groups (ORTHOMCL68, ORTHOMCL93, ORTHOMCL207, and ORTHOMCL9830) of genes (e.g., Kaladp0550s0020, Kaladp0011s0363, Kaladp0037s0421, Kaladp0055s0317, respectively) belonging to the CAZyme families GH100, GT20, GT2, and GT5, respectively, displayed convergent changes in their patterns of diel transcript abundance in two CAM species (*K. fedtschenkoi* and *A. comosus*) compared with the  $C_3$  photosynthesis species (*A. thaliana*) (Supplementary Data 3). Specifically, the *K. fedtschenkoi* CAZyme genes with convergent changes in diel transcript abundance pattern (e.g., Kaladp0550s0020 [GH100], Kaladp0011s0363 [GH20], Kaladp0037s0421 [GT2], and Kaladp0055s0317 [GT5]) showed higher transcript abundance in the dark and early light period (Supplementary Fig. 19). In particular, two genes (Kaladp0011s0363 and Kaladp0055s0317) were predicted to be involved in starch and sucrose metabolism (Supplementary Fig. 20). Kaladp0011s0363 encodes a probable trehalose phosphate synthase. Trehalose 6-phosphate is an important sugar signaling metabolite and is thought to link starch degradation to demand for sucrose and growth<sup>56</sup>. Kaladp0550s0020 encodes an alkaline-neutral invertase that catalyzes the hydrolysis of sucrose to glucose and fructose. This invertase has also been implicated in metabolic signaling processes as an important regulator of plant growth and development<sup>57</sup>. Taken together, these data suggest that the evolution of CAM from  $C_3$  photosynthesis requires re-scheduling of the transcription of metabolic and signaling genes that regulate the partitioning of carbohydrates between reserves that provide substrates for CAM and carbohydrates required for growth.

In addition to the above convergent changes in expression pattern of four CAZyme genes, we also identified convergent



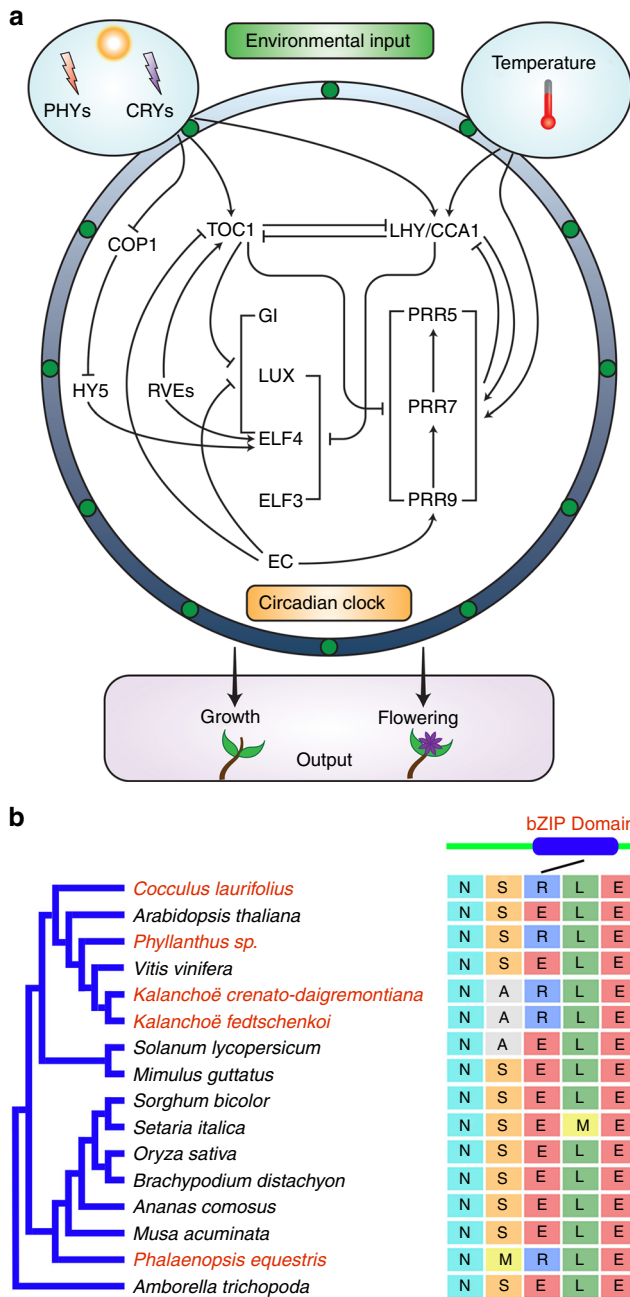
**Fig. 8** Convergent changes in diel transcript expression of heat-shock proteins (HSPs) in CAM species in comparison with  $C_3$  species. **a** Schematic representation of the possible roles of HSP40, HSP60, and HSP70 in leaf heat tolerance. **b** Comparison of diel transcript expression pattern of HSP70 between CAM species (*Kalanchoë fedtschenkoi* and pineapple) and  $C_3$  species (*Arabidopsis*). Left panel shows the diel transcript expression patterns. Right panel shows enrichment triangle network, in which a *K. fedtschenkoi* gene and a pineapple ortholog had significantly enriched expression in the same time-window, whereas an *Arabidopsis* ortholog had significantly enriched expression in the opposite time-window. The numbers are the time shifts in diel transcript expression pattern between genes connected by each edge. White and black bars indicate daytime (12-h) and nighttime (12-h), respectively. X-axis represents the time after the beginning of the light period. RuBisCO: Ribulose-1,5-bisphosphate carboxylase/oxygenase; RCA: rubisco activase; RuBP: ribulose-1,5-bisphosphate; PGA: 3-phosphoglycerate

changes in protein sequences of another two CAZyme genes (Kaladp0016s0058 [GT29] and Kaladp0067s0114 [GH35]) that were under positive selection (CodeML implemented in PosiGene<sup>58</sup>,  $P < 0.05$ ) in the dicot CAM species *K. fedtschenkoi* and one of the two monocot CAM species (*A. comosus* and *P. equestris*) (Supplementary Figs. 16–17). Kaladp0016s0058 encodes a putative sialyltransferase-like protein. Two single amino-acid mutations were found in Kaladp0016s0058 and its *A. comosus* ortholog Aco018360, as compared with the orthologous protein sequences of non-CAM species (Supplementary Fig. 16). These two mutations are close to each other (i.e., within a four-amino-acid distance), suggesting the possibility that the two mutations affect the same functional domain. Kaladp0067s0114 encodes a beta-galactosidase protein that hydrolyses the glycosidic bond between two or more carbohydrates. Two single amino-acid mutations were identified in

Kaladp0067s0114 and its *P. equestris* ortholog PEQU\_04899, as compared with the orthologous protein sequences of non-CAM species (Supplementary Fig. 17). These two mutations are close to each other (i.e., within an 11-amino-acid distance) in the middle of galactose-binding domain (Supplementary Fig. 17), which can bind to specific ligands and carbohydrate substrates for enzymatic catalytic reactions<sup>59</sup>. The relevance of these convergent changes in protein sequence to CAM evolution needs further investigation.

## Discussion

The CAM pathway has been found in 36 families of vascular plants<sup>4</sup>, among which Crassulaceae plays a unique role in CAM research because the pathway was first discovered in this succulent plant family and was thus named<sup>60</sup>. Within Crassulaceae, the



**Fig. 9** A convergent change in elongated hypocotyl 5 (HY5) protein sequences in CAM species. **a** An overview of the signaling pathway involved in circadian rhythm in plants. **b** Convergent change in HY5 protein sequences in diverse species (highlighted in red font). The black line indicates the protein sequence alignment position (located within the bZIP domain) where the mutation (E-to-R) occurred. CCA1 circadian clock associated 1, COP1 constitutive photomorphogenic 1, CRY cryptochrome, EC evening complex, ELF 3/4 early flowering 3/4, GI gigantea, LHY late elongated hypocotyl, LUX lux arrhythmo, PRR5/7/9 pinorosinol reductase 5/7/9, PHYs phytochromes, RVEs reveilles, TOC1 timing of cab expression 1

genus *Kalanchoë* has been the most widely used for CAM research. As a model species for research into the molecular biology and functional genomics of CAM, *K. fedtschenkoi* stands out due to its relatively small genome, low repetitive sequence content, and efficient stable transformation protocols<sup>20</sup>. The genome sequence presented in this study renders *K. fedtschenkoi*

as a new model for plant evolutionary and comparative genomics research, both for CAM photosynthesis and beyond. Although this study focused on genome-wide analysis of convergent evolution in CAM plants, the *K. fedtschenkoi* genome data can be used to facilitate CAM research related to: (1) generating loss-of-function mutants for functional characterization of CAM-related genes using genome-editing technology; (2) deciphering the regulation of CAM genes through identification of transcription factors and promoters of their target genes; (3) analyzing CAM gene expression by serving as a template for mapping of RNA sequencing reads and protein mass spectrometry data; and (4) identifying DNA polymorphisms related to genetic diversity of plants in the genus *Kalanchoë*.

Our genome-wide comparison of CAM species and non-CAM species revealed two types of convergent changes that could be informative with respect to the evolution of CAM: protein sequence convergence and convergent changes in the diel re-scheduling of transcript abundance. In the present study, a total of 60 genes exhibited convergent evolution in divergent eudicot and monocot CAM lineages. Specifically, we identified protein sequence convergence in six genes involved in nocturnal CO<sub>2</sub> fixation, circadian rhythm, carbohydrate metabolism, and so on (Supplementary Table 10 and Supplementary Figs. 16–17). Also, we identified convergent diel expression changes in 54 genes that are involved in stomatal movement, heat stress response, carbohydrate metabolism, and so on (Supplementary Data 3). These results provide strong support for our hypothesis that convergent evolution in protein sequence or gene temporal expression underpins the multiple and independent emergences of CAM from C<sub>3</sub> photosynthesis. New systems biology tools and genome-editing technologies<sup>52,61</sup> offer great potential for plant functional genomics research based on loss- or gain-of-function mutants to characterize the role of the genes predicted here to have undergone convergent evolution.

Convergent gene function can arise by (1) a mutation or mutations in the same gene or genes that result in homoplasy in organisms or (2) independent causal mutation or mutations in different genes in each lineage<sup>10,62</sup>. We identified four genes that showed convergent changes in protein sequences, none of which were shared by the three CAM species *A. comosus*, *K. fedtschenkoi*, and *P. equestris* (Supplementary Table 10 and Supplementary Figs. 12–15), suggesting that CAM convergences result mainly from the second scenario. Alternatively, *K. fedtschenkoi* shares the convergent mutation in the PEPC2 protein sequence with *P. equestris* (Fig. 6), whereas it shares the convergent change in the pattern of diel transcript abundance of *PPCK1* with *A. comosus* (Fig. 4b). These results suggest that two alternative modes of convergent evolution could have occurred in pathways for nocturnal CO<sub>2</sub> fixation. First, PPCK shifted from light period to dark period to promote the activation of PEPC1 (the most abundant isoform), as exemplified by *K. fedtschenkoi* and *A. comosus*. Second, a single amino-acid mutation from R/K/H to D to maintain the active state of PEPC2, without the need for phosphorylation, then occurred, as in *K. fedtschenkoi* and *P. equestris*.

According to the constrained selection theory of Morris<sup>11</sup>, we expected to see convergent changes in protein sequences in all the three CAM species. However, in this study, single-site mutations were found in only two of the three CAM species. Our additional positive selection analysis revealed that *Kalanchoë* did share convergent sequence mutation with the other two CAM species, but at alternate sites (Supplementary Figs. 16–17). This is consistent with a recent report showing that single amino-acid mutations were not shared by all the bird species that displayed convergent evolution of hemoglobin function as an adaptation to high-altitude environments<sup>14</sup>. Alternatively, our results, to some extent, support the contingent adaptation theory of Gould<sup>15</sup>. The relevance of these predicted convergent changes to CAM needs to

be investigated using experimental approaches, such as transferring the convergent CAM genes to  $C_3$  photosynthesis species to test the effect of these genes on  $C_3$ -to-CAM photosynthesis transition.

In this study, we did not identify any gene that exhibited both convergent changes in transcript abundance patterns (Supplementary Data 3) and convergent changes in protein sequence (Supplementary Table 10 and Supplementary Figs. 16–17), suggesting that convergent evolution of a gene in CAM species was achieved through either protein sequence convergence or rewiring of gene expression. Indeed, we have not seen any reports showing that both protein convergence and convergent gene expression change occurred in the same gene in any type of organisms to date. Thus, we can hypothesize that convergent evolution follows the “law of parsimony” that emphasizes the fewest possible assumptions for explaining a thing or event<sup>63</sup>. An implication of this hypothesis is that reuse of the key genes via altered diel expression patterns would be the shortest path for  $C_3$ -to-CAM photosynthesis evolution; and on the other hand, mutations in some key sites of protein sequences, while keeping the temporal gene expression pattern unchanged, would be the shortest path for evolving new protein function required by CAM. Although our data fit this hypothesis, additional screens for genes that have convergent changes in both protein sequence and expression pattern in the future are merited.

Increasing human population and changes in global temperature and precipitation are creating major challenges for the sustainable supply of food, fiber, and fuel in the years to come. As a proven mechanism for increasing WUE in plants, CAM offers great potential for meeting these challenges. Engineering of CAM-into- $C_3$  photosynthesis plants could be a viable strategy to improve WUE in non-CAM crops for food and biomass production<sup>4,6</sup>. The genes predicted here to have undergone convergent evolution during the emergence of CAM are crucial candidates for CAM-into- $C_3$  photosynthesis engineering. Our results suggest that CAM-into- $C_3$  photosynthesis engineering requires rewiring of the diel transcript abundance patterns for most of the candidate genes in the target  $C_3$  photosynthesis species, along with amino-acid mutations in the protein sequences of several other candidate genes. Specifically, CAM-into- $C_3$  photosynthesis engineering efforts should be focused on changing the temporal patterns of transcript expression of endogenous genes in the target  $C_3$  photosynthesis species corresponding to the *K. fedtschenkoi* genes listed in Supplementary Data 3. CRISPR/Cas9-based knock-in approach<sup>52</sup> can be used to replace the original endogenous promoters of the target genes with temporal promoters that confer temporal expression patterns similar to those of their orthologous genes in the CAM species. For example, dark-inducible promoters such as Din10<sup>64</sup> can be used to drive the expression of carboxylation gene modules during the nighttime and light-inducible promoters, such as GT1-GATA-NOS101<sup>65</sup>, can be used to drive the expression of decarboxylation gene modules during the daytime. To make the protein sequence changes needed for CAM-into- $C_3$  photosynthesis engineering, transferring the *K. fedtschenkoi* genes listed in Supplementary Table 10 to target  $C_3$  photosynthesis species via the *Agrobacterium*-mediated transformation could provide a relatively straightforward path to an efficient engineered CAM pathway. Alternatively, one could mutate the amino acids shown in Supplementary Figs. 12–15 using a knock-in strategy with emerging genome-editing technology<sup>52</sup>.

In summary, this study provides an important model genome for studying plant comparative, functional, and evolutionary genomics, as well as significant advances in our understanding of CAM evolution. Our findings hold tremendous potential to accelerate the genetic improvement of crops for enhanced

drought avoidance and sustainable production of food and bioenergy on marginal lands.

## Methods

**Plant material.** *Kalanchoë fedtschenkoi* ‘M2’ plants were purchased from Mass Spectrum Botanicals (Tampa, FL, USA) (Supplementary Method 2).

**Estimation of DNA content.** The DNA contents of young leaf tissue samples were analyzed using flow cytometry analysis service provided by Plant Cytometry Services (The Netherlands). The internal standard was *Vinca minor* (DNA = 1.51 pg/2 C = 1477 Mbp/2 C).

**Chromosome counting.** Images were collected using an Olympus FluoView FV1000 confocal microscope (Center Valley, PA, USA) with a 60× objective. Images were sharpened using Adobe Photoshop and chromosomes were counted using ImageJ software (Supplementary Method 3).

**Illumina sequencing of genome.** The genomic DNA libraries of *K. fedtschenkoi* were sequenced on a MiSeq instrument (Illumina, CA, USA) using MiSeq Reagent Kit v3 (600-cycle) (Illumina, CA, USA) (Supplementary Method 4).

**Transcriptome sequencing.** In order to capture mRNA abundance changes responsive to diel conditions, samples were collected in triplicate from mature *K. fedtschenkoi* leaves (i.e., the fifth and sixth mature leaf pairs counting from the top) every 2 h over a 24 h time course under 12 h light/12 h dark photoperiod. Additional tissues were sampled in triplicate including roots, flowers, shoot tips plus young leaves, and stems at one time point, 4 h after the beginning of the light period (Supplementary Method 5).

**Genome assembly and improvement.** The *K. fedtschenkoi* genome was initially assembled using platanus<sup>66</sup> from 70X Illumina paired-end reads (2 × 300 bp reads; unamplified 540 bp whole-genome shotgun fragment library), and three mate-libraries (3 kb, 14X; 6 kb, 12X; 11 kb, 11X). Further genome scaffolding was performed using MeDuSa<sup>67</sup> sequentially with the genome assemblies of *K. laxiflora* v1.1 (Phytozome), *Vitis vinifera* Genoscope.12X (Phytozome), and *Solanum tuberosum* v3.4 (Phytozome).

**Protein-coding gene annotation.** The genome annotation for *K. fedtschenkoi* was performed using homology-based predictors facilitated with transcript assemblies (Supplementary Method 6).

**Construction of orthologous groups.** The protein sequences of 26 plant species were selected for ortholog group construction (Supplementary Method 7).

**Construction of species phylogeny.** The phylogeny of plant species was constructed from the protein sequences of 210 single-copy genes identified through analysis of orthologous groups (see “Construction of orthologous groups” section). The details for species phylogeny construction are described in Supplementary Method 8.

**Construction of protein tribes and phylogenetic analysis.** The protein sequences used for ortholog analysis (see “Construction of orthologous groups”) were also clustered into tribes using TRIBE-MCL<sup>68</sup>, with a BLASTp E-value cutoff of 1e-5 and an inflation value of 5.0. Phylogenetic analysis of the protein tribes is described in Supplementary Method 9.

**Analysis of convergence in protein sequences in CAM species.** The phylogenetic trees of protein tribes (see aforementioned “Construction of protein tribes and phylogenetic analysis”) were examined to identify the “CAM-convergence” clade, which was defined to contain genes from *K. fedtschenkoi* (dicot) and at least one of the two monocot CAM species (*A. comosus* and *P. equestris*) without any genes from  $C_3$  or  $C_4$  species. The rationale for defining the “CAM-convergence” clade is that the dicot CAM species *K. fedtschenkoi* should be separated from the monocot CAM species if there is no convergence between *Kalanchoë* and the monocot CAM species (Supplementary Method 10).

**Gene Ontology analysis and pathway annotation.** Whole-genome gene ontology (GO) term annotation was performed using BLAST2GO<sup>69,70</sup> with a BLASTp E-value hit filter of  $1 \times 10^{-6}$ , an annotation cutoff value of 55, and GO weight of 5. The enrichment of GO biological process and pathway annotation are described in Supplementary Method 11.

**Analysis of carbohydrate active enzymes.** The protein sequences were searched against the dbCAN database<sup>55</sup> using HMMER3 (<http://hmmer.org/>). The HMMER search outputs were parsed to keep significant hits with E-value <1e-23 (calculated

by HMMER) and coverage >0.2 (calculated on the HMM, which is equal to (end position - start position)/total length of HMM), as suggested by a large scale benchmark analysis<sup>71</sup>.

**Estimation of transcript abundance in *Kalanchoë*.** The RNA-seq data in fastq format were mapped to the *K. fedtschenkoi* genome using TopHat2<sup>72</sup>. Transcript abundance in FPKM (Fragments Per Kilobase of transcript per Million mapped reads) was estimated using Cufflinks<sup>73</sup>. All mapped read counts of the transcripts were counted by using htseq-count, a subprogram of HTseq<sup>74</sup>.

**Co-expression network analysis in *Kalanchoë*.** The expression data of 16 samples in triplicates were used for co-expression network analysis, which included time-course data (12 time points: 2, 4, 6, ..., 24 h after the beginning of the light period) from mature leaf and one time point data (4 h after the beginning of the light period) from roots, flowers, stems, and shoot tips plus young leaves collected in triplicate from the *K. fedtschenkoi* plants grown under 12 h light/12 h dark photoperiod. The details for co-expression network analysis are described in Supplementary Method 12.

**Cluster analysis of gene expression in *Kalanchoë*.** Count values for each RNA-seq library were used to calculate polynomial regressions across time (Supplementary Method 13).

**Comparative analysis of gene expression.** The diurnal expression data with 4-h intervals for *Arabidopsis thaliana* were obtained from Mockler et al.<sup>75</sup> and adjusted to 2-h interval time series by interpolation using the SRS1 cubic spline function (<http://www.srs1software.com/>). The diurnal expression data with 2-h intervals for *K. fedtschenkoi* was generated in this study. The diurnal expression data with 2-h intervals for *Ananas comosus* was obtained from Ming et al.<sup>4</sup>. The gene expression data were normalized by Z-score transformation. The hierarchical clustering of gene expression was performed for genes in each ortholog group using the Bioinformatics Toolbox in Matlab (Mathworks, Inc.) based on Spearman correlation (Supplementary Method 14).

**Genome synteny analysis.** Pairwise genome alignments were performed between grape genome (Genoscope.12X; <https://phytozome.jgi.doe.gov>) and *K. fedtschenkoi* (Supplementary Method 15).

**Protein 3D structural simulation.** The protein structural models were built using the iterative threading assembly refinement (I-TASSER, V4.3) structural modeling toolkit<sup>76,77</sup>.

**Gas chromatography-mass spectrometry metabolite profiling.** For the major metabolites of *K. fedtschenkoi*, a total of 36 leaf samples (the 5th and 6th fully expanded leaf pairs counting from the top) were collected with three biological replicates sampled every 2-h for a 24-h diurnal cycle. Additionally, three biological replicate samples of stems, roots, shoot tips plus young leaves, and flowers were also collected (Supplementary Method 16).

**In vitro protein expression and analysis of enzyme activity.** The PEPC proteins were expressed in bacterial BL21 strains (Novagen BL21 (DE3) pLysS Singles), and purified via Glutathione Sepharose 4B beads (GE Healthcare Life Sciences, Pittsburgh, PA, USA). The protein quality was checked via western blot using anti-PEPC antibody (Agrisera, Sweden) and the PEPC activity was determined (Supplementary Method 17).

**Data availability.** The Department of Energy (DOE) will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>). The *K. fedtschenkoi* genome sequence and annotation are deposited in Phytozome (<https://phytozome.jgi.doe.gov>). The *K. fedtschenkoi* genome sequence is also deposited at NCBI GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>) under the accession code NQLW00000000. The genome sequencing reads are deposited in NCBI Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra>) with the BioSample accession SAMN07509503, which is the combination of the five individual BioSamples: SAMN07453935, SAMN07453936, SAMN07453937, SAMN07453938, and SAMN07453939. The RNA-Seq reads are deposited in NCBI SRA with the BioSample accession codes SAMN07453940 - SAMN07453987. The metabolite data is deposited at MetaboLights (<http://www.ebi.ac.uk/metabolights/>) under the accession code MTBLS519.

Received: 12 January 2017 Accepted: 21 September 2017

Published online: 01 December 2017

## References

- West-Eberhard, M., Smith, J. & Winter, K. Photosynthesis, reorganized. *Science* **332**, 311–312 (2011).
- Borland, A. M., Griffiths, H., Hartwell, J. & Smith, J. A. C. Exploiting the potential of plants with crassulacean acid metabolism for bioenergy production on marginal lands. *J. Exp. Bot.* **60**, 2879–2896 (2009).
- Cushman, J. C., Davis, S. C., Yang, X. & Borland, A. M. Development and use of bioenergy feedstocks for semi-arid and arid lands. *J. Exp. Bot.* **66**, 4177–4193 (2015).
- Yang, X. et al. A roadmap for research on crassulacean acid metabolism (CAM) to enhance sustainable food and bioenergy production in a hotter, drier world. *New Phytol.* **207**, 491–504 (2015).
- Owen, N. A. & Griffiths, H. A system dynamics model integrating physiology and biochemical regulation predicts extent of crassulacean acid metabolism (CAM) phases. *New Phytol.* **200**, 1116–1131 (2013).
- Borland, A. M. et al. Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends Plant Sci.* **19**, 327–338 (2014).
- Silvera, K. et al. Evolution along the crassulacean acid metabolism continuum. *Funct. Plant Biol.* **37**, 995–1010 (2010).
- Christopher, J. & Holtum, J. Patterns of carbohydrate partitioning in the leaves of crassulacean acid metabolism species during deacidification. *Plant Physiol.* **112**, 393–399 (1996).
- Holtum, J. A. M., Smith, J. A. C. & Neuhaus, H. E. Intracellular transport and pathways of carbon flow in plants with crassulacean acid metabolism. *Funct. Plant Biol.* **32**, 429–449 (2005).
- Washburn, J. D., Bird, K. A., Conant, G. C., Pires, J. C. & Herendeen, P. S. Convergent evolution and the origin of complex phenotypes in the age of systems biology. *Int. J. Plant Sci.* **177**, 305–318 (2016).
- Morris S. C. *Life's Solution: Inevitable Humans In A Lonely Universe*. Cambridge University (2003).
- Foot, A. D. et al. Convergent evolution of the genomes of marine mammals. *Nat. Genet.* **47**, 272–275 (2015).
- Hu, Y. B. et al. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proc. Natl Acad. Sci. USA* **114**, 1081–1086 (2017).
- Natarajan, C. et al. Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354**, 336–339 (2016).
- Gould S. J. *Wonderful Life: The Burgess Shale And The Nature Of Life*. Norton (1989).
- Pfenning, A. R. et al. Convergent transcriptional specializations in the brains of humans and song-learning birds. *Science* **346**, 1256846 (2014).
- Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol.* **207**, 437–453 (2015).
- Ming, R. et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **47**, 1435–1442 (2015).
- Cai, J. et al. The genome sequence of the orchid *Phalaenopsis equestris*. *Nat. Genet.* **47**, 65–72 (2015).
- Hartwell, J., Dever, L. V. & Boxall, S. F. Emerging model systems for functional genomics analysis of crassulacean acid metabolism. *Curr. Opin. Plant Biol.* **31**, 100–108 (2016).
- Soltis, D. E., Soltis, P. S., Endress, P. K., & Chase, M. W. *Phylogeny And Evolution Of Angiosperms*. Sinauer Associates Inc. (2005).
- Soltis, D. E. et al. Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. *Am. J. Bot.* **100**, 916–929 (2013).
- The Angiosperm Phylogeny Group. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. *Bot. J. Linn. Soc.* **181**, 1–20 (2016).
- Zeng, L. et al. Resolution of deep eudicot phylogeny and their temporal diversification using nuclear genes from transcriptomic and genomic datasets. *New Phytol.* **214**, 1338–1354 (2017).
- Moore, M. J., Soltis, P. S., Bell, C. D., Burleigh, J. G. & Soltis, D. E. Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots. *Proc. Natl Acad. Sci. USA* **107**, 4623–4628 (2010).
- Maddison, W. P. & Knowles, L. L. Inferring phylogeny despite incomplete lineage sorting. *Syst. Biol.* **55**, 21–30 (2006).
- Degnan, J. H. & Rosenberg, N. A. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
- Murat, F. et al. Karyotype and gene order evolution from reconstructed extinct ancestors highlight contrasts in genome plasticity of modern rosid crops. *Genome Biol. Evol.* **7**, 735–749 (2015).
- Jaillon, O. et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–467 (2007).
- Paterson, A. H. et al. Repeated polyploidization of *Gossypium* genomes and the evolution of spinnable cotton fibres. *Nature* **492**, 423–427 (2012).
- Amborella Genome Project. The *Amborella* genome and the evolution of flowering plants. *Science* **342**, 1241089 (2013).

32. Ledger, S., Strayer, C., Ashton, F., Kay, S. A. & Putterill, J. Analysis of the function of two circadian-regulated *CONSTANS-LIKE* genes. *Plant J.* **26**, 15–22 (2001).
33. Hsu, P. Y. & Harmer, S. L. Wheels within wheels: the plant circadian system. *Trends Plant Sci.* **19**, 240–249 (2014).
34. Hartwell J. The Circadian Clock in CAM Plants. in *Annual Plant Reviews: Endogenous Plant Rhythms* (ed Hall A. J. W., McWatters H.). Blackwell Publishing (2006).
35. Hartwell, J., Nimmo, G., Wilkins, M., Jenkins, G. & Nimmo, H. Phosphoenolpyruvate carboxylase kinase is a novel protein kinase regulated at the level of gene expression. *Plant J.* **20**, 333–342 (1999).
36. Taybi, T., Patil, S., Chollet, R. & Cushman, J. A minimal Ser/Thr protein kinase circadianly regulates phosphoenolpyruvate carboxylase activity in CAM-induced leaves of *Mesembryanthemum crystallinum*. *Plant Physiol.* **123**, 1471–1482 (2000).
37. Zhang, L. et al. Origin and mechanism of crassulacean acid metabolism in orchids as implied by comparative transcriptomics and genomics of the carbon fixation pathway. *Plant J.* **86**, 175–185 (2016).
38. Schlieper, D., Förster, K., Paulus, J. K. & Groth, G. Resolving the activation site of positive regulators in plant phosphoenolpyruvate carboxylase. *Mol. Plant J.* **7**, 437–440 (2014).
39. Nimmo, H. G. The regulation of phosphoenolpyruvate carboxylase in CAM plants. *Trends Plant Sci.* **5**, 75–80 (2000).
40. Dever, L. V., Boxall, S. F., Kneřová, J. & Hartwell, J. Transgenic perturbation of the decarboxylation phase of crassulacean acid metabolism alters physiology and metabolism but has only a small effect on growth. *Plant Physiol.* **167**, 44–59 (2015).
41. Dittrich, P. Nicotinamide adenine dinucleotide-specific “malic” enzyme in *Kalanchoë daigremontiana* and other plants exhibiting crassulacean acid metabolism. *Plant Physiol.* **57**, 310–314 (1976).
42. Kinoshita, T. et al. Phot1 and phot2 mediate blue light regulation of stomatal opening. *Nature* **414**, 656–660 (2001).
43. Krause, G. H., Winter, K., Krause, B. & Virgo, A. Protection by light against heat stress in leaves of tropical crassulacean acid metabolism plants containing high acid levels. *Funct. Plant Biol.* **43**, 1061–1069 (2016).
44. Berry, J. & Björkman, O. Photosynthetic response and adaptation to temperature in higher plants. *Annu. Rev. Plant Physiol.* **31**, 491–543 (1980).
45. Salvucci, M. E. & Crafts-Brandner, S. J. Mechanism for deactivation of Rubisco under moderate heat stress. *Physiol. Plant.* **122**, 513–519 (2004).
46. Wang, W., Vinocur, B., Shoseyov, O. & Altman, A. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. *Trends Plant Sci.* **9**, 244–252 (2004).
47. Wang, G. et al. A tomato chloroplast-targeted DnaJ protein protects Rubisco activity under heat stress. *J. Exp. Bot.* **66**, 3027–3240 (2015).
48. Liu, C. et al. Coupled chaperone action in folding and assembly of hexadecameric Rubisco. *Nature* **463**, 197–202 (2010).
49. Nijhawan, A., Jain, M., Tyagi, A. K. & Khurana, J. P. Genomic survey and gene expression analysis of the basic leucine zipper transcription factor family in rice. *Plant Physiol.* **146**, 333–350 (2008).
50. Ram, H. & Chattopadhyay, S. Molecular interaction of bZIP domains of GBF1, HY5 and HYH in Arabidopsis seedling development. *Plant Signal. Behav.* **8**, e22703 (2013).
51. Chen, X. et al. Shoot-to-root mobile transcription factor HY5 coordinates plant carbon and nitrogen acquisition. *Curr. Biol.* **26**, 640–646 (2016).
52. Liu, D., Hu, R., Palla, K. J., Tuskan, G. A. & Yang, X. Advances and perspectives on the use of CRISPR/Cas9 systems in plant genomics research. *Curr. Opin. Plant Biol.* **30**, 70–77 (2016).
53. Borland, A. M., Guo, H.-B., Yang, X. & Cushman, J. C. Orchestration of carbohydrate processing for crassulacean acid metabolism. *Curr. Opin. Plant Biol.* **31**, 118–124 (2016).
54. Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42**, D490–D495 (2014).
55. Yin, Y. et al. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–W451 (2012).
56. Martins, M. C. M. et al. Feedback inhibition of starch degradation in *Arabidopsis* leaves mediated by trehalose 6-phosphate. *Plant Physiol.* **163**, 1142–1163 (2013).
57. Xiang, L. et al. Exploring the neutral invertase–oxidative stress defence connection in *Arabidopsis thaliana*. *J. Exp. Bot.* **62**, 3849–3862 (2011).
58. Sahn, A., Bens, M., Platzer, M., Szafranski, K. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Res.* **45**, e100, (2017).
59. Ito, N. & Phillips, S. E. Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase. *Nature* **350**, 87 (1991).
60. Black, C. C. & Osmond, C. B. Crassulacean acid metabolism photosynthesis: ‘working the night shift’. *Photosynth. Res.* **76**, 329–341 (2003).
61. De Paoli, H. C., Tuskan, G. A. & Yang, X. H. An innovative platform for quick and flexible joining of assorted DNA fragments. *Sci. Rep.* **6**, 19278 (2016).
62. Wake, D. B., Wake, M. H. & Specht, C. D. Homoplasy: from detecting pattern to determining process and mechanism of evolution. *Science* **331**, 1032–1035 (2011).
63. Laird, J. The law of parsimony. *The Monist* **29**, 321–344 (1919).
64. Fujiki, Y. et al. Dark-inducible genes from *Arabidopsis thaliana* are associated with leaf senescence and repressed by sugars. *Physiol. Plant.* **111**, 345–352 (2001).
65. Puente, P., Wei, N. & Deng, X. W. Combinatorial interplay of promoter elements constitutes the minimal determinants for light and developmental control of gene expression in *Arabidopsis*. *EMBO J.* **15**, 3732 (1996).
66. Kajitani, R. et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* **24**, 1384–1395 (2014).
67. Bosi, E. et al. MeDuSa: a multi-draft based scaffold. *Bioinformatics* **31**, (2443–2451) (2015).
68. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
69. Gotz, S. et al. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
70. Conesa, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
71. Ekstrom, A., Taujale, R., McGinn, N. & Yin, Y. PlantCAZyme: a database for plant carbohydrate-active enzymes. *Database (Oxford)* **2014**, bau079 (2014).
72. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, 1 (2013).
73. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
74. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
75. 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).
76. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* **5**, 725–738 (2010).
77. Yang, J. et al. The I-TASSER Suite: protein structure and function prediction. *Nat. Methods* **12**, 7–8 (2015).
78. Stamatakis, A. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
79. Mirarab, S. & Warnow, T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, i44–i52 (2015).

## Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. This research was supported by the U.S. Department of Energy, Office of Science, Genomic Science Program under Award Number DE-SC0008834. Additional support was provided by the UK Biotechnology and Biological Sciences Research Council (grant no. BB/F009313/1) and the Laboratory Directed Research and Development (LDRD) Program (Project ID: 7758) of Oak Ridge National Laboratory. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory. This research also used the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory. We thank Daniel Rokhsar, Mary Ann Cushman, and Lee Gunter for critical review and comments on the manuscript and Lori Kunder (Kunder Design Studio) for assistance with figure preparation. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE-AC05-00OR22725.

## Author contributions

X.Y. conceived and initiated the *Kalanchoë* genome project, supervised the study and interpreted the data, and wrote the manuscript; R.H. carried out protein function characterization, data analysis, and wrote the manuscript; H.Y., S.J., and P.P. carried out genome sequencing and RNA-seq; J.J. and J.S. carried out genome assembly; S.S. and D.M.G. carried out genome annotation; H.T. carried out genome duplication, synteny analysis, and wrote the manuscript; J.H., S.F.B., and L.V.D. contributed to material, interpreted the data, and wrote the manuscript; C.-M. W., R.V.B., and R.M. contributed pineapple genomics and gene expression data; D.A.W., R.C.M., and D.A.J. carried out convergent expression analysis; P.E.A. and R.L.H. carried out GO and metabolite data analysis; K.W. interpreted the data; J.A.C.S. interpreted the data and wrote the

manuscript; E.F., R.M., H.C.D.P., A.M.B., and Y.Y. carried out metabolic pathway analysis; Z.Z. and T.J.T. carried out metabolite profiling; J.D.B. carried out phylogenetic analysis; H.-B.G. and H.G. carried out phylogenetic analysis and protein structure modeling; K.H. and J.H.L.-M. carried out phylogenetic analysis, protein structure clustering, and wrote the manuscript; J.M.H. and K.J.P. carried out ploidy analysis; M.X., J.-G. C. and W.M. contributed to protein function characterization; W.C.Y. carried out RNA-seq data analysis and interpreted the data; D.L. carried out stomatal and circadian gene analysis; R.A., T.G., J.A.M., and S.-D.L. contributed to transcriptome and genome sequencing; J.G. contributed to transcriptome sequencing; J.C.C. contributed to transcriptome, genome sequencing, and wrote the manuscript; G.A.T. conceived the study and interpreted the data. All authors read and commented on manuscript.

### Additional information

**Supplementary Information** accompanies this paper at doi:10.1038/s41467-017-01491-7.

**Competing interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

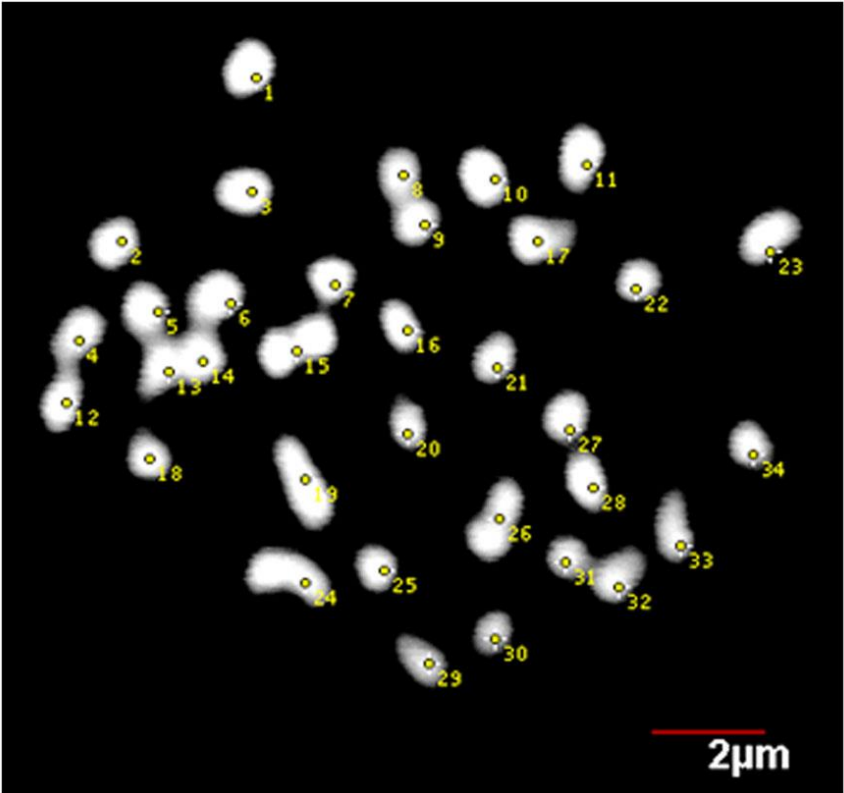
© The Author(s) 2017

Xiaohan Yang<sup>1,2</sup>, Rongbin Hu<sup>1</sup>, Hengfu Yin<sup>1</sup>, Jerry Jenkins<sup>3</sup>, Shengqiang Shu<sup>4</sup>, Haibao Tang<sup>5</sup>, Degao Liu<sup>1</sup>, Deborah A. Weighill<sup>1,2</sup>, Won Cheol Yim<sup>6</sup>, Jungmin Ha<sup>6</sup>, Karolina Heyduk<sup>7</sup>, David M. Goodstein<sup>4</sup>, Hao-Bo Guo<sup>8</sup>, Robert C. Moseley<sup>1,2</sup>, Elisabeth Fitzek<sup>9</sup>, Sara Jawdy<sup>1</sup>, Zhihao Zhang<sup>1</sup>, Meng Xie<sup>1</sup>, James Hartwell<sup>10</sup>, Jane Grimwood<sup>3</sup>, Paul E. Abraham<sup>11</sup>, Ritesh Mewalal<sup>1</sup>, Juan D. Beltrán<sup>12</sup>, Susanna F. Boxall<sup>10</sup>, Louisa V. Dever<sup>10</sup>, Kaitlin J. Palla<sup>1,2</sup>, Rebecca Albion<sup>6</sup>, Travis Garcia<sup>6</sup>, Jesse A. Mayer<sup>6</sup>, Sung Don Lim<sup>6</sup>, Ching Man Wai<sup>13</sup>, Paul Peluso<sup>14</sup>, Robert Van Buren<sup>15</sup>, Henrique Cestari De Paoli<sup>1,16</sup>, Anne M. Borland<sup>1,17</sup>, Hong Guo<sup>8</sup>, Jin-Gui Chen<sup>1</sup>, Wellington Muchero<sup>1</sup>, Yanbin Yin<sup>9</sup>, Daniel A. Jacobson<sup>1,2</sup>, Timothy J. Tschaplinski<sup>1</sup>, Robert L. Hettich<sup>11</sup>, Ray Ming<sup>5,13</sup>, Klaus Winter<sup>18</sup>, James H. Leebens-Mack<sup>7</sup>, J. Andrew C. Smith<sup>12</sup>, John C. Cushman<sup>6</sup>, Jeremy Schmutz<sup>3,4</sup> & Gerald A. Tuskan<sup>1</sup>

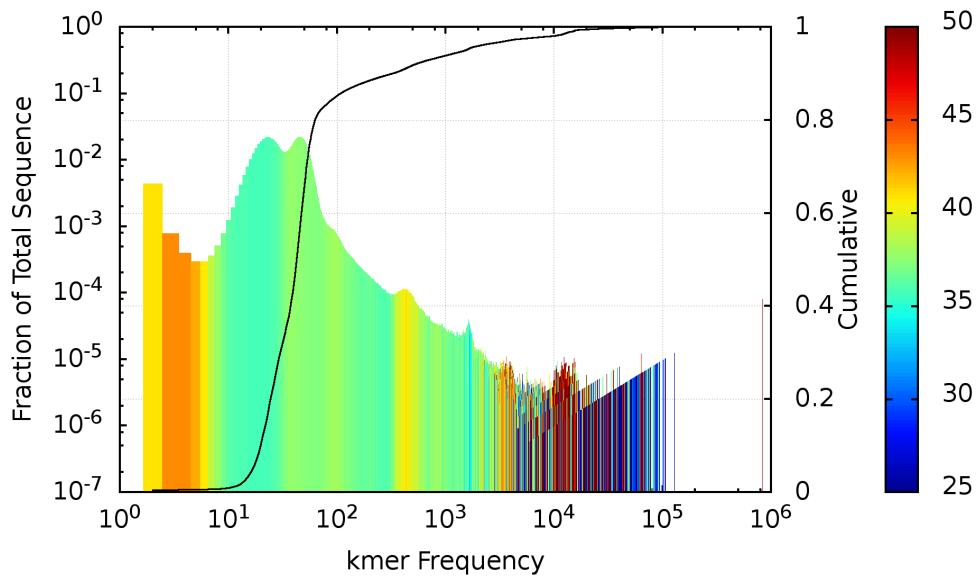
<sup>1</sup>Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. <sup>2</sup>The Bredesen Center for Interdisciplinary Research and Graduate Education, University of Tennessee, Knoxville, TN 37996, USA. <sup>3</sup>HudsonAlpha Institute for Biotechnology, 601 Genome Way, Huntsville, AL 35801, USA. <sup>4</sup>US Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. <sup>5</sup>Center for Genomics and Biotechnology, Fujian Provincial Key Laboratory of Haixia Applied Plant Systems Biology, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China. <sup>6</sup>Department of Biochemistry and Molecular Biology, University of Nevada, Reno, NV 89557, USA. <sup>7</sup>Department of Plant Biology, University of Georgia, Athens, GA 30602, USA. <sup>8</sup>Department of Biochemistry & Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA. <sup>9</sup>Department of Biological Sciences, Northern Illinois University, DeKalb, IL 60115, USA. <sup>10</sup>Department of Plant Sciences, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK. <sup>11</sup>Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. <sup>12</sup>Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK. <sup>13</sup>Department of Plant Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. <sup>14</sup>Pacific Biosciences, Inc., 940 Hamilton Avenue, Menlo Park, CA 94025, USA. <sup>15</sup>Department of Horticulture, Michigan State University, East Lansing, MI 48824, USA. <sup>16</sup>Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996, USA. <sup>17</sup>School of Natural and Environmental Science, Newcastle University, Newcastle upon Tyne NE1 7RU, UK. <sup>18</sup>Smithsonian Tropical Research Institute, Apartado, Balboa, Ancón 0843-03092, Republic of Panama



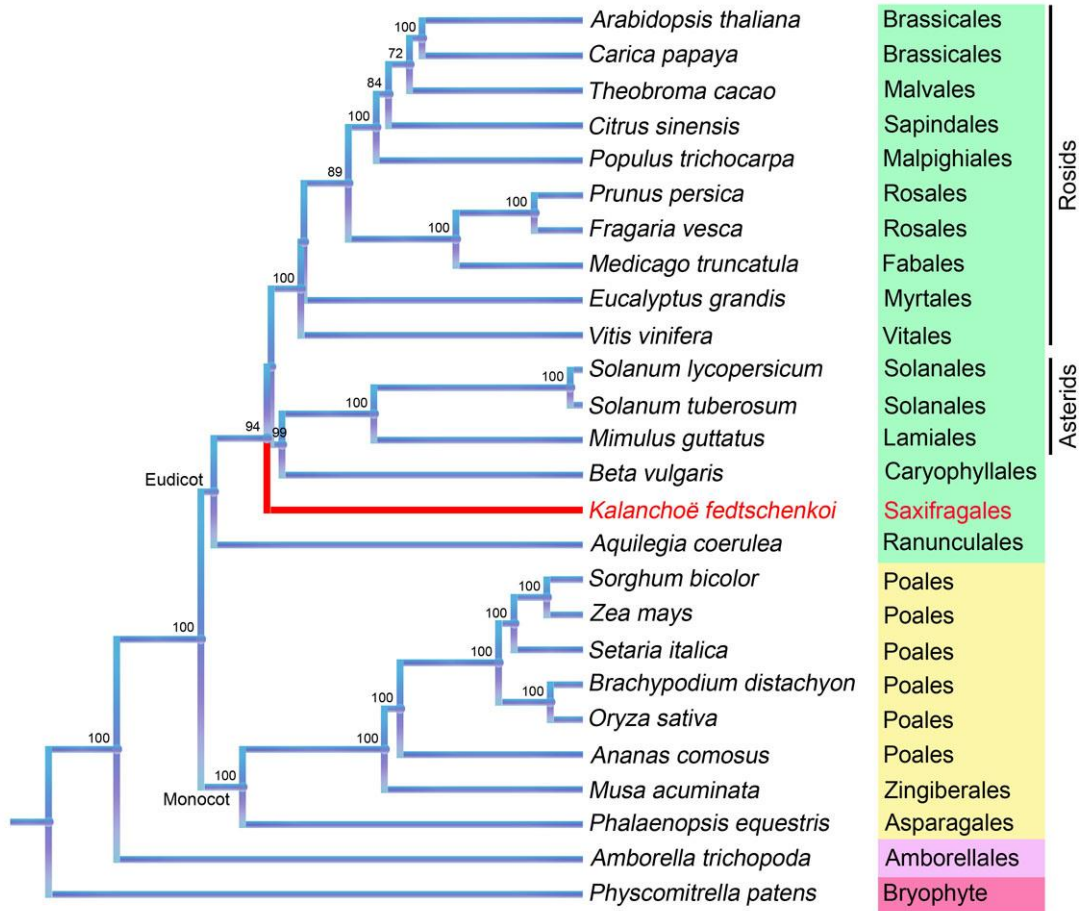
Supplementary Figures



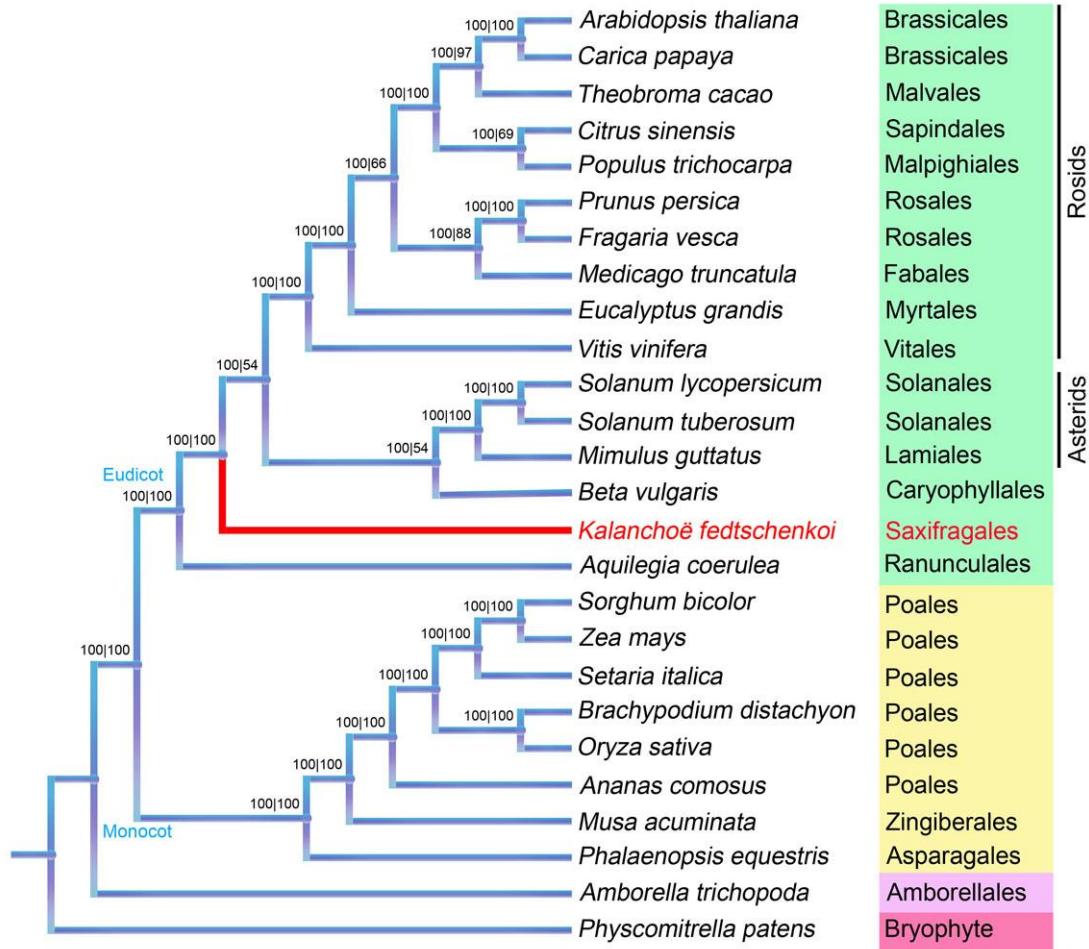
Supplementary Figure 1. Chromosome count of *Kalanchoë fedtschenkoi*.



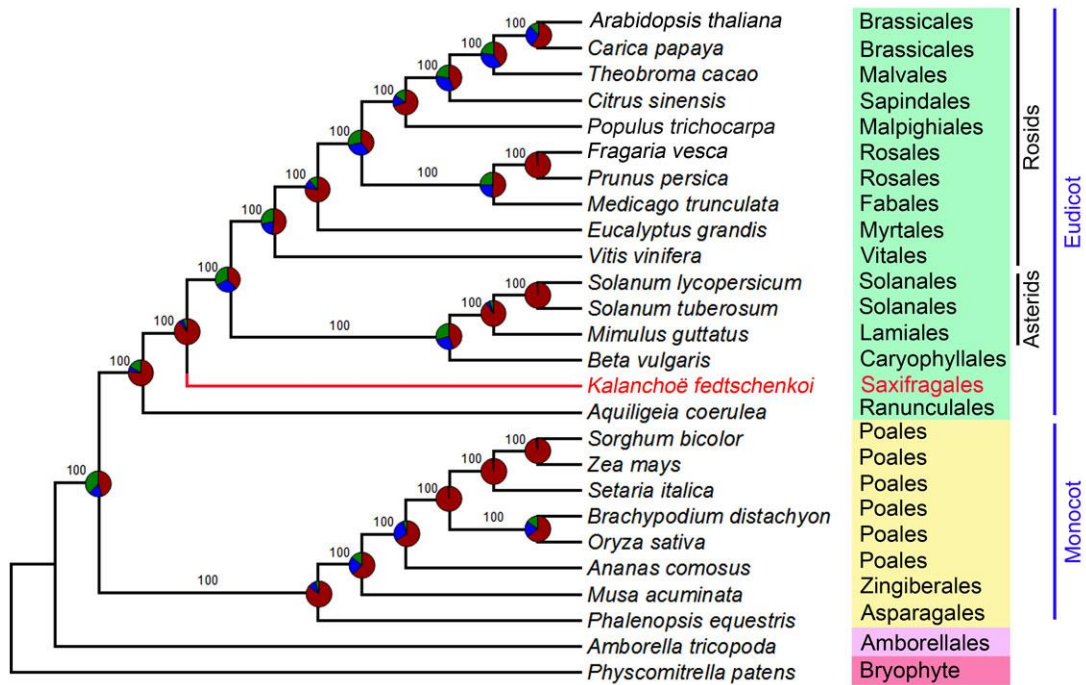
**Supplementary Figure 2. Plot of the fraction of total sequence against 24-mer frequency for *Kalanchoë fedtschenkoi*.** Each of the individual 24mer frequency bins are colored by the average GC for k-mers in that bin. The solid black line is the cumulative fraction of total sequence for the particular library on the secondary y-axis.



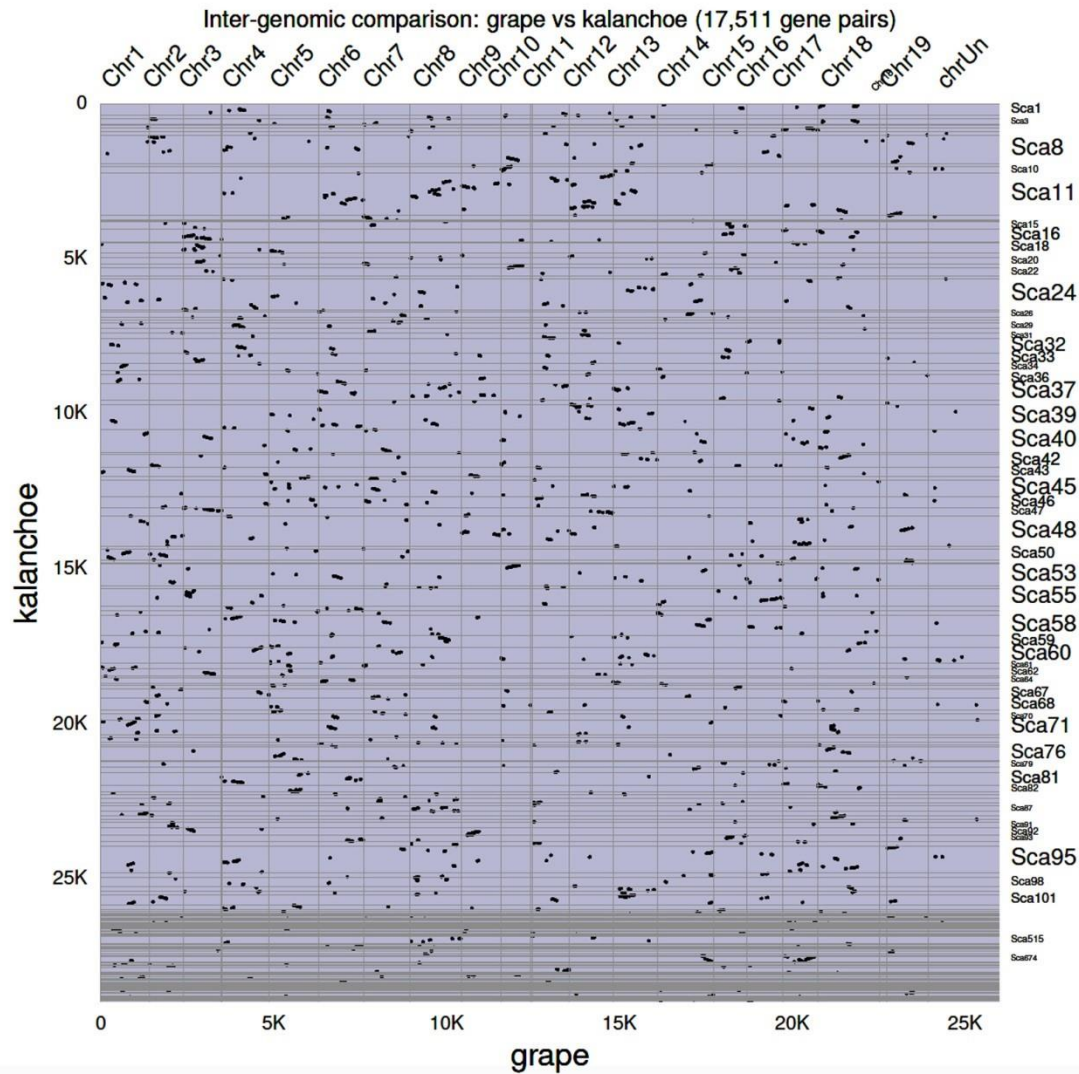
**Supplementary Figure 3. A phylogenetic tree created from concatenated protein sequence alignment of 210 single-copy nuclear genes using maximum-likelihood method <sup>1</sup>. The numbers at each node are percent bootstrap support values from the maximum likelihood analysis.**



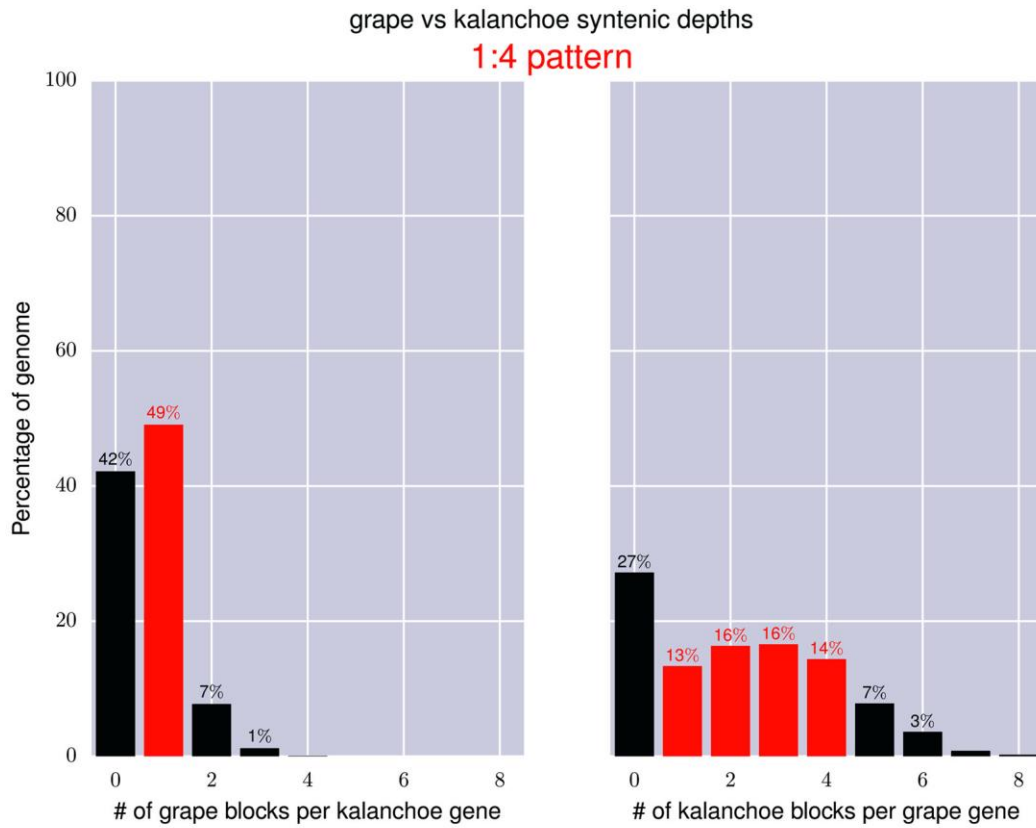
**Supplementary Figure 4. Phylogenetic tree created from partitioned analysis of concatenated protein sequence alignment of 210 single-copy nuclear genes.** The numbers on the left side of “|” are the percent probability from Bayesian inference using MrBayes <sup>2</sup>. The numbers on the right side of “|” are the bootstrap support values from maximum likelihood inference using IQ-TREE <sup>1</sup>.



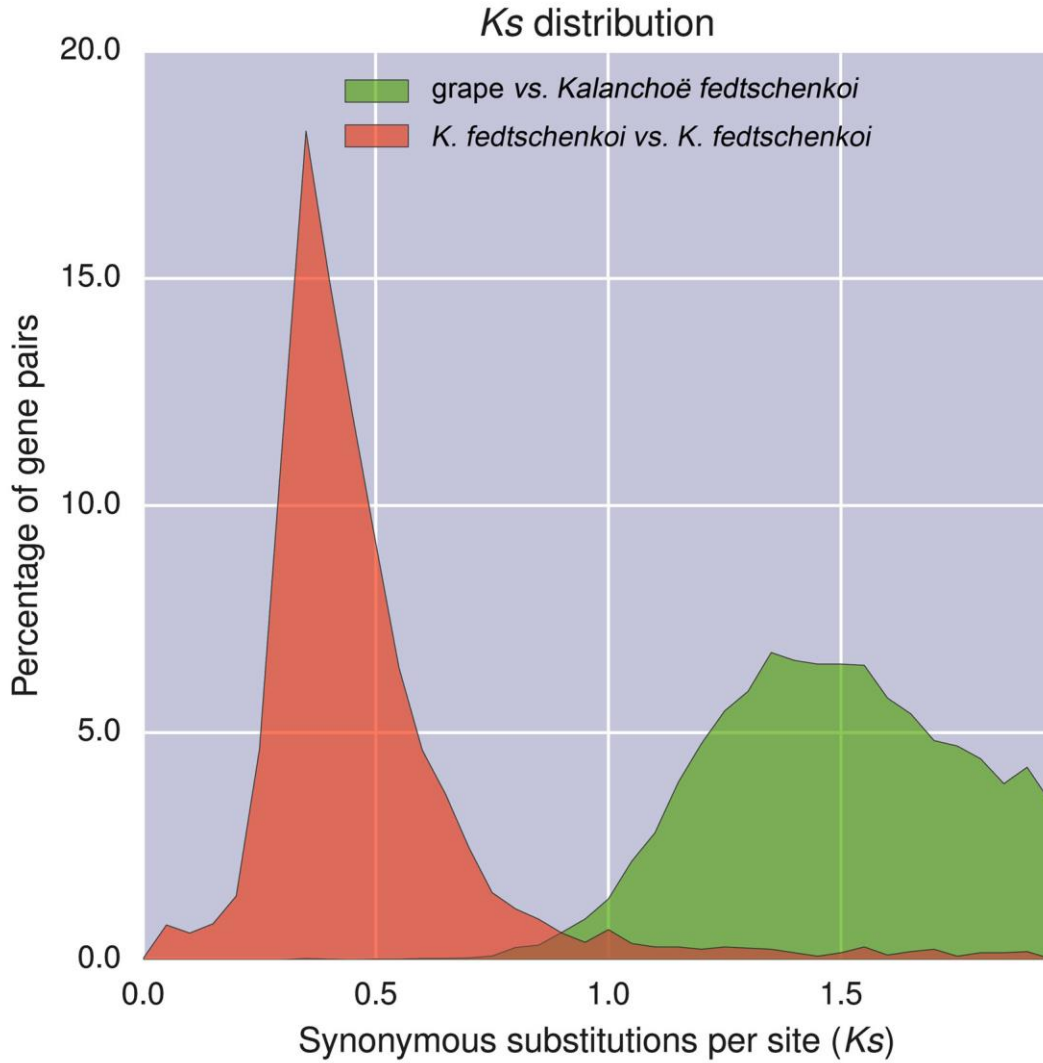
**Supplementary Figure 5. A species tree reconstructed from 210 single copy genes using a fully Bayesian multispecies coalescent method.** The coding-sequence (CDS) alignments for each of the 210 single-copy-gene ortholog groups were used to reconstruct the multispecies coalescent tree using starBEAST2<sup>3</sup>. Pie graphs on nodes represent the proportion of gene trees that support the various quartets at every node, with red for the main topology shown in this tree, blue for the first alternative and green for the second alternative, respectively. Quartet frequencies displayed in pie graphs and the numbers representing percent posterior-probability at each node were calculated by ASTRAL-II<sup>4</sup>.



**Supplementary Figure 6. Syntenic dot plot between the grape genome and the *Kalanchoë fedtschenkoi* genome.** Each dot represents a homologous gene pair retained in a synteny block with at least four gene pairs per block.



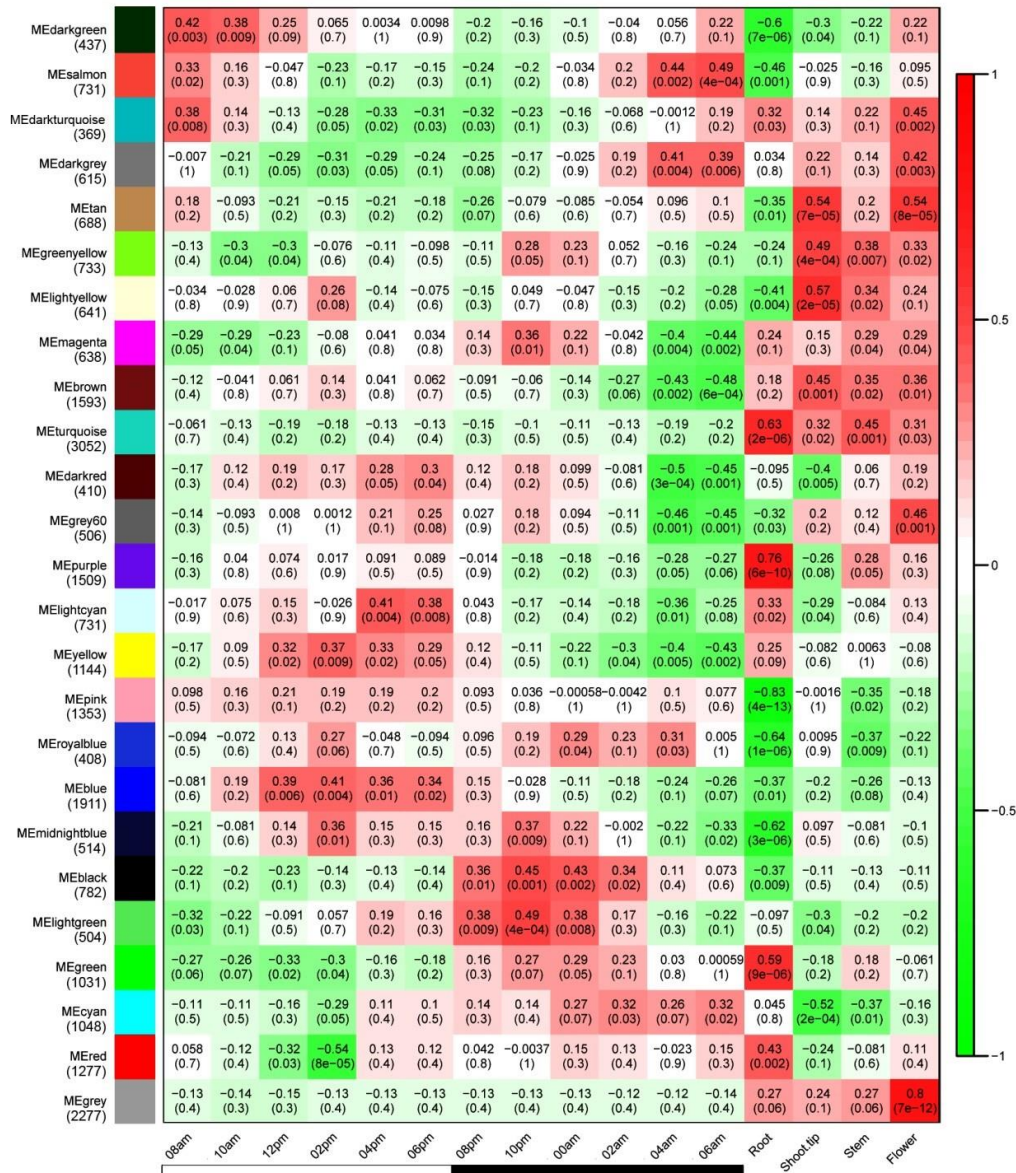
**Supplementary Figure 7. Syntenic depths in grape vs. *Kalanchoë* genome comparison.** The mode of each count distribution corresponds to the synteny patterns. From *Kalanchoë* point of view (left), we observed predominantly one single grape-*Kalanchoë* synteny block, suggesting no duplications in the grape lineage since the divergence; from grape point of view (right), we observed up to four grape-*Kalanchoë* synteny blocks, suggesting two whole-genome duplications (WGDs) in the *Kalanchoë* lineage since their divergence. We only included reciprocal best hits in order to remove orthologs due to shared WGDs.



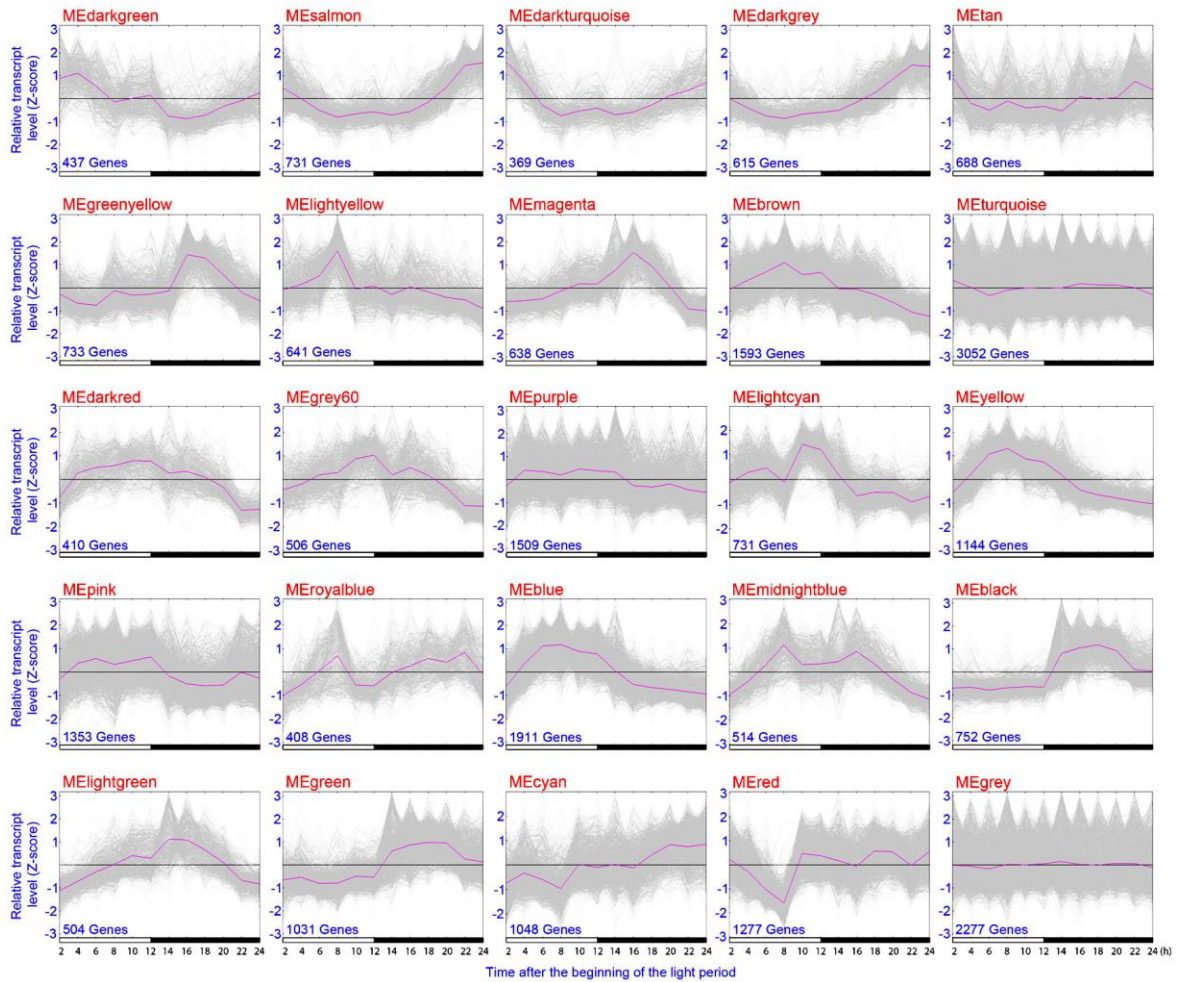
**Supplementary Figure 8. Synonymous substitutions per synonymous site ( $K_s$ ) distribution for syntenic gene pairs.**  $K_s$  values were calculated per syntenic gene pair in grape-*Kalanchoë fedtschenkoi* and *K. fedtschenkoi*-*K. fedtschenkoi* comparisons.



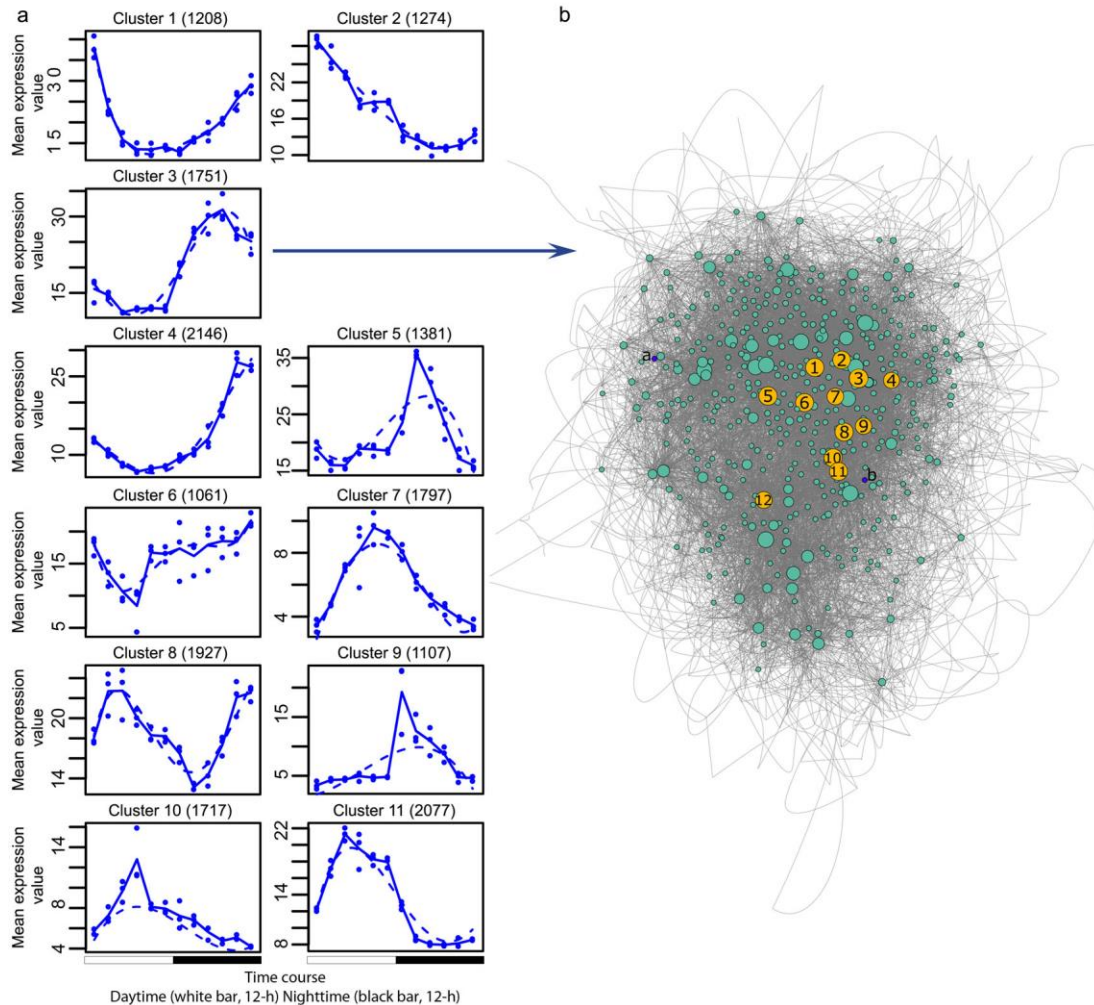
## Module-trait relationships



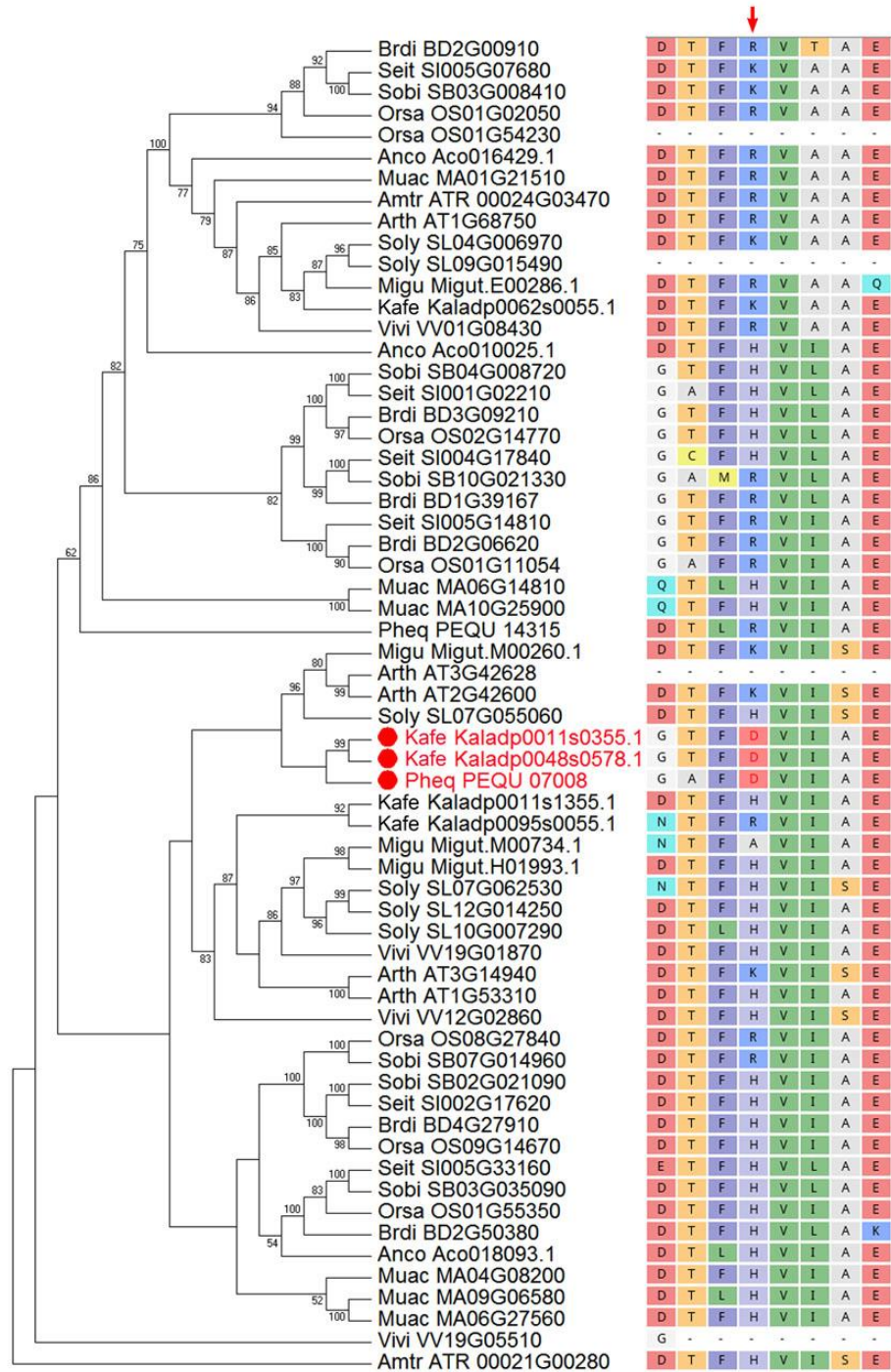
**Supplementary Figure 9. The relationship between co-expression module and trait in *Kalanchoë fedtschenkoi*.** The color blocks on the left side are the co-expression modules (named as MEdarkgreen, MEsalmon, etc.) constructed from RNA-seq data. The numbers underneath the module names are the number of genes in each module. The number represents the correlation between co-expression module and trait. The numbers in parentheses represent the *p*-value (Student's t-test) of the correlation. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



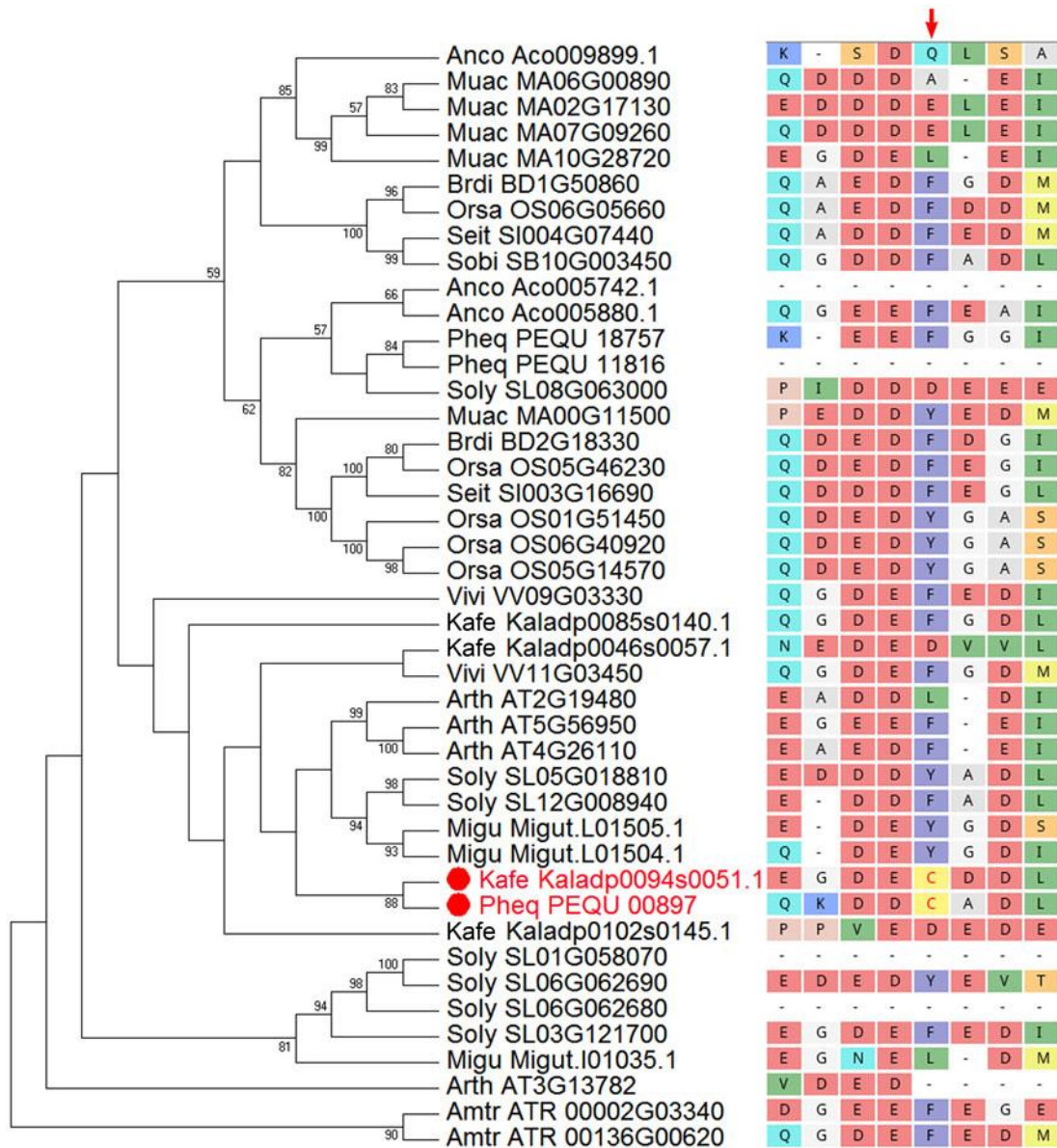
**Supplementary Figure 10. Transcript changes of genes in each co-expression modules** (Supplementary Fig. 9). The grey lines in each module indicate expression pattern of individual gene in the modules. The purple line represents the median pattern of expression and total gene number was shown in individual module figures. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



**Supplementary Figure 11. Cluster analysis of diel expression of *Kalanchoë fedtschenkoi* genes that showed significant changes in transcript abundance over a 24-h period. (A)** Diel transcript expression profile of the 11 gene clusters constructed, using maSigPro<sup>5,6</sup>, from the transcripts that showed significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) “non-flat” diel expression patterns as determined by a polynomial regression. Dots are the median expression value for all genes in the cluster at a given time-point, the solid line is the median fit line across all time-points, and the dashed line is the cluster’s polynomial fit. **(B)** A network constructed for Cluster 3 using ARACNE<sup>7</sup>. Node sizes are scaled by the number of directed connections, or edges. The top 1% of each network’s nodes are highlighted and numbered, with annotations found in [Supplementary Data 2](#). Number labels are not indicative of connectivity of the node. Nodes with less than 10 directed edges are not pictured. Blue colored nodes have putative CAM function, and are also annotated in [Supplementary Data 2](#).

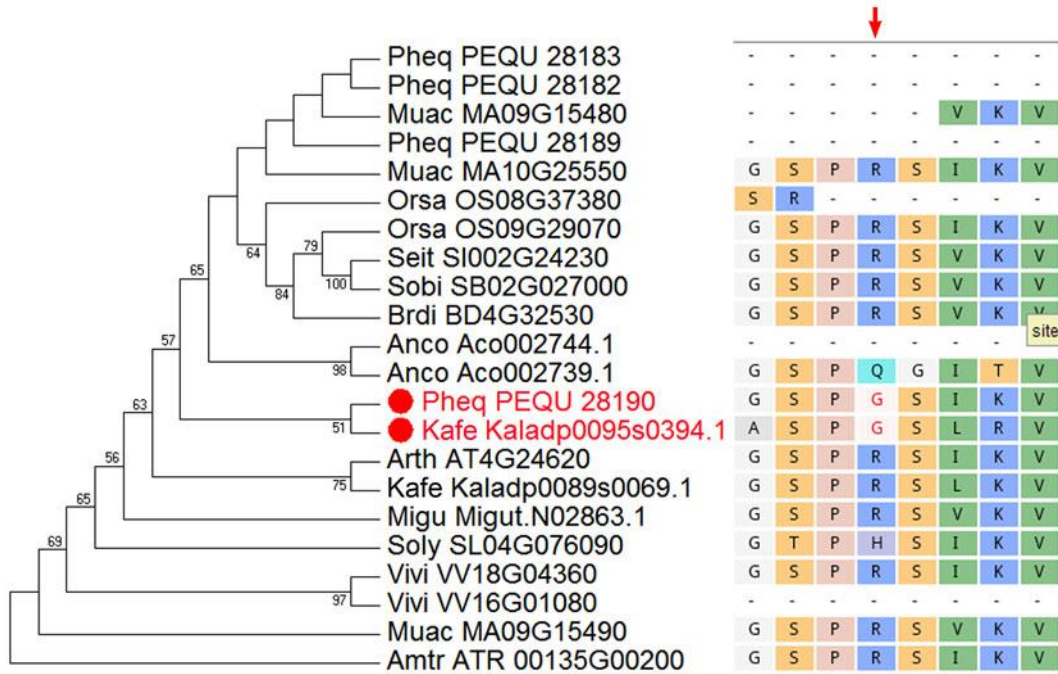


**Supplementary Figure 12. A maximum-likelihood phylogeny of tribe I50\_F000807 that contains the gene encoding phosphoenolpyruvate carboxylase (PEPC).** The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. Red dots highlight the genes showing convergent evolution in protein sequence. The red arrow indicates the protein sequence alignment position where the mutation (H/K/R-to-D) occurred.

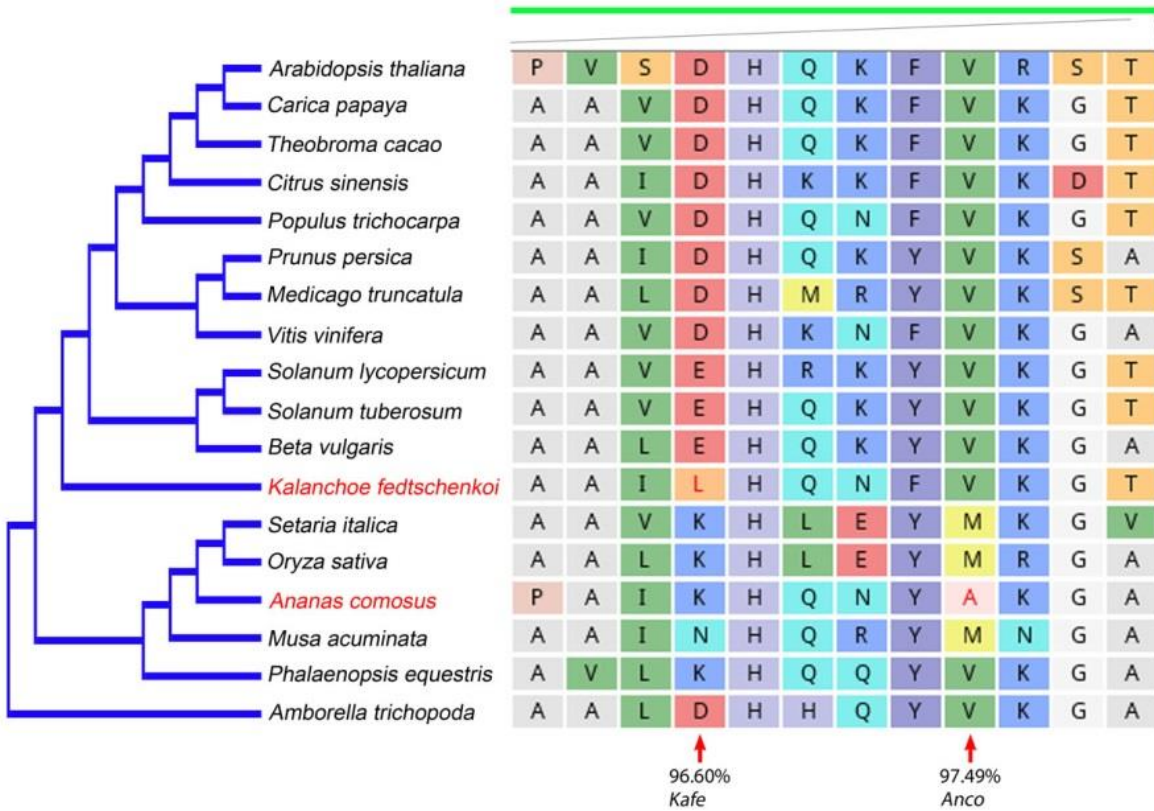


**Supplementary Figure 13. A maximum-likelihood phylogeny of tribe I50\_F001102 that contains the gene encoding nucleosome assembly protein 1 4-like (NAP1L4).** The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. Red dots highlight the genes showing convergent evolution in protein sequence. The red arrow indicates the protein sequence alignment position where the mutation (F/Y-to-C) occurred.



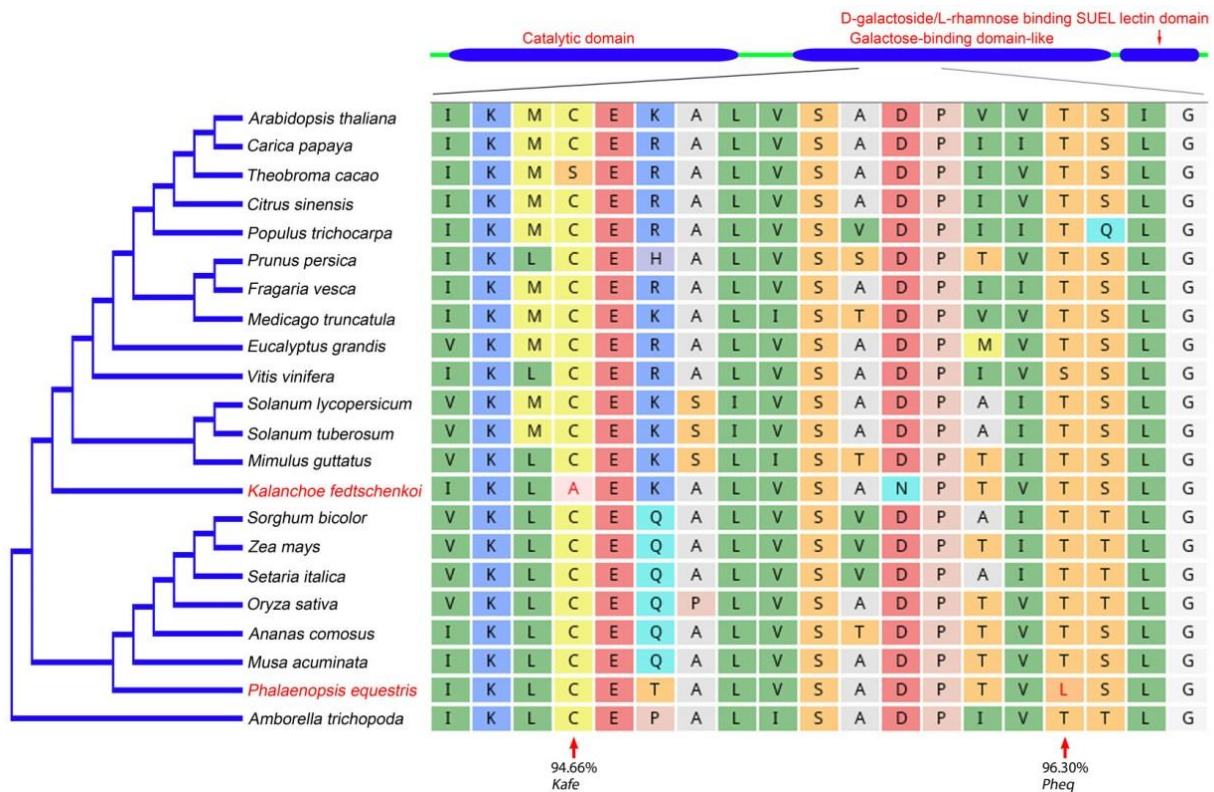


**Supplementary Figure 15. A maximum-likelihood phylogeny of tribe I50\_F003508 that contains the gene encoding a putative chloroplast-localized glucose-6-phosphate isomerase (GPI).** The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. Red dots highlight the genes showing convergent evolution in protein sequence. The red arrow indicates the protein sequence alignment where the mutation (H/Q/R-to-G) occurred.

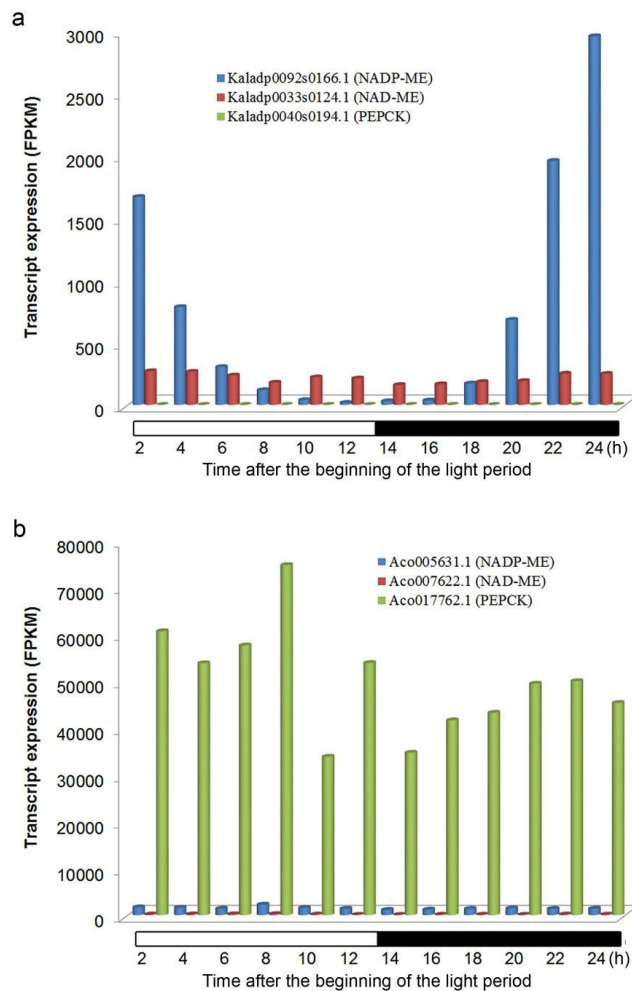


**Supplementary Figure 16. A gene encoding a putative sialyltransferase-like protein (SIA2) under positive selection in *Kalanchoë fedtschenkoi* and *Ananas comosus*.** Specifically, two genes, Kaladp0016s0058 in *K. fedtschenkoi* and Aco018360.1 in *A. comosus*, were identified to be under positive selection (CodeML implemented in PosiGene<sup>8</sup>,  $P < 0.05$ ) in comparison with their orthologs in the non-CAM species listed in this figure. The CAM species that under positive selection were highlighted in red. The arrows indicate the specific mutation positions in individual CAM species and the percentage indicates the mutation probability in that position as compared to non-CAM species. “Kafe” is the abbreviation for *Kalanchoë fedtschenkoi* and “Anco” is the abbreviation for *Ananas comosus*.

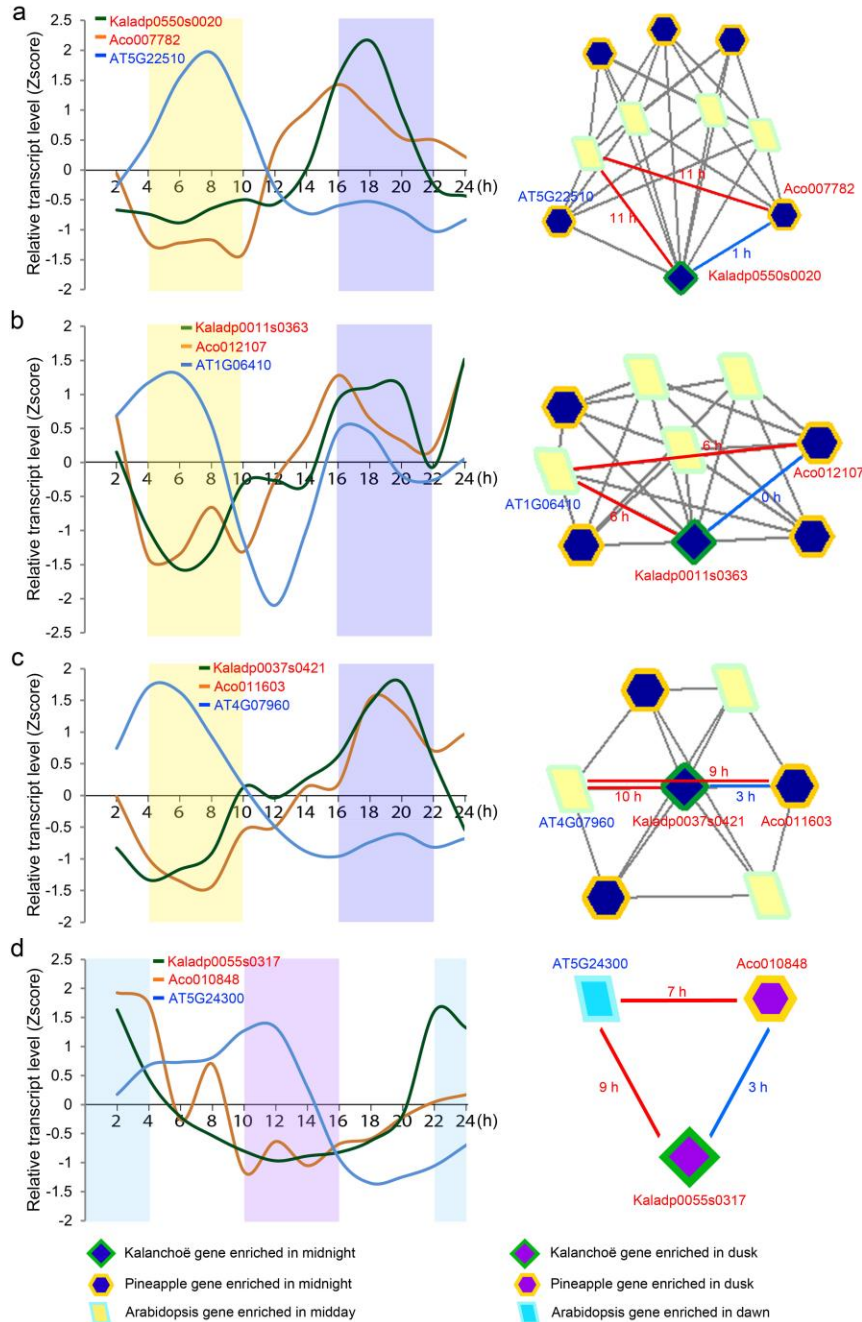




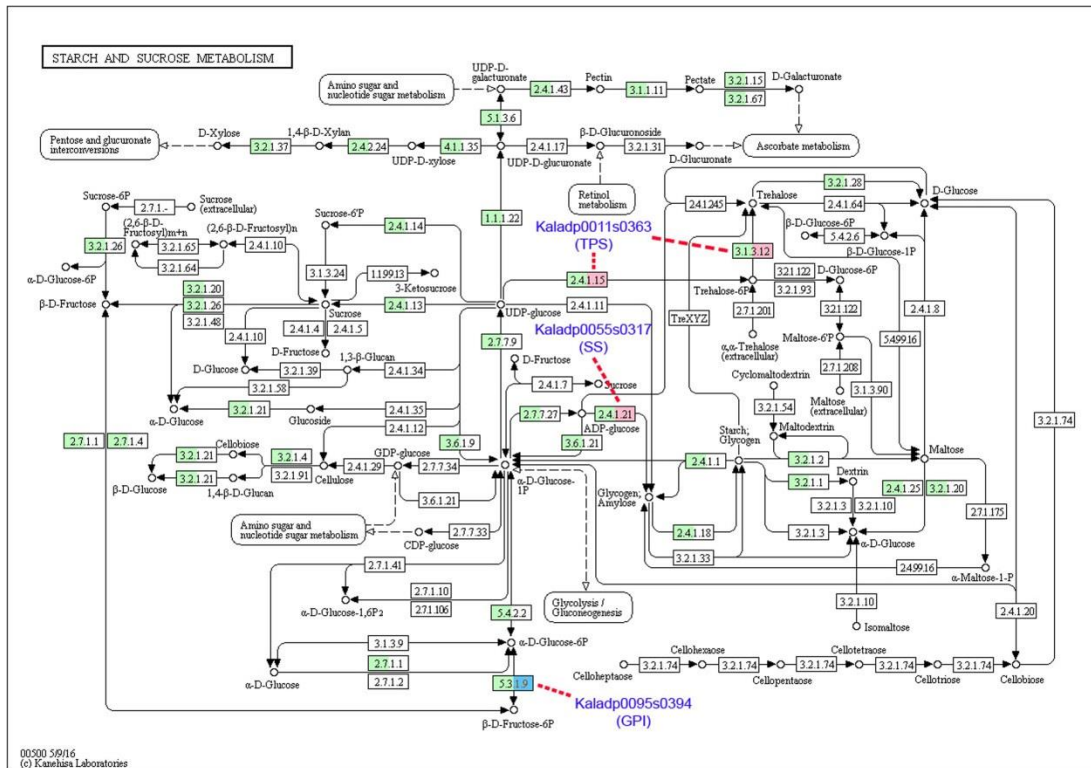
**Supplementary Figure 17. A gene encoding a beta-galactosidase protein (BGAL3) under positive selection in *Kalanchoë fedtschenkoi* and *Phalaenopsis equestris*.** Specifically, two genes, Kaladp0067s0114 in *K. fedtschenkoi* and PEQU\_04899 in *P. equestris*, were identified to be under positive selection (CodeML implemented in PosiGene<sup>8</sup>,  $P < 0.05$ ) in comparison with their orthologs in the non-CAM species listed in this figure. The CAM species that under positive selection were highlighted in red. The arrows indicate the specific mutation positions in individual CAM species and the percentage indicates the mutation probability in that position as compared to non-CAM species. “Kafe” is the abbreviation for *Kalanchoë fedtschenkoi* and “Pheq” is the abbreviation for *Phalaenopsis equestris*.



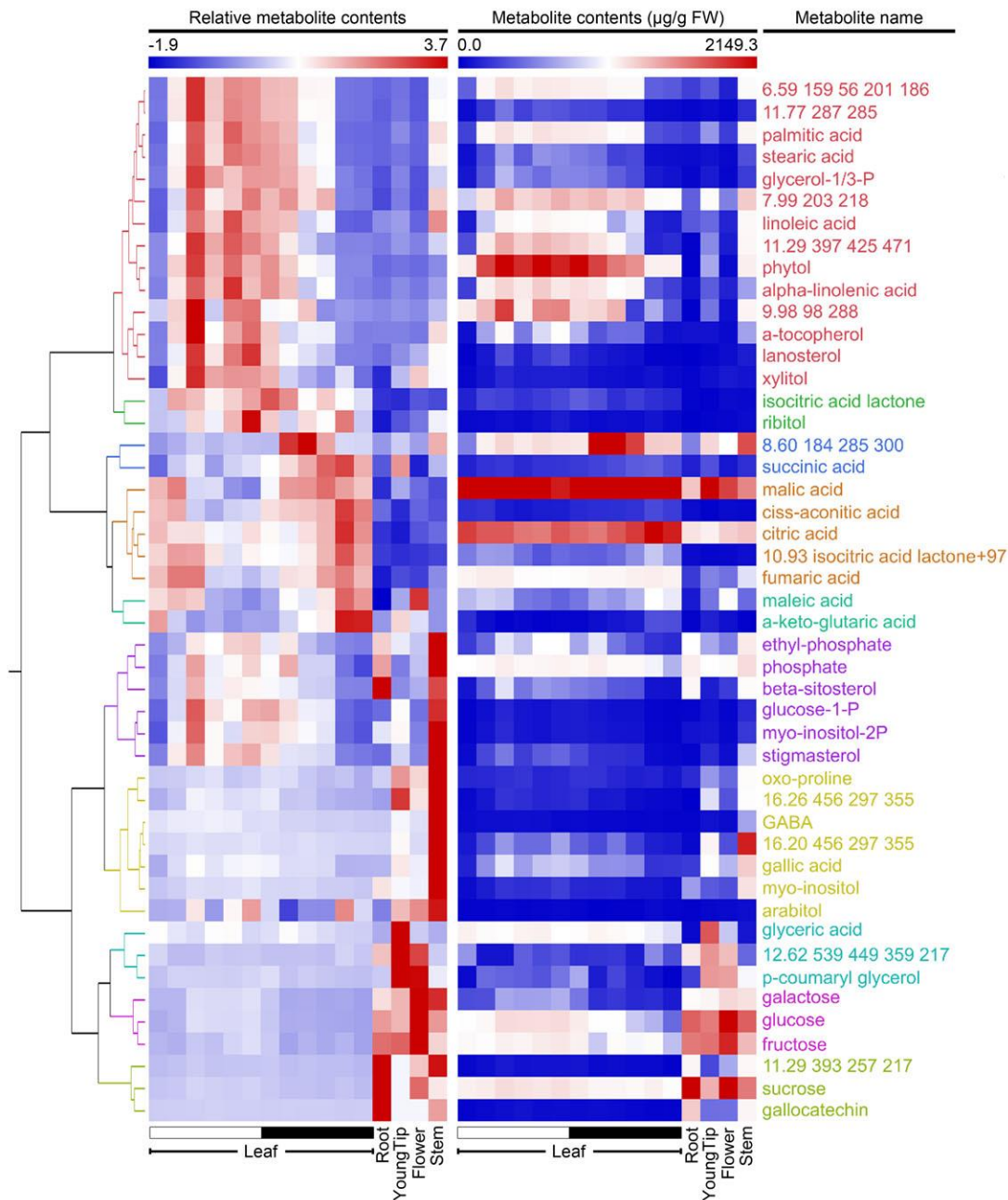
**Supplementary Figure 18. Diel expression of genes involved in the decarboxylation process in *Kalanchoë fedtschenkoi* and pineapple. (a) Diel transcript expression of decarboxylation genes in *K. fedtschenkoi*. (b) Diel transcript expression of decarboxylation genes in pineapple. NAD-ME, NAD-malic enzyme; NADP-ME, NADP-malic enzyme; PEPCK, phosphoenolpyruvate carboxykinase. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.**



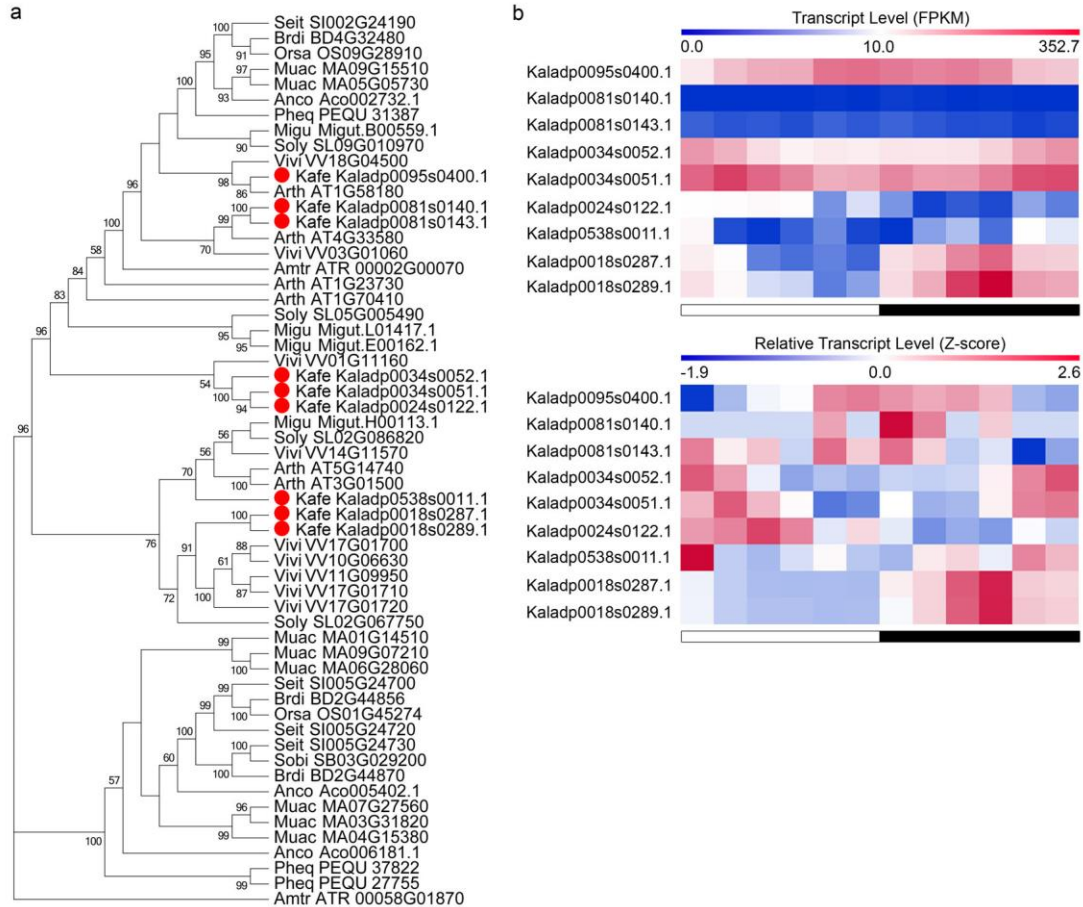
**Supplementary Figure 19. Enrichment of triangle networks of genes relevant to CAZyme in *Kalanchoë fedtschenkoi*, pineapple and *Arabidopsis*.** (a) Network of orthogroup (ORTHOMCL68) containing GH100 genes (CINV1). (b) Network of orthogroup (ORTHOMCL93) containing GT20 genes (TPS). (c) Network of orthogroup (ORTHOMCL207) containing GT2 genes (CSLC12). (d) Network of orthogroup (ORTHOMCL9830) containing GT5 genes (Starch synthase). White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively. X-axis represents the time after the beginning of the light period.



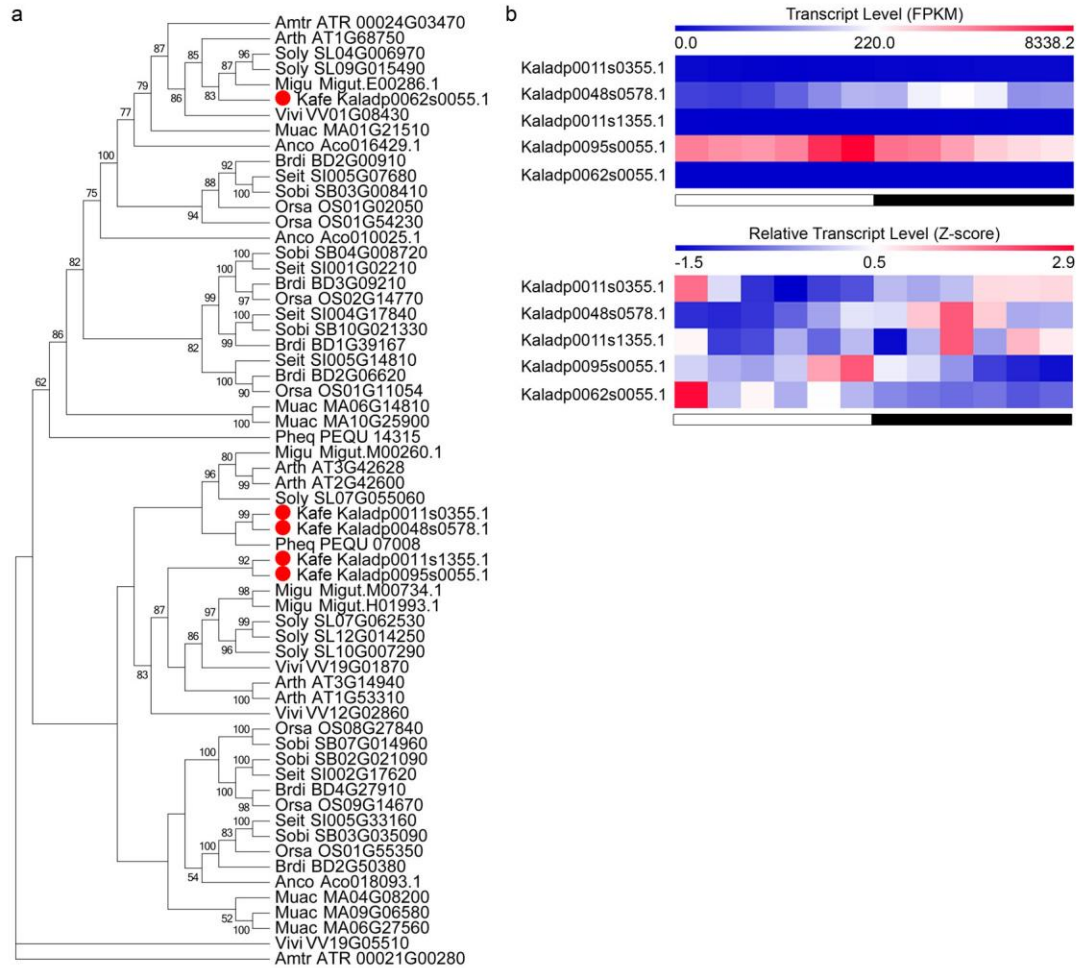
**Supplementary Figure 20. An overview KEGG map of starch and sucrose metabolism. Green color indicates the genes identified in *Kalanchoë fedtschenkoi*. Pink color highlights the *K. fedtschenkoi* genes (Kaladp0011s0363 and Kaladp0055s0317) showing convergent evolution with *Ananas comosus* (pineapple) in diel expression pattern. Blue color highlights the *K. fedtschenkoi* gene (Kaladp0095s0394) showing convergent evolution in protein sequence with *Phalaenopsis equestris* (orchid). GPI: glucose-6-phosphate isomerase chloroplast-like; SS: starch synthase chloroplast amyloplast-like; TPS: trehalose phosphate synthase.**



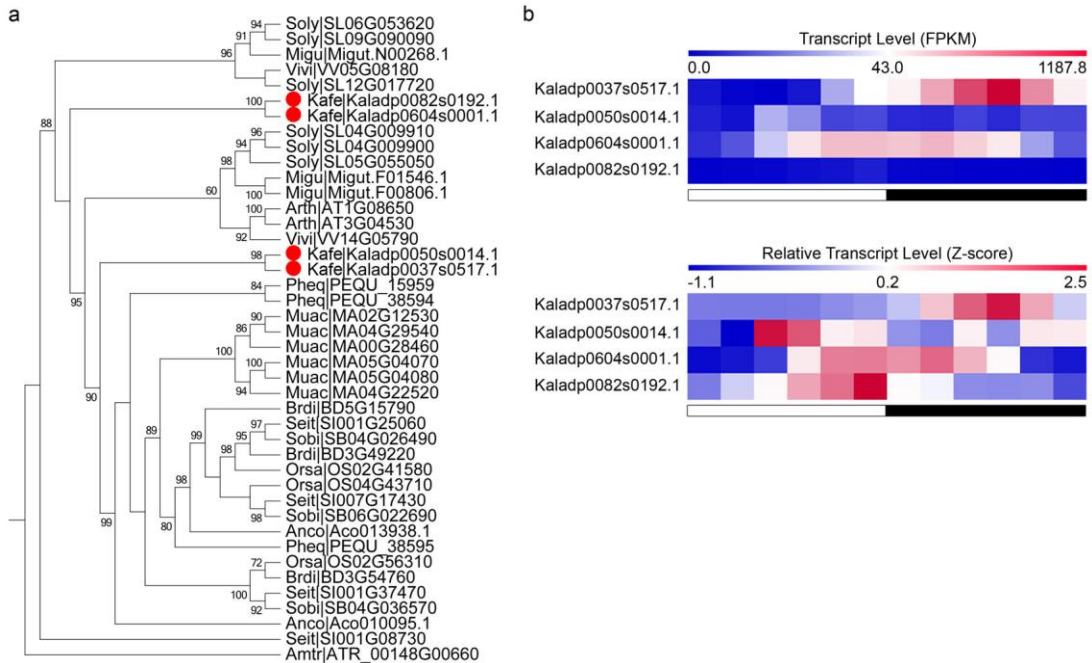
**Supplementary Figure 21. A hierarchical clustering (Fast Ward method) heatmap highlights 47 quantified metabolite contents in *Kalanchoë fedtschenkoi*.** The black and white bars indicate nighttime (12-hour) and daytime (12-hour), respectively. The numbers (e.g., “11.77 287 285”, “9.98 98 288”) on the right side indicate retention times of metabolites with unknown identity.



**Supplementary Figure 22. Phylogeny of the beta-carbonic anhydrase ( $\beta$ -CA) proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the  $\beta$ -CA genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.

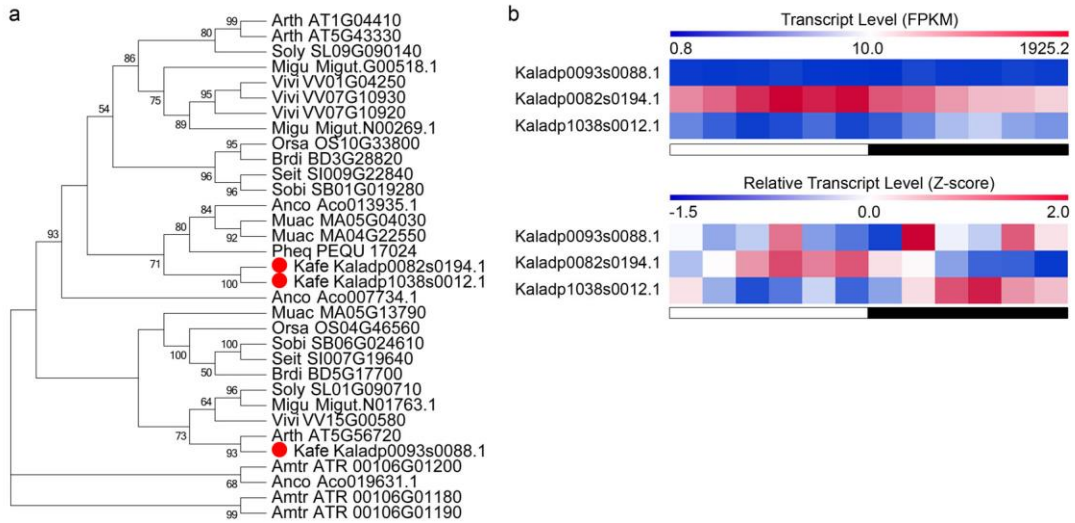


**Supplementary Figure 23. Phylogeny of the phosphoenolpyruvate carboxylase (PEPC) proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *PEPC* genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.

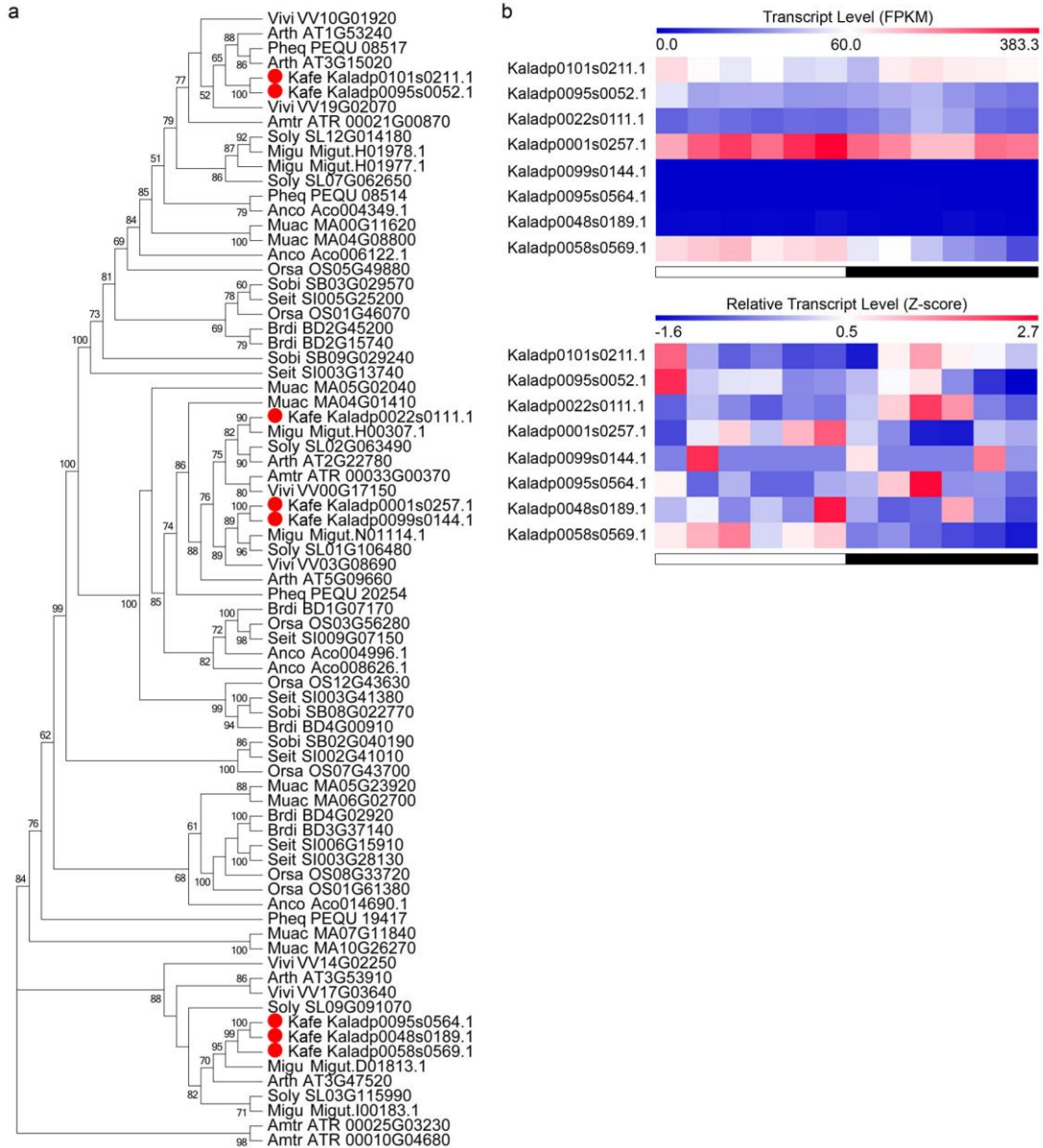


**Supplementary Figure 24. Phylogeny of the phosphoenolpyruvate carboxylase kinase (PPCK) proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *PPCK* genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.

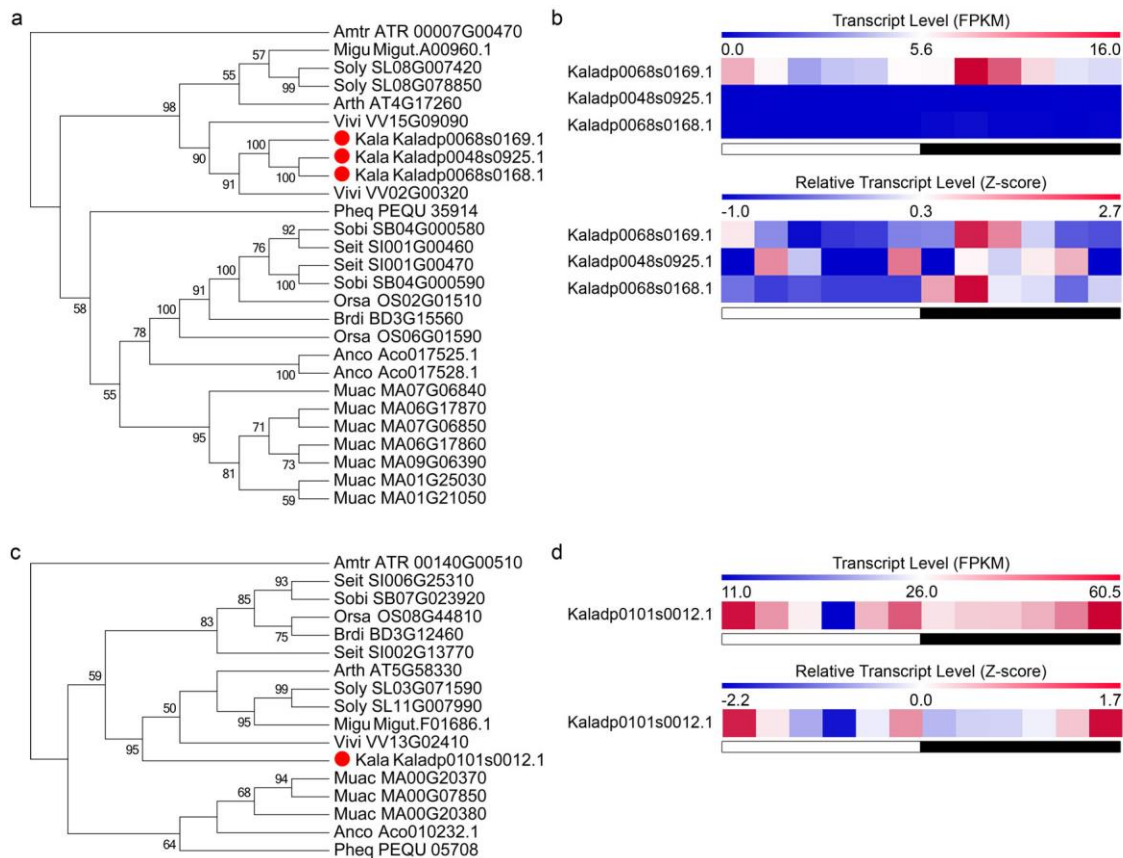




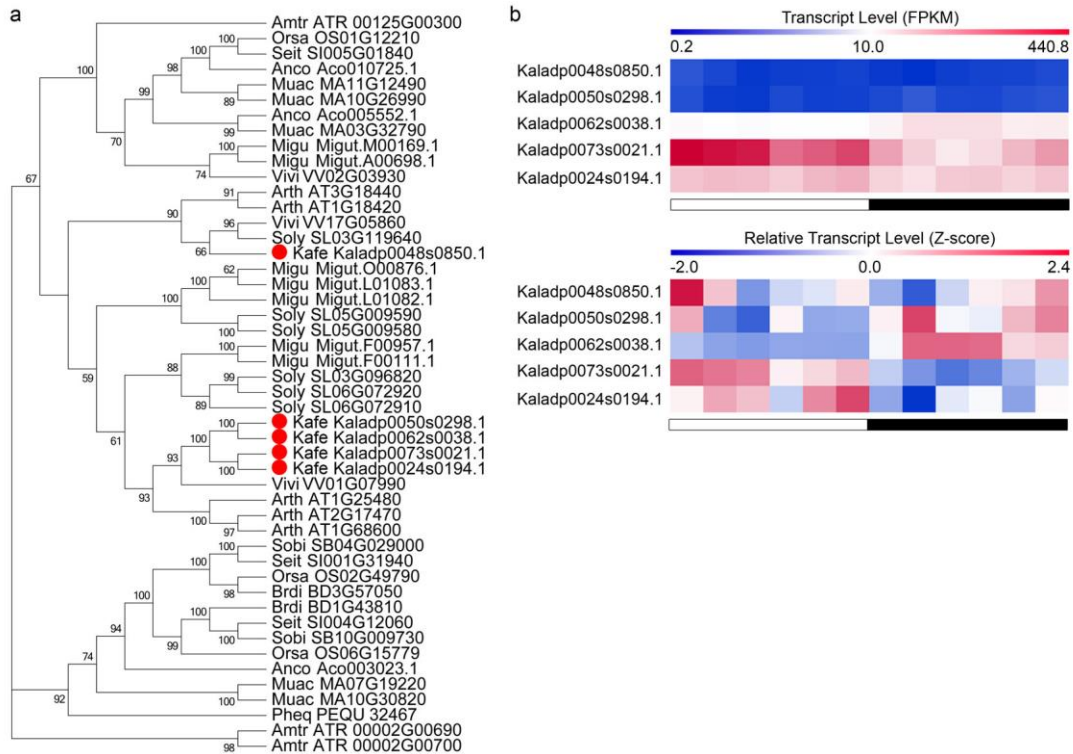
**Supplementary Figure 25. Phylogeny of the NAD(P)-malate dehydrogenase (MDH) subfamily-1 proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *MDH* subfamily-1 genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



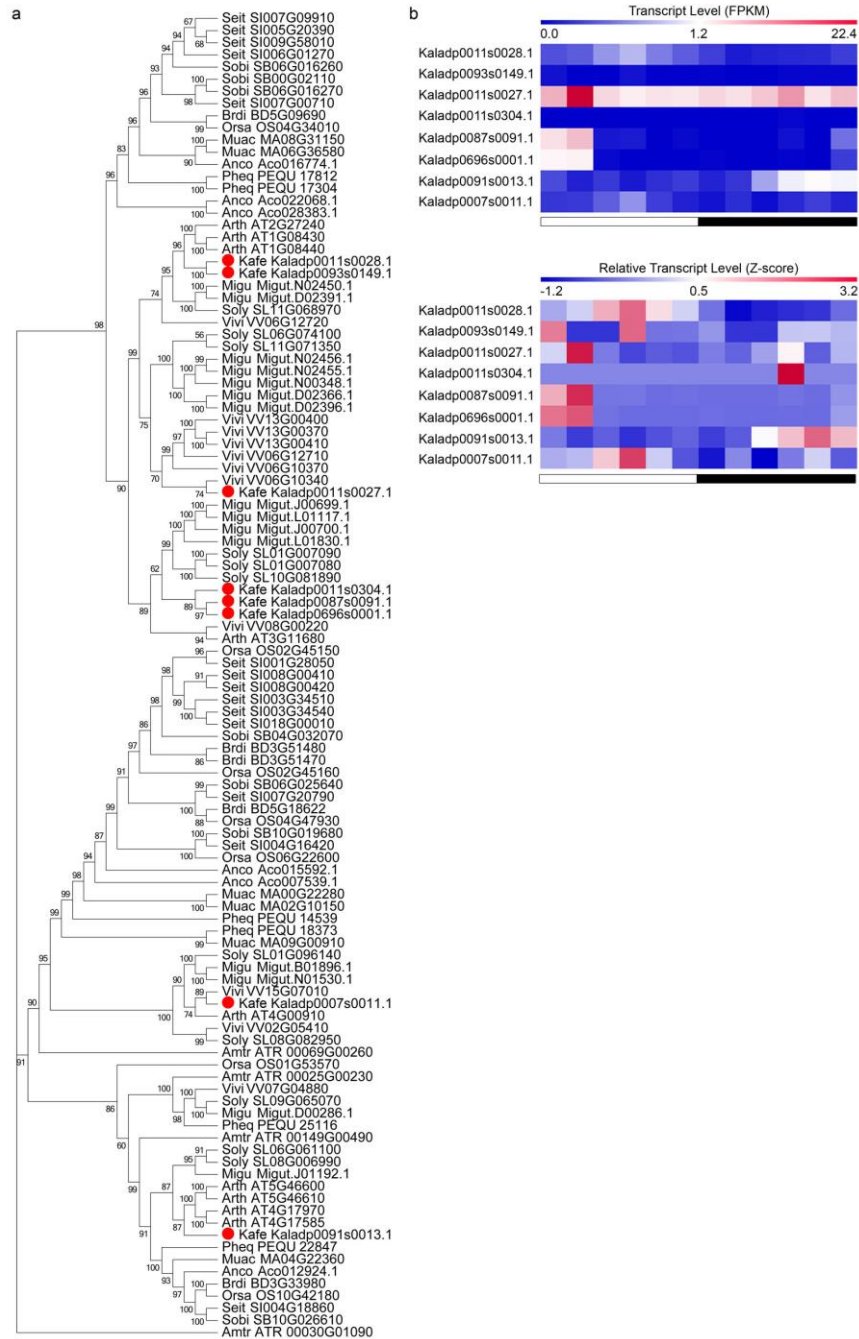
**Supplementary Figure 26. Phylogeny of the NAD(P)-malate dehydrogenase (MDH) subfamily-2 proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *MDH* subfamily-2 genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see Supplementary Table 9) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



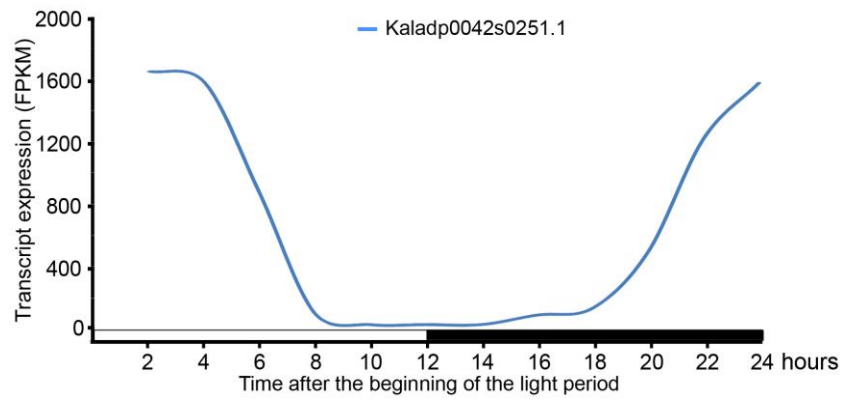
**Supplementary Figure 27. Phylogeny of the NAD(P)-malate dehydrogenase (MDH) subfamily-3 and 4 proteins and diel transcript expression in *Kalanchoë fedtschenkoi*.** (a) and (c) Maximum-likelihood phylogenetic trees constructed from protein sequences of MDH subfamily-3 and 4, respectively. (b) and (d) Diel transcript expression of the MDH subfamily-3 and 4, respectively. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see [Supplementary Table 9](#)) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



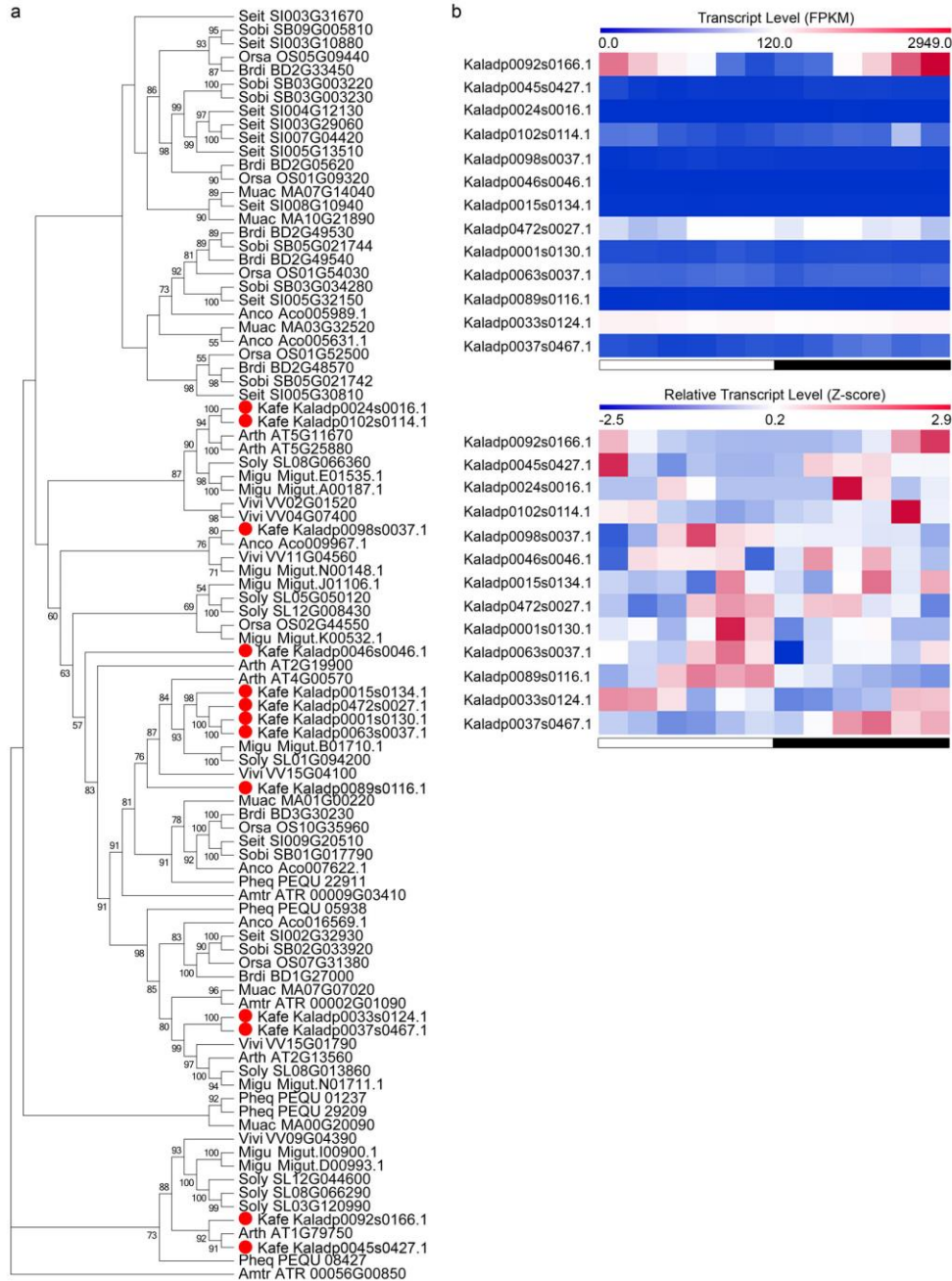
**Supplementary Figure 28. Phylogeny of the aluminum-activated malate transporter (ALMT) subfamily-1 proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *ALMT* subfamily-1 genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see [Supplementary Table 9](#)) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



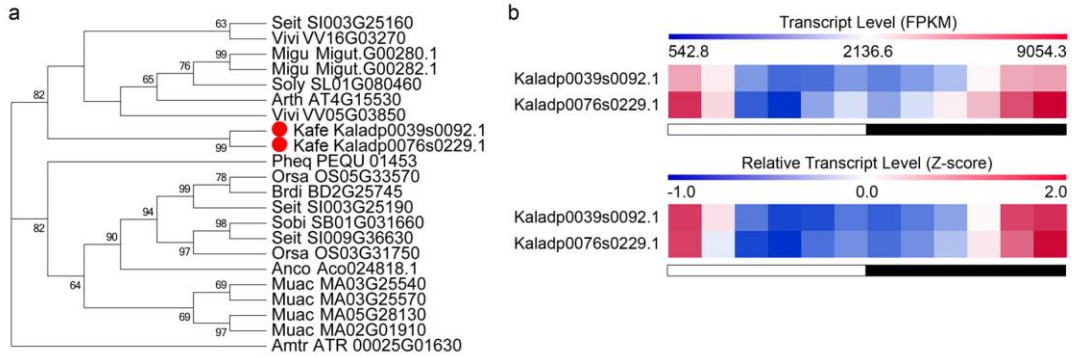
**Supplementary Figure 29. Phylogeny of the aluminum-activated malate transporter (ALMT) subfamily-2 proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *ALMT* subfamily-2 genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see [Supplementary Table 9](#)) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



**Supplementary Figure 30. Diel transcript expression of the tonoplast dicarboxylate transporter (*TDT*) gene in *Kalanchoë fedtschenkoi*.** White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.

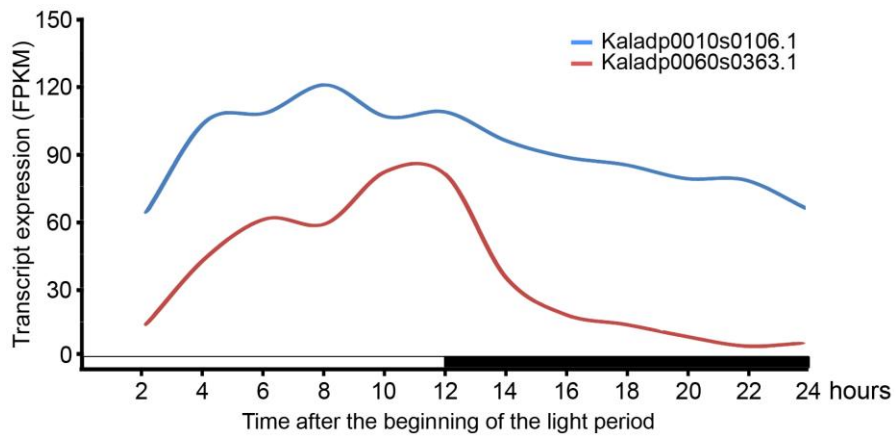


**Supplementary Figure 31. Phylogeny of the NAD(P)-malic enzyme (ME) proteins and diel transcript expression. (a)** A maximum-likelihood phylogenetic tree constructed from protein sequences. **(b)** Diel transcript expression of the *NAD(P)-ME* genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see [Supplementary Table 9](#)) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.

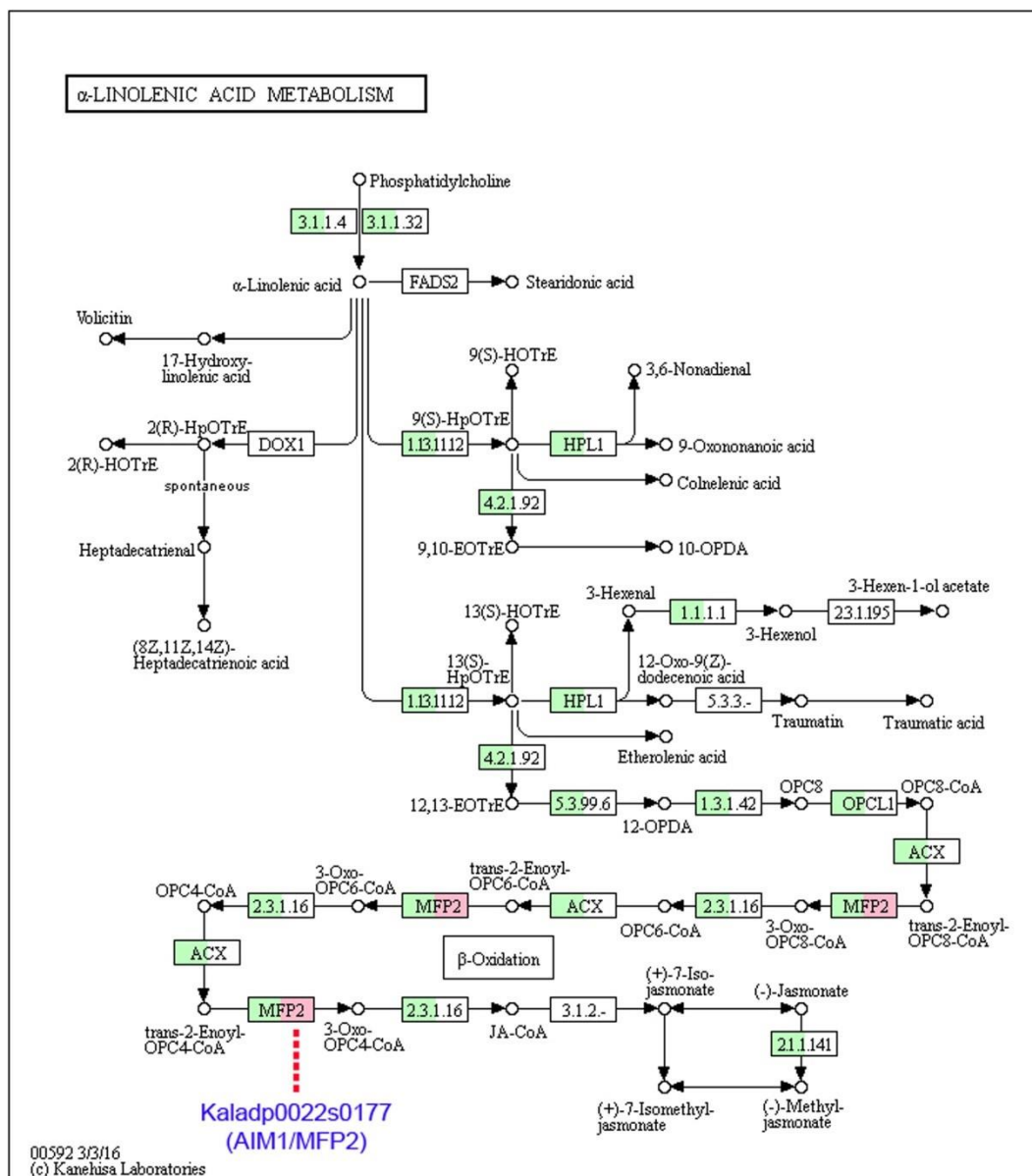


**Supplementary Figure 32. Phylogeny of the pyruvate phosphate dikinase (PPDK) proteins and diel transcript expression.** (a) A maximum-likelihood phylogenetic tree constructed from protein sequences. (b) Diel transcript expression of the *PPDK* genes in *Kalanchoë fedtschenkoi*. The taxon names in the phylogenetic tree are listed as species abbreviation (the first four letters, see [Supplementary Table 9](#)) followed by gene/transcript name. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.

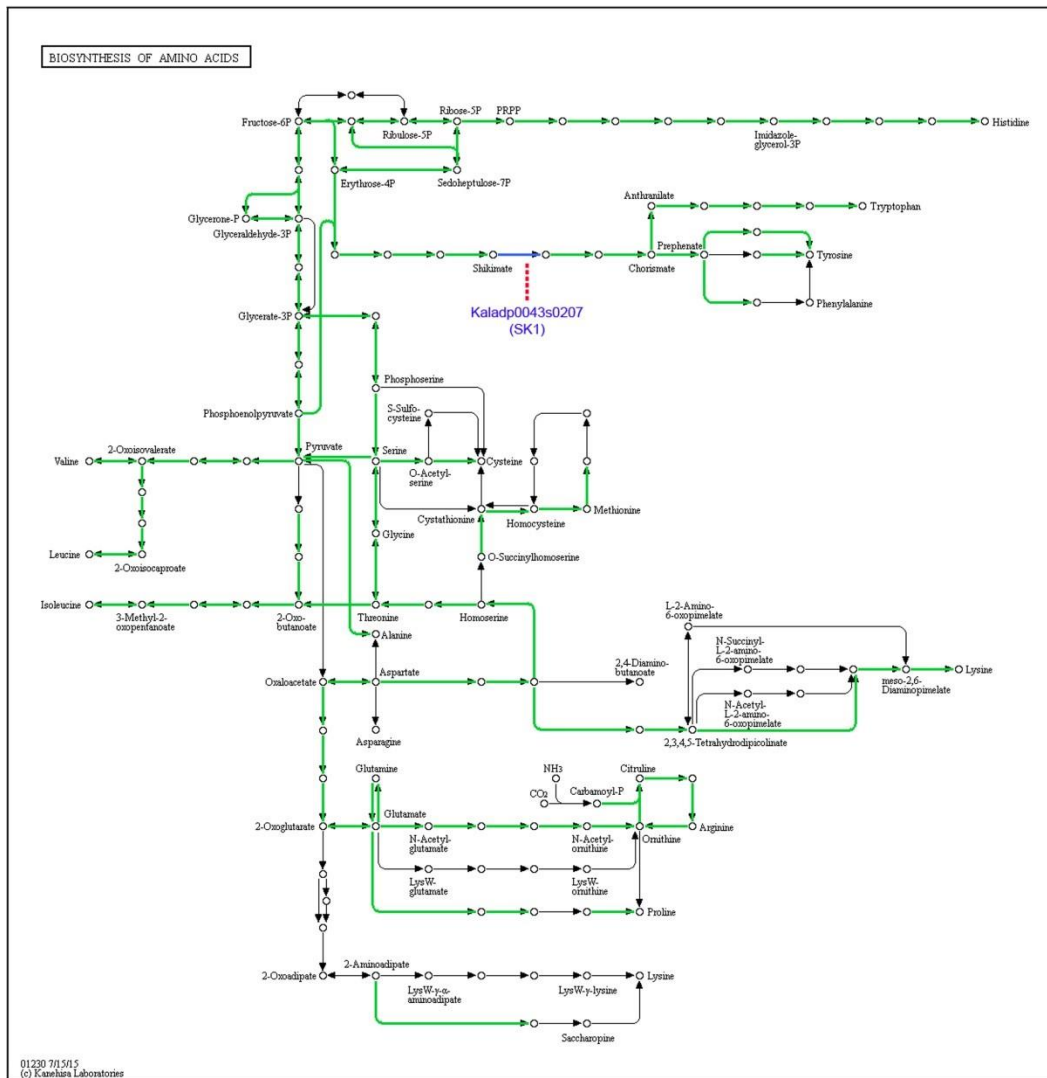




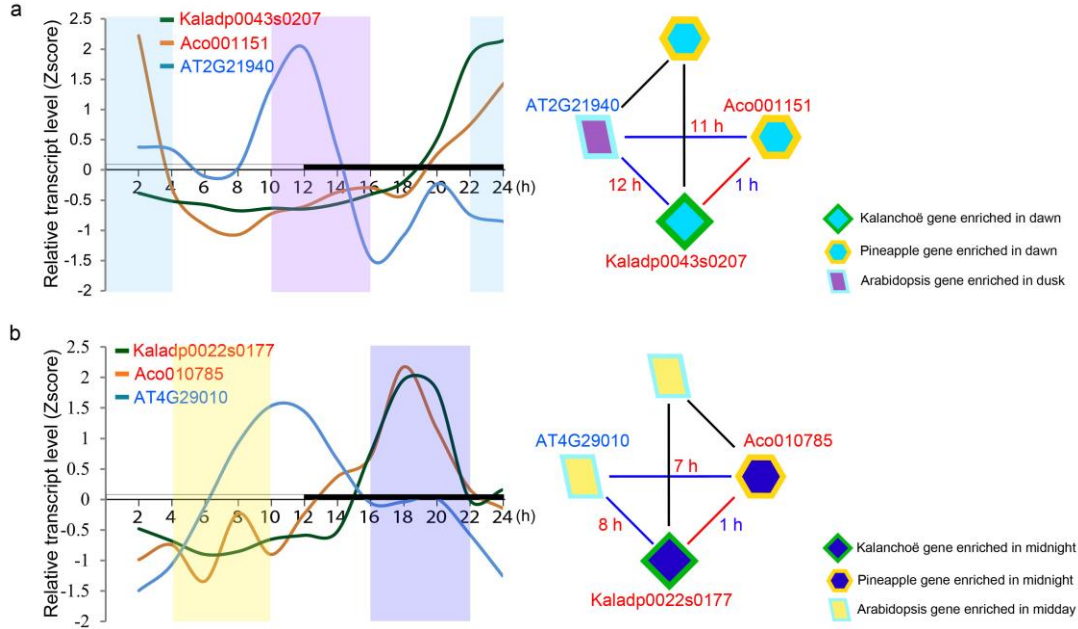
**Supplementary Figure 33. Diel transcript expression of pyruvate, phosphate dikinase regulatory protein (*PPDK-RP*) genes in *Kalanchoë fedtschenkoi*.** White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively.



**Supplementary Figure 34. An overview KEGG map of jasmonic acid biosynthesis.** Green color indicates the genes identified in *Kalanchoë fedtschenkoi*. Pink color highlights the gene (Kaladp0022s0177) showing convergent evolution with *Ananas comosus* (pineapple) in diel expression pattern. AIM1: peroxisomal fatty acid beta-oxidation multifunctional protein AIM1.



**Supplementary Figure 35. An overview KEGG map of aromatic amino acid biosynthesis.** Green color indicates the genes identified in *Kalanchoë fedtschenkoi*. Blue color highlights the *K. fedtschenkoi* gene (Kaladp0043s0207) showing convergent evolution with *Ananas comosus* (pineapple) in diel expression pattern. SK1: shikimate kinase 1.



**Supplementary Figure 36. Enrichment of triangle networks of genes relevant to secondary metabolism in *Kalanchoë fedtschenkoi*, pineapple and *Arabidopsis*.** (a) The *K. fedtschenkoi* gene (Kaladp0043s0207) encodes a shikimate kinase (SS) to regulate aromatic amino acid biosynthesis. (b) The *K. fedtschenkoi* gene (Kaladp0022s0177) encodes AIM1/MFP2 that is involved in the biosynthesis of jasmonic acid. White and black bars indicate daytime (12-hour) and nighttime (12-hour), respectively. X-axis represents the time after the beginning of the light period.

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

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

	Midday	Midnight
Gene $i$	$m_{i1}$	$m_{i2}$
$\neg$ Gene $i$	$\sum_{j \neq i} m_{j1}$	$\sum_{j \neq i} m_{j2}$

	Midnight	Midday
Gene $i$	$m_{i2}$	$m_{i1}$
$\neg$ Gene $i$	$\sum_{j \neq i} m_{j2}$	$\sum_{j \neq i} m_{j1}$

**Supplementary Figure 37. Contingency tables for Fisher Exact Test to determine the enrichment of gene expression between contrasting time windows.** Midday: 4, 6, and 8 h from the starting of the light period; midnight: 16, 18, and 20 h from the starting of the light period; dawn: 22, 24, and 2 h from the starting of the light period; and dusk: 10, 12, and 14 h from the starting of the light period.

## Supplementary Tables

**Supplementary Table 1. Estimated DNA content in *Kalanchoë fedtschenkoi*.**

Sample	DNA ratio with internal standard*	DNA pg/2C	DNA (Mbp/C)
KLM-1	0.342	0.52	253
KLM-2	0.351	0.53	259
KLM-3	0.345	0.52	255
<b>Average</b>	<b>0.346</b>	<b>0.52</b>	<b>256</b>

\* Internal standard: *Vinca minor* (DNA= 1.51 pg/2C = 1477 Mbp/2C).

**Supplementary Table 2. Sequencing reads generated for *Kalanchoë fedtschenkoi* using by Illumina MiSeq platform.**

<b>Library type</b>	<b>Fragment length</b>	<b>Reads</b>	<b>Coverage, X</b>
Paired-end	540 bp	2x 300 bp	70
Mate-pair	3 Kb	2x 300 bp	14
Mate-pair	6 Kb	2x 300 bp	12
Mate-pair	11 Kb	2x 300 bp	11

**Supplementary Table 3. Statistics of genome assembly (Scaffolds) for *Kalanchoë fedtschenkoi*.**

---

Total number of sequences	1,324
Total bases	256,351,415
Min sequence length, bp	1,002
Max sequence length, bp	10,491,024
Average sequence length, bp	193,619
Median sequence length, bp	8,120
N25 length, bp	5,483,823
N50 length, bp	2,451,343
N75 length, bp	1,051,141
N90 length, bp	258,600
N95 length, bp	77,720
As	29.83%
Ts	29.86%
Gs	17.23%
Cs	17.23%
(A + T)s	59.69%
(G + C)s	34.47%
Ns	5.85%

---



**Supplementary Table 4. Statistics of genome annotation for *Kalanchoë fedtschenkoi*.**

Primary transcripts (loci)	30,964
Alternative transcripts	14,226
Total transcripts	45,190
<hr/>	
<b>For primary transcripts:</b>	
Average number of exons	5.1
Median exon length	161
Median intron length	126
<hr/>	
<b>Gene model support:</b>	
Any EST support	25,740
EST support over 100% of their lengths	22,751
EST support over 95% of their lengths	23,476
EST support over 90% of their lengths	23,775
EST support over 75% of their lengths	24,330
EST support over 50% of their lengths	24,870
Peptide homology coverage of 100%	1,729
Peptide homology coverage of over 95%	16,118
Peptide homology coverage of over 90%	19,939
Peptide homology coverage of over 75%	24,305
Peptide homology coverage of over 50%	27,502
Pfam annotation	21,794
Panther annotation	27,197
KOG annotation	12,093
KEGG Orthology annotation	4,656
E.C. number annotation	2,452

**Supplementary Table 5. List of genes involved in CAM carboxylation process in *Kalanchoë fedtschenkoi*.**

Tribe_id	Name	Gene locus	Definition	Gene expression (FPKM) <sup>a</sup>	Gene Co-expression module <sup>b</sup>	Cluster <sup>c</sup>	Profile-variation <sup>d</sup>
150_F000796	β-CA	Kaladp0018s0287.1	Beta carbonic anhydrase	219.37	MEgreen	3	non-flat
150_F000796	β-CA	Kaladp0018s0289.1	Beta carbonic anhydrase	352.68	MEblack	3	non-flat
150_F000796	β-CA	Kaladp0024s0122.1	Beta carbonic anhydrase	18.35	MEpink	8	non-flat
150_F000796	β-CA	Kaladp0034s0051.1	Beta carbonic anhydrase	277.08	MEsalmon	N/A	non-flat
150_F000796	β-CA	Kaladp0081s0140.1	Beta carbonic anhydrase	1.14	MEagenta	10	non-flat
150_F000796	β-CA	Kaladp0081s0143.1	Beta carbonic anhydrase	2.39	MElightcyan	N/A	flat
150_F000796	β-CA	Kaladp0538s0011.1	Beta carbonic anhydrase	66.8	MEtan	1	flat
150_F000022	PPCK1	Kaladp0037s0517.1	Phosphoenolpyruvate carboxylase kinase	1187.84	MEblack	3	non-flat
150_F000022	PPCK	Kaladp0050s0014.1	Phosphoenolpyruvate carboxylase kinase	37.58	MElightyellow	N/A	flat
150_F000022	PPCK	Kaladp0082s0192.1	Phosphoenolpyruvate carboxylase kinase	6.62	MEpurple	7	non-flat
150_F000022	PPCK2	Kaladp0604s0001.1	Phosphoenolpyruvate carboxylase kinase	376.3	MElightgreen	7	non-flat
150_F000807	PEPC	Kaladp0011s0355.1	Phosphoenolpyruvate carboxylase	10.79	MEtan	6	non-flat
150_F000807	PEPC	Kaladp0011s1355.1	Phosphoenolpyruvate carboxylase	3.29	MEgrey	N/A	flat
150_F000807	PEPC2	Kaladp0048s0578.1	Phosphoenolpyruvate carboxylase 2	280.38	MEblack	3	non-flat
150_F000807	PEPC	Kaladp0062s0055.1	Phosphoenolpyruvate carboxylase	18.51	MEgrey	2	non-flat
150_F000807	PEPC1	Kaladp0095s0055.1	Phosphoenolpyruvate carboxylase 1	8338.11	MEpink	7	non-flat
150_F000614	MDH	Kaladp0001s0257.1	Malate dehydrogenase	383.29	MEpink	11	non-flat
150_F000614	MDH	Kaladp0022s0111.1	Malate dehydrogenase	43.82	MEblack	3	non-flat
150_F000614	MDH	Kaladp0048s0189.1	Malate dehydrogenase	4.27	MEgrey	N/A	flat
150_F000614	MDH	Kaladp0058s0569.1	Malate dehydrogenase	150.19	MEblue	2	non-flat
150_F000614	MDH	Kaladp0095s0052.1	Malate dehydrogenase	107.8	MEbrown	1	non-flat
150_F000614	MDH	Kaladp0095s0564.1	Malate dehydrogenase	1.17	MEgreenyellow	N/A	flat
150_F000614	MDH	Kaladp0101s0211.1	Malate dehydrogenase	236.21	MEtan	1	non-flat
150_F001882	MDH	Kaladp0082s0194.1	Malate dehydrogenase	1925.18	MEblue	10	non-flat
150_F001882	MDH	Kaladp0093s0088.1	Malate dehydrogenase	3.08	MEturquoise	N/A	flat
150_F001882	MDH	Kaladp1038s0012.1	Malate dehydrogenase	19.58	MEturquoise	N/A	flat
150_F002831	MDH	Kaladp0068s0169.1	Malate dehydrogenase	226.83	MEturquoise	5	non-flat
150_F004944	MDH	Kaladp0101s0012.1	Malate dehydrogenase	60.4	MEsalmon	N/A	flat
150_F001112	ALMT	Kaladp0024s0194.1	Tonoplast aluminum-activated malate transporter	147.01	MEpink	N/A	flat
150_F001112	ALMT	Kaladp0048s0850.1	Tonoplast aluminum-activated malate transporter	8.52	MEturquoise	N/A	flat
150_F001112	ALMT	Kaladp0050s0298.1	Tonoplast aluminum-activated malate transporter	2.49	MEtan	N/A	flat
150_F001112	ALMT6	Kaladp0062s0038.1	Tonoplast aluminum-activated malate transporter	66.11	MEblack	3	non-flat
150_F001112	ALMT6	Kaladp0073s0021.1	Tonoplast aluminum-activated malate transporter 6	440.84	MEpink	2	non-flat
150_F000383	ALMT	Kaladp0007s0011.1	Tonoplast aluminum-activated malate transporter	1.11	MEbrown	10	non-flat
150_F000383	ALMT	Kaladp0011s0027.1	Tonoplast aluminum-activated malate transporter	123.96	MEturquoise	2	non-flat
150_F000383	ALMT	Kaladp0011s0028.1	Tonoplast aluminum-activated malate transporter	2.97	MEgrey	N/A	flat
150_F000383	ALMT	Kaladp0087s0091.1	Tonoplast aluminum-activated malate transporter	6.68	MEdarkgreen	2	non-flat
150_F000383	ALMT	Kaladp0091s0013.1	Tonoplast aluminum-activated malate transporter	1.54	MEsalmon	4	non-flat
150_F000383	ALMT	Kaladp0093s0149.1	Tonoplast aluminum-activated malate transporter	17.78	MEturquoise	N/A	flat

<sup>a</sup>The maximum expression level in the mature leaf during 24-h period, as revealed by RNA-seq analysis.

<sup>b</sup>The gene co-expression module listed in [Supplementary Fig. 9](#).

<sup>c</sup>Gene clusters constructed, using maSigPro<sup>5,6</sup>, from the transcripts that showed significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) “non-flat” diel expression patterns.

<sup>d</sup>The significance of variation of transcript abundance over a 24-h period was determined by a polynomial regression. “Non-flat” indicates that the transcript abundance varied significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) across time points. “Flat” indicates that the transcript abundance did not vary significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) across time points.

**Supplementary Table 6. List of genes involve in CAM decarboxylation process in *Kalanchoë fedtschenkoi*.**

Tribe_id	Name	Gene locus	Definition	Gene expression (FPKM) <sup>a</sup>	Gene Co-expression module <sup>b</sup>	Cluster <sup>c</sup>	Profile-variation <sup>d</sup>
150_F004828	TDT	Kaladp0042s0251.1	Tonoplast dicarboxylate transporter	1617.63	MEsalmon	1	non-flat
150_F003433	PPDK	Kaladp0039s0092.1	Pyruvate, orthophosphate dikinase	4744.54	MEpink	1	non-flat
150_F003433	PPDK	Kaladp0076s0229.1	Pyruvate, orthophosphate dikinase	9054.27	MEsalmon	1	non-flat
150_F005114	PPDK-RP	Kaladp0010s0106.1	PPDK regulatory protein	121.13	MEblue	10	non-flat
150_F005114	PPDK-RP	Kaladp0060s0363.1	PPDK regulatory protein	81.78	MEblue	11	non-flat
150_F001811	PEPCK	Kaladp0023s0088.1	Phosphoenolpyruvate carboxykinase	17.45	MEgrey	N/A	flat
150_F001811	PEPCK	Kaladp0040s0194.1	Phosphoenolpyruvate carboxykinase	78.13	MEturquoise	4	non-flat
150_F001811	PEPCK	Kaladp1116s0004.1	Phosphoenolpyruvate carboxykinase	9.04	MEgrey	3	non-flat
150_F000472	NAD-ME	Kaladp0001s0130.1	NAD-dependent malic enzyme	21.76	MEcyan	N/A	flat
150_F000472	NAD-ME	Kaladp0015s0134.1	NAD-dependent malic enzyme	20.74	MEgrey	N/A	flat
150_F000472	NADP-ME	Kaladp0024s0016.1	NADP-malic enzyme	1.01	MEgrey	N/A	flat
150_F000472	NAD-ME	Kaladp0033s0124.1	NAD-dependent malic enzyme	269.31	MEpink	1	non-flat
150_F000472	NAD-ME	Kaladp0037s0467.1	NAD-dependent malic enzyme	83.61	MEgreen	3	non-flat
150_F000472	NADP-ME	Kaladp0045s0427.1	NADP-malic enzyme	204.39	MEturquoise	1	non-flat
150_F000472	NAD-ME	Kaladp0063s0037.1	NAD-dependent malic enzyme	36.49	MEcyan	1	non-flat
150_F000472	NADP-ME	Kaladp0092s0166.1	NADP-malic enzyme	2948.96	MEsalmon	4	non-flat
150_F000472	NAD-ME	Kaladp0472s0027.1	NAD-dependent malic enzyme	129.8	MEmidnightblue	N/A	flat

<sup>a</sup>The maximum expression level in the mature leaf during 24-h period, as revealed by RNA-seq analysis.

<sup>b</sup>The gene co-expression module listed in [Supplementary Fig. 9](#).

<sup>c</sup>Gene clusters constructed, using maSigPro<sup>5,6</sup>, from the transcripts that showed significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) “non-flat” diel expression patterns.

<sup>d</sup>The significance of variation of transcript abundance over a 24-h period was determined by a polynomial regression. “Non-flat” indicates that the transcript abundance varied significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) across time points. “Flat” indicates that the transcript abundance did not vary significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $P < 0.05$ ) across time points.

**Supplementary Table 7: Number of genes shared between *Kalanchoë fedtschenkoi* co-expression modules (Supplementary Fig. 9) and genes clusters (Supplementary Fig. 11).**

The gene clusters were constructed, using maSigPro<sup>5,6</sup>, from the transcripts that showed significantly (ANOVA of glm models where  $H_0 =$  a flat line,  $p$ -value<0.05) “non-flat” diel expression patterns. The numbers in parentheses are number of genes in each co-expression module or gene cluster.

Module	Cluster 1 (1208)	Cluster 2 (1274)	Cluster 3 (1751)	Cluster 4 (2146)	Cluster 5 (1381)	Cluster 6 (1061)	Cluster 7 (1797)	Cluster 8 (927)	Cluster 9 (1107)	Cluster 10 (1717)	Cluster 11 (2077)
MEblack (782)	1	0	298	43	80	8	2	0	181	1	1
MEblue (1911)	0	75	0	4	3	2	285	100	12	297	717
MEbrown(1593)	6	254	9	1	70	1	136	20	17	380	247
MEcyan (1048)	17	2	100	348	4	125	11	30	8	2	3
MEdarkgreen (437)	79	100	6	15	0	5	1	81	1	5	12
MEdarkgrey (615)	24	0	63	341	1	21	0	2	1	3	0
MEdarkred (410)	2	20	1	0	20	1	113	5	19	48	57
MEdarkturquoise (369)	176	29	5	34	0	3	0	16	0	2	0
MEgreen (1031)	12	2	555	105	337	71	7	3	171	9	0
MEgreenyellow (733)	7	1	200	6	249	18	4	0	38	8	0
MEgrey (2277)	85	108	108	110	95	43	186	37	79	74	70
MEgrey60 (506)	0	42	3	0	59	9	134	1	18	35	34
MElightcyan (731)	6	90	0	2	0	13	190	31	3	2	89
MElightgreen (504)	1	1	12	1	97	12	81	0	149	13	2
MElightyellow (641)	12	56	12	5	22	1	6	42	6	171	18
MEmagenta (638)	0	1	25	3	295	16	35	1	121	8	2
EMidnightblue (514)	2	4	8	1	77	1	73	2	76	124	9
MEpink (1353)	64	93	22	111	12	18	59	245	16	121	155
MEpurple (1509)	36	114	22	41	24	22	153	94	52	48	119
MEred (1277)	43	7	25	72	7	520	15	5	7	1	6
MEroyalblue (408)	3	1	72	86	12	2	3	26	8	36	4
MEsalmon (731)	128	0	10	403	0	7	1	15	0	0	1
MEtan (688)	196	24	39	101	6	12	3	15	3	7	2
MEturquoise (3052)	259	203	262	162	172	120	75	66	100	81	45
MEyellow (1144)	1	35	0	2	1	1	210	27	8	192	425

**Supplementary Table 8: Time-shifts in diel transcript expression profiles between orthologs in *Kalanchoë fedtschenkoi* (Kafe), *Ananas comosus* (Anco) and *Arabidopsis thaliana* (Arth).** Only the gene pairs containing *K. fedtschenkoi* and *A. comosus* genes that showed convergent changes in diel expression pattern were shown here. “Xcorr” is the cross correlation. “Time” is time-shift in hours.

Name	<i>Kalanchoë</i>	Pineapple	<i>Arabidopsis</i>	Kafe vs. Arth		Anco vs. Arth		Kafe vs. Anco	
				Time	Xcorr	Time	Xcorr	Time	Xcorr
PHOT2	Kaladp0033s0113.1	Aco014242.1	AT5G58140	11	0.58	9	0.63	1	0.87
PPR	Kaladp0058s0382.1	Aco019522.1	AT1G02060	12	0.67	11	0.72	0	0.91
DAG	Kaladp0130s0003.1	Aco002485.1	AT3G06790	11	0.82	11	0.90	0	0.92
TPX2	Kaladp0062s0086.1	Aco009655.1	AT3G23090	9	0.88	9	0.85	1	0.90
MR11	Kaladp0053s0416.1	Aco007956.1	AT2G05830	11	0.87	11	0.72	0	0.78
NUTF2	Kaladp0040s0332.1	Aco015511.1	AT2G46100	7	0.84	6	0.79	1	0.88
bcrC	Kaladp0040s0040.1	Aco007440.1	AT2G34670	8	0.69	8	0.70	0	0.91
	Kaladp0045s0260.1	Aco004226.1	AT2G38820	10	0.90	11	0.94	1	0.85
PLT1	Kaladp0040s0745.1	Aco000080.1	AT2G16120	7	0.77	8	0.50	1	0.84
EXS	Kaladp0008s0539.1	Aco017367.1	AT1G55610	9	0.79	9	0.73	1	0.88
HSP70	Kaladp0060s0296.1	Aco031458.1	AT5G02500	10	0.64	9	0.71	1	0.80
CSLC12	Kaladp0037s0421.1	Aco011603.1	AT4G07960	10	0.85	9	0.89	1	0.90
PLC2	Kaladp0059s0034.1	Aco001556.1	AT3G08510	9	0.93	8	0.81	0	0.85
Lycopene cyclase protein	Kaladp0056s0132.1	Aco024665.1	AT3G10230	9	0.83	10	0.80	0	0.85
CDK17	Kaladp0028s0063.1	Aco011817.1	AT1G49620	10	0.83	8	0.68	0	0.86
NPF	Kaladp0033s0087.1	Aco013262.1	AT2G26690	11	0.93	8	0.74	2	0.83
TIM23	Kaladp1244s0001.1	Aco007546.1	AT1G72750	6	0.92	7	0.86	1	0.83
IAA29	Kaladp0048s0752.1	Aco030721.1	AT4G32280	7	0.86	7	0.69	0	0.87
PRN12	Kaladp0026s0118.1	Aco015046.1	AT5G64760	9	0.84	12	0.66	1	0.68
PDK	Kaladp0068s0282.1	Aco017179.1	AT3G06483	10	0.91	11	0.88	1	0.76
OMP	Kaladp0040s0029.1	Aco018518.1	AT3G52420	9	0.76	12	0.86	2	0.82
TMEM245	Kaladp0048s0390.1	Aco003087.1	AT5G55960	12	0.81	12	0.73	1	0.81
TMEM184C	Kaladp1262s0005.1	Aco014396.1	AT1G11200	9	0.64	7	0.76	0	0.66
ACLB2	Kaladp0045s0074.1	Aco001770.1	AT3G06650	8	0.77	6	0.75	1	0.85
NPH3 family	Kaladp0040s0264.1	Aco016939.1	AT5G67440	10	0.95	9	0.77	0	0.78
SS	Kaladp0043s0207.1	Aco001151.1	AT2G21940	12	0.83	11	0.87	1	0.90
Ash21	Kaladp0045s0419.1	Aco000634.1	AT1G51450	11	0.78	11	0.95	1	0.88
WRC	Kaladp0011s0273.1	Aco020634.1	AT2G42040	8	0.58	8	0.74	1	0.90
SNRPC1	Kaladp0349s0001.1	Aco016127.1	AT4G03120	8	0.87	11	0.88	2	0.89
COL2	Kaladp0039s0496.1	Aco002998.1	AT1G68520	11	0.73	11	0.94	1	0.85
GK	Kaladp0037s0359.1	Aco011530.1	AT1G80460	9	0.81	9	0.87	0	0.92
FH	Kaladp0040s0015.1	Aco002502.1	AT5G50950	8	0.93	9	0.87	1	0.87
PIP1-2	Kaladp0059s0048.1	Aco010251.1	AT4G23400	9	0.80	8	0.74	0	0.85
AIM1	Kaladp0022s0177.1	Aco010785.1	AT4G29010	8	0.86	7	0.82	1	0.92
HAD	Kaladp0071s0244.1	Aco001521.1	AT3G45740	11	0.75	12	0.76	0	0.82
SURE	Kaladp0427s0023.1	Aco023553.1	AT4G14930	7	0.52	8	0.74	0	0.86
EDR2L	Kaladp0098s0203.1	Aco027425.1	AT5G10750	9	0.87	12	0.74	2	0.83
CYB2	Kaladp0035s0007.1	Aco006777.1	AT5G38630	10	0.55	12	0.66	1	0.93
SEC14	Kaladp0067s0305.1	Aco003265.1	AT1G72160	11	0.73	11	0.64	1	0.88
Gdap2	Kaladp0098s0113.1	Aco002356.1	AT3G10210	9	0.56	9	0.62	1	0.84
FAH	Kaladp0060s0413.1	Aco015620.1	AT1G12050	9	0.60	10	0.47	0	0.82
	Kaladp0081s0088.1	Aco014627.1	AT1G04650	10	0.52	9	0.79	0	0.71
UFC1	Kaladp0095s0421.1	Aco014085.1	AT1G27530	9	0.89	9	0.85	1	0.77
ABCG15	Kaladp0045s0418.1	Aco007271.1	AT1G17840	11	0.80	10	0.60	1	0.81
	Kaladp0034s0049.1	Aco005409.1	AT1G23760	8	0.95	7	0.88	1	0.90
SAG	Kaladp0808s0026.1	Aco022311.1	AT2G44670	11	0.89	10	0.85	1	0.97
PPCK1	Kaladp0037s0517.1	Aco013938.1	AT1G08650	11	0.84	12	0.94	2	0.81
Starch synthase	Kaladp0055s0317.1	Aco010848.1	AT5G24300	9	0.76	7	0.73	3	0.85
SNRPC	Kaladp0018s0298.1	Aco007284.1	AT1G09070	10	0.88	10	0.79	0	0.89
Leucine rich repeat	Kaladp1251s0003.1	Aco008844.1	AT3G47090	7	0.90	7	0.74	2	0.80
CHR5	Kaladp0011s0810.1	Aco001980.1	AT2G13370	6	0.78	6	0.53	0	0.86
Leucine rich repeat	Kaladp0054s0031.1	Aco008313.1	AT2G07040	8	0.78	9	0.51	0	0.72
CINV1	Kaladp0550s0020.1	Aco007782.1	AT5G22510	11	0.83	11	0.88	1	0.79
TPS	Kaladp0011s0363.1	Aco012107.1	AT1G06410	6	0.70	6	0.56	0	0.84

**Supplementary Table 9. The representative species used for identification of convergent change in protein sequences.**

<b>Photosynthesis</b>	<b>Species name</b>	<b>Abbreviation</b>
C <sub>3</sub>	<i>Amborella trichopoda</i>	Amtr
C <sub>3</sub>	<i>Arabidopsis thaliana</i>	Arth
C <sub>3</sub>	<i>Brachypodium distachyon</i>	Brdi
C <sub>3</sub>	<i>Mimulus guttatus</i>	Migu
C <sub>3</sub>	<i>Musa acuminata</i>	Muac
C <sub>3</sub>	<i>Oryza sativa</i>	Orsa
C <sub>3</sub>	<i>Solanum lycopersicum</i>	Soly
C <sub>3</sub>	<i>Vitis vinifera</i>	Vivi
C <sub>4</sub>	<i>Setaria italica</i>	Seit
C <sub>4</sub>	<i>Sorghum bicolor</i>	Sobi
CAM	<i>Ananas comosus</i>	Anco
CAM	<i>Kalanchoë fedtschenkoi</i>	Kafe
CAM	<i>Phalaenopsis equestris</i>	Pheq

**Supplementary Table 10. List of *Kalanchoë fedtschenkoi* genes showing convergent changes in protein sequences.**

Tribe_id	Name	Gene locus	Definition	Gene expression (FPKM) <sup>a</sup>	Gene Co-expression module <sup>b</sup>	Cluster <sup>c</sup>	Profile-variation <sup>d</sup>
150_F000807	PEPC2	Kaladp0048s0578.1	Phosphoenolpyruvate carboxylase 2	280.38	MEblack	3	non-flat
150_F001102	NAP1L4	Kaladp0094s0051.1	Nucleosome assembly protein 14-like	3.4	MEturquoise	NA	flat
150_F001653	HY5	Kaladp0060s0460.1	Transcription factor hy5-like protein	59.76	MEblack	NA	flat
150_F003508	GPI	Kaladp0095s0394.1	Glucose-6-phosphate isomerase chloroplastic-like	38.42	MEturquoise	1	non-flat

<sup>a</sup>The maximum expression level in the mature leaf during 24-h period, as revealed by RNA-seq analysis.

<sup>b</sup>The gene co-expression module listed in [Supplementary Fig. 9](#).

<sup>c</sup>Gene clusters constructed, using maSigPro<sup>5,6</sup>, from the transcripts that showed significantly (ANOVA of glm models where  $H_0$  = a flat line,  $P < 0.05$ ) “non-flat” diel expression patterns.

<sup>d</sup>The significance of variation of transcript abundance over a 24-h period was determined by a polynomial regression. “Non-flat” indicates that the transcript abundance varied significantly (ANOVA of glm models where  $H_0$  = a flat line,  $P < 0.05$ ) across time points. “Flat” indicates that the transcript abundance did not vary significantly (ANOVA of glm models where  $H_0$  = a flat line,  $P < 0.05$ ) across time points.

**Supplementary Table 11. A list of *Kalanchoë fedtschenkoi* genes involved in stomatal movement.** The role of *K. fedtschenkoi* genes was predicted based on homology to the *Arabidopsis* genes that have been shown to be involved in regulation of stomatal movement.

Name	<i>Kalanchoë</i>	Description	<i>Arabidopsis</i>	References
ABI1	Kaladp0011s0443.1	ABA INSENSITIVE 1	AT4G26080	9
ABI2	Kaladp0048s0509.1	ABA INSENSITIVE 2	AT5G57050	9
AHA2	Kaladp0008s0304.1	PLASMA MEMBRANE PROTON ATPASE 2	AT4G30190	10
AKT1	Kaladp0055s0506.1	POTASSIUM TRANSPORTER 1	AT2G26650	11
ALMT9	Kaladp0062s0038.1	ALUMINUM-ACTIVATED MALATE TRANSPORTER 9	AT3G18440	12
BAK1	Kaladp0043s0196.1	BRI1-ASSOCIATED RECEPTOR KINASE	AT4G33430	13
BLUS1	Kaladp0062s0090.1	BLUE LIGHT SIGNALING1	AT4G14480	14
CPK23	Kaladp0040s0351.1	CALCIUM-DEPENDENT PROTEIN KINASE 23	AT4G04740	15
CPK3	Kaladp0042s0341.1	CALCIUM-DEPENDENT PROTEIN KINASE 3	AT4G23650	16
CPK6	Kaladp0055s0096.1	CALCIUM-DEPENDENT PROTEIN KINASE 6	AT2G17290	16
HT1	Kaladp0073s0100.1	HIGH LEAF TEMPERATURE 1	AT1G62400	17
KAT1	Kaladp0008s0789.1	POTASSIUM CHANNEL IN ARABIDOPSIS THALIANA 1	AT5G46240	18
KAT2	Kaladp0840s0007.1	POTASSIUM CHANNEL IN ARABIDOPSIS THALIANA 2	AT4G18290	19
OST1	Kaladp0016s0289.1	OPEN STOMATA 1	AT4G33950	20
OST2/AH A1	Kaladp0098s0188.1	OPEN STOMATA 2	AT2G18960	21
PHOT1	Kaladp0071s0248.2	PHOTOTROPIN 1	AT3G45780	22
PHOT2	Kaladp0033s0113.1	PHOTOTROPIN 2	AT5G58140	22
PYL9	Kaladp0008s0082.1	PYRABACTIN RESISTANCE 1-LIKE 9	AT1G01360	23
QUAC1/A LMT12	Kaladp0091s0013.1	QUICK-ACTIVATING ANION CHANNEL 1	AT4G17970	24
SLAC1	Kaladp0050s0214.1	SLOW ANION CHANNEL-ASSOCIATED 1	AT1G12480	25



**Supplementary Table 12. A list of *Kalanchoë fedtschenkoi* genes involved in circadian rhythm.** The role of *K. fedtschenkoi* genes was predicted based on homology to the *Arabidopsis* genes that have been shown to be involved in circadian rhythm.

Name	<i>Kalanchoë</i>	Description	<i>Arabidopsis</i>	References
COP1	Kaladp0011s0927.1	CONSTITUTIVE PHOTOMORPHOGENIC 1	AT2G32950	25
CRY1	Kaladp0071s0308.1	CRYPTOCHROME 1	AT4G08920	26
CRY2	Kaladp0082s0193.1	CRYPTOCHROME 2	AT1G04400	27
ELF3	Kaladp0039s0732.1	EARLY FLOWERING 3	AT2G25930	28
FKF1	Kaladp0036s0214.1	FLAVIN-BINDING, KELCH REPEAT, F BOX 1	AT1G68050	29
PHYA	Kaladp0034s0172.1	PHYTOCHROME A	AT1G09570	30
PHYB	Kaladp0039s0298.1	PHYTOCHROME B	AT2G18790	31
CCA1	Kaladp0496s0018.2	CIRCADIAN CLOCK ASSOCIATED 1	AT2G46830	32
CHE	Kaladp0032s0054.1	CCA1 HIKING EXPEDITION	AT5G08330	33
GI	Kaladp0040s0489.1	GIGANTEA	AT1G22770	34
LUX	Kaladp0033s0047.1	LUX ARRHYTHMO	AT3G46640	35
PIF3	Kaladp0057s0097.1	PHYTOCHROME INTERACTING FACTOR 3	AT1G09530	36
PRR3	Kaladp0058s0661.1	PSEUDO-RESPONSE REGULATOR 3	AT5G60100	37
PRR5	Kaladp0032s0115.1	PSEUDO-RESPONSE REGULATOR 5	AT5G24470	38
PRR7	Kaladp0101s0041.1	PSEUDO-RESPONSE REGULATOR 7	AT5G02810	38
PRR9	Kaladp0032s0115.1	PSEUDO-RESPONSE REGULATOR 9	AT2G46790	38
TOC1	Kaladp0040s0446.2	TIMING OF CAB EXPRESSION 1	AT5G61380	39
ZTL	Kaladp0809s0098.1	ZEITLUPE	AT5G57360	40
LNK1	Kaladp0607s0046.1	NIGHT LIGHT-INDUCIBLE AND CLOCK-REGULATED 1	AT5G64170	41
LNK2	Kaladp0099s0129.1	NIGHT LIGHT-INDUCIBLE AND CLOCK-REGULATED 2	AT3G54500	41
RVE1	Kaladp0574s0015.1	REVEILLE 1	AT5G17300	42
RVE6	Kaladp0055s0349.1	REVEILLE 6	AT5G52660	43
RVE8	Kaladp0577s0020.1	REVEILLE 8	AT3G09600	44
CKB4	Kaladp0016s0180.1	CASEIN KINASE II BETA SUBUNIT 4	AT2G44680	45
ELF4	Kaladp0045s0206.1	EARLY FLOWERING 4	AT2G40080	35
FIO1	Kaladp0089s0025.1	FIONA1	AT2G21070	46
HY5	Kaladp0060s0460.1	ELONGATED HYPOCOTYL 5	AT5G11260	47
JMJD5	Kaladp0076s0198.1	LIGHT INSENSITIVE PERIOD1	AT5G64813	48
LWD1	Kaladp0048s0797.1	LIGHT-REGULATED WD 1	AT1G12910	49
PRMT5	Kaladp0056s0075.1	PROTEIN ARGININE METHYLTRANSFERASE 5	AT4G31120	50
SKIP	Kaladp0040s0680.1	SNW/SKI-INTERACTING PROTEIN	AT1G77180	51
STIPL1	Kaladp0071s0383.1	SPLICEOSOMAL TIMEKEEPER LOCUS1	AT1G17070	52
TEJ	Kaladp0040s0530.1	POLY(ADP-RIBOSE) GLYCOHYDROLASE 1	AT2G31870	53

## Supplementary Data Files

**Supplementary data 1.** Biological processes enriched in co-expression modules. See separate Excel file.

**Supplementary data 2.** The hub genes with the top 1% most connected nodes as well as putative CAM genes with at least 10 directed edges from each gene cluster shown in Supplementary Table 7 and Supplementary Fig. 11. See separate Excel file

**Supplementary data 3.** List of *Kalanchoë* genes showing convergent changes in diel expression pattern. See separate Excel file

**Supplementary data 4.** Gene count of CAZyme module genes in thirteen plant species. See separate Excel file.

**Supplementary data 5.** List of predicted CAZyme genes in *Kalanchoë fedtschenkoi*. See separate Excel file.

## Supplementary Notes

### Supplementary Note 1

#### CAM pathway genes in *Kalanchoë*

The CAM pathway can be divided into two temporally separated processes that bring about the photosynthetic assimilation of CO<sub>2</sub>: the carboxylation reactions at night that initially capture atmospheric CO<sub>2</sub> and convert it to malic acid, and the subsequent decarboxylation reactions during the day that regenerate this CO<sub>2</sub> and allow its assimilation via the Calvin-Benson cycle. Our metabolite profiling revealed that malic acid accumulation from nocturnal CO<sub>2</sub> fixation peaked during the dark period in mature *K. fedtschenkoi* leaf tissue (Supplementary Fig. 21). *Kalanchoë* species were instrumental in early biochemical studies that established the basic pathways of carbon flow during the CAM cycle<sup>54</sup>, and more recent studies have identified several key regulatory steps involved in controlling flux through these pathways<sup>55</sup>. Key components of the nocturnal carboxylation process include beta-carbonic anhydrase ( $\beta$ -CA), phosphoenolpyruvate carboxylase (PEPC), phosphoenolpyruvate carboxylase kinase (PPCK), NAD(P)-malate dehydrogenase (MDH), and the tonoplast membrane proteins responsible for the vacuolar accumulation of malic acid, which include the two energy-dependent H<sup>+</sup> pumps, the V-ATPase, and V-PPiase, and the inward-rectifying malate transporters such as aluminum-activated malate transporter (ALMT). In the subsequent light period, malic acid is remobilized from the vacuole by vacuolar malate transporters such as tonoplast dicarboxylate transporter (TDT) and decarboxylated in *Kalanchoë* via NAD(P)-malic enzyme (ME) and is further converted to phosphoenolpyruvate (PEP) by pyruvate phosphate dikinase (PPDK), which is activated in the light through dephosphorylation by PPDK regulatory protein (PPDK-RP), and CO<sub>2</sub> is released for assimilation by RuBisCO in the Calvin cycle (Fig. 3a). Transcriptome sequencing (RNA-seq) data was generated from mature CAM leaves of *K. fedtschenkoi* that were collected at 2 h intervals over a 12h light/12h dark cycle. Analysis of the resulting light/ dark cycles of transcript abundance profiles revealed candidate gene family members corresponding to each of the proposed core components of the CAM cycle (Supplementary Tables 5 and 6).

Gene duplication can be a major source of genetic novelty driving evolutionary innovation in plants<sup>56</sup>. Several CAM-related gene families in *K. fedtschenkoi*, including *PEPC*, *MDH*, *ALMT*, *NAD-ME*, and *NADP-ME* families, contain genes from recent genome duplication events, with 4DTV values ranging from 0.11 – 0.20 (Fig. 3a). Furthermore, differences in both transcript abundance and diel expression patterns between duplicated genes were identified (Fig. 3b). These results suggest that both recent WGD and functional diversification of genes by reprogramming of diel gene expression after duplication have contributed to the evolution of CAM-related gene families. This contrasts strikingly with the scenario inferred from the analysis of CAM gene evolution in the pineapple genome, where CAM-recruited genes were not found to have occurred through duplication and neofunctionalization, but instead the ancestral copy of each gene shared with grasses was simply elevated at the transcript level in CAM tissues, and in some cases also placed under light/dark and potentially circadian clock control<sup>57</sup>.

Nine  $\beta$ -CA genes are predicted in the *K. fedtschenkoi* genome (Supplementary Fig. 22a), among which three genes (Kaladp0095s0400, Kaladp0018s0287, and Kaladp0018s0289) have high transcript abundance (Supplementary Fig. 22b). The transcript expression of Kaladp0018s0287 and Kaladp0018s0289 showed pronounced diel changes (Supplementary Table 5), reaching peak abundance towards the middle of the night (Supplementary Fig. 22b). Because carboxylation of PEP by PEPC at night depends on the initial hydration of CO<sub>2</sub> by  $\beta$ -

CA, a CAM-specific role for this isoform of  $\beta$ -CA in supporting nocturnal carbon fixation is implied. In addition, the peak transcript abundance of Kaladp0095s0400 occurred in the late afternoon ([Supplementary Fig. 22b](#)), indicating that this gene might also be relevant to CAM under the assumption that protein expression is delayed relative to transcript expression. The relative importance of Kaladp0018s0287, Kaladp0018s0289, and Kaladp0095s0400 in nocturnal carbon fixation needs further investigation using genetic approaches (i.e., single and double mutants) in the future.

Five genes in the *K. fedtschenkoi* genome are predicted to encode PEPC ([Supplementary Fig. 23a](#)), which is one of the most abundant soluble proteins in CAM plants<sup>58</sup>. Among these, Kaladp0095s0055 was by far the most abundant transcript and it displayed a clear light/dark rhythm in transcript abundance with a peak during the late afternoon just before the start of the dark period ([Supplementary Fig. 23b](#)), making this gene a strong candidate for the CAM-specific member of the PEPC gene family. Kaladp0048s0578 was the second most abundant PEPC transcript among the five PEPC genes, with transcript abundance peaking around midnight ([Supplementary Fig. 23b](#)). Therefore, both Kaladp0095s0055 and Kaladp0048s0578 might contribute to nocturnal carbon fixation in *K. fedtschenkoi*.

The *K. fedtschenkoi* genome contains four PPCK genes (Kaladp0037s0517, Kaladp0050s0014, Kaladp0604s0001, and Kaladp0082s0192; [Supplementary Fig. 24a](#)) that are candidates for encoding the protein kinase that catalyzes the phosphorylation of PEPC. Kaladp0037s0517 corresponds to the *K. fedtschenkoi* gene that was the first to be characterized as encoding the functional PPCK in any CAM plant<sup>59</sup>. This gene showed the highest transcript abundance, and displayed a transcript abundance peak in the middle of the dark period ([Supplementary Fig. 24b](#)). These results indicate that Kaladp0037s0517 is the CAM-specific, nocturnally active PPCK gene.

The 15 MDH genes predicted in the *K. fedtschenkoi* genome can be divided into four tribes: tribe MDH1 comprising three genes ([Supplementary Fig. 25a](#)), tribe MDH2 comprising eight genes ([Supplementary Fig. 26a](#)), tribe MDH3 ([Supplementary Fig. 27a](#)) comprising three genes, and tribe MDH4 ([Supplementary Fig. 27c](#)) comprising one gene from *K. fedtschenkoi*. The transcript abundance varies among these genes ([Supplementary Figs. 25b, 26b, 27b and 27d](#)). Among these MDH genes, Kaladp0082s0194 (in tribe MDH1) had the highest transcript abundance and showed a peak transcript level in the late afternoon before the onset of the dark period ([Supplementary Fig. 25b](#)), temporally similar in phase to the expression of the PEPC gene Kaladp0095s0055 ([Supplementary Fig. 23b](#)). Kaladp0001s0257 (in tribe MDH2) had the second-most transcript abundance of MDH and also displayed a similar diel pattern of gene expression, with transcript levels peaking before dusk ([Supplementary Fig. 26b](#)).

Malate influx into the vacuole in *Kalanchoë* occurs via an anion-selective inward-rectifying ion channel<sup>60</sup>. The ALMT genes predicted in the *K. fedtschenkoi* genome can be divided into two tribes ([Supplementary Figs. 28a and 29a](#)) and there are variations in transcript abundance among these genes ([Supplementary Figs. 28b and 29b](#)). Among these ALMT genes, Kaladp0073s0021 is the most highly expressed and its transcript shows a strong diel change in abundance, with peak transcript occurring in the early part of the light period ([Supplementary Fig. 28b](#)). Another ALMT gene, Kaladp0062s0038, has lower transcript abundance than Kaladp0073s0021, but it has peak expression in the middle of the dark period ([Supplementary Fig. 28b](#)). Both genes are orthologous to the ALMT4/5/6 clade of *A. thaliana* genes, which includes ALMT6, a tonoplast-localized malate channel that mediates voltage-gated malate influx into the vacuole<sup>61</sup>. With the onset of the light period, malate efflux from the vacuole begins as

the nocturnally accumulated malic acid is remobilized as substrate for internal CO<sub>2</sub> release. The strong inward-rectification of the vacuolar malate channel suggests that a separate class of transport protein may be involved in malate export<sup>62</sup>. One candidate is the *K. fedtschenkoi* ortholog (Kaladp0042s0251) of the *Arabidopsis* TDT protein<sup>63</sup>, which showed a clear diel oscillation in transcript abundance (Supplementary Table 6), peaking during the first 4 hours of the light period (Supplementary Fig. 30).

Upon release of malate from the vacuole back into the cytoplasm, malate decarboxylation in the light phase of CAM in *Kalanchoë* is brought about by ME<sup>64, 65, 66</sup>. The alternative decarboxylase PEP carboxykinase (PEPCK) showed relatively low transcript abundance (Supplementary Table 6). Among the 13 ME family genes predicted in the *K. fedtschenkoi* genome (Supplementary Fig. 31a), Kaladp0092s0166 (*NADP-ME*) showed the highest transcript level, with abundance peaking at dawn (Supplementary Fig. 31b). PPDK, which converts pyruvate to PEP and initiates the gluconeogenic recovery of storage carbohydrate starch, was represented by only a single full-length gene in *K. fedtschenkoi* (Kaladp0076s0229). Kaladp0076s0229 showed a strong light/dark cycle of transcript abundance, with peak transcript level phased to dawn (Supplementary Fig. 32), consistent with the predicted function of this enzyme during pyruvate recycling to starch in the light period<sup>65, 66</sup>. Two genes (Kaladp0010s0106 and Kaladp0060s0363) in the *K. fedtschenkoi* genome are predicted for PPDK-RP, both of which showed higher transcript levels during the light period relative to the dark, and Kaladp0060s0363, in particular, displayed a peak transcript abundance phased to dusk (Supplementary Fig. 33), consistent with the regulatory phosphorylation of PPDK in the dark to inactivate it<sup>65</sup>.

## Supplementary Note 2

### Convergent evolution of genes in secondary metabolism

Secondary metabolism plays an important role in plant-environmental interactions. Plants synthesize various types of secondary metabolites, such as phenylpropanoids, glucosinolates, terpenoids, and phytoalexins/alkaloids<sup>67</sup>. Among the genes showing convergent changes in patterns of diel expression in two CAM species (*K. fedtschenkoi* and *A. comosus*) compared with a C<sub>3</sub> species (*A. thaliana*) (Supplementary Data 3), two (Kaladp0022s0177 and Kaladp0043s0207) were predicted to be involved in multiple secondary metabolic processes, including jasmonic acid biosynthesis (Supplementary Fig. 34), and aromatic amino acid biosynthesis via the shikimate pathway (Supplementary Fig. 35). Kaladp0043s0207 was classified into the co-expression module MEdarkgrey (Supplementary Data 3), which was associated with the leaf samples collected during the last two hours of the dark period (Supplementary Fig. 9), with higher transcript abundance than other diel time periods (Supplementary Fig. 10). Kaladp0022s0177 was classified into the co-expression module MEblack (Supplementary Data 3), which was associated with leaf samples collected 2-6 h after the beginning of the dark period (Supplementary Fig. 9), with higher transcript abundance than other diel time periods (Supplementary Fig. 10). The diel transcript abundance profiles of Kaladp0043s0207 and Kaladp0022s0177 and their corresponding *A. comosus* orthologs clustered together, distinct from the diel transcript abundance of their *Arabidopsis* ortholog (Supplementary Fig. 36). There was only a 1-h shift in diel transcript abundance pattern of Kaladp0043s0207 relative to its *A. comosus* ortholog Aco001151, whereas Kaladp0043s0207 and Aco001151 showed 12- and 11-h shifts relative to their best-matched *A. thaliana* ortholog AT2G21940, respectively (Supplementary Fig. 36a). The diel abundance of transcript of

Kaladp0022s0177 displayed only 1-h shift relative to its *A. comosus* ortholog Aco010785, whereas Kaladp0022s0177 and Aco010785 showed 8- and 7-h shifts relative to their best-matched *A. thaliana* ortholog AT4G29010, respectively ([Supplementary Fig. 36b](#)). Kaladp0043s0207 encodes shikimate kinase, an enzyme in the shikimate pathway that phosphorylates the C<sub>3</sub> hydroxyl group of shikimate to yield shikimate-3-phosphate, possibly providing a regulatory link between the energy-requiring shikimate pathway and cellular energy balance in plants <sup>68</sup>. Kaladp0022s0177 encodes the fatty acid beta-oxidation multifunctional protein AIM1/MFP2, which is involved in the biosynthesis of jasmonic acid (JA), an important regulator of plant development and stress response <sup>69</sup>. Because methyl jasmonate (MeJA), derived from JA, has a role in the induction of stomatal closure in various plant species <sup>70</sup>, this gene might influence the regulation of stomatal movement in CAM species.

## Supplementary Methods

### Supplementary Method 1

#### Analysis of positive selection in genes in CAM species

The protein-coding sequences of 24 plant species including *Amborella trichopoda* (PLAZA 3.0<sup>71</sup>; available at <http://bioinformatics.psb.ugent.be/plaza/>), *Ananas comosus*<sup>72</sup>, *Arabidopsis thaliana* (PLAZA 3.0), *Beta vulgaris* (PLAZA 3.0), *Brachypodium distachyon* (PLAZA 3.0), *Carica papaya* (PLAZA 3.0), *Citrus sinensis* (PLAZA 3.0), *Eucalyptus grandis* (PLAZA 3.0), *Fragaria vesca* (PLAZA 3.0), *Kalanchoë fedtschenkoi* (v1.1; available at <https://phytozome.jgi.doe.gov>), *Medicago truncatula* (PLAZA 3.0), *Mimulus guttatus* (PLAZA 3.0), *Musa acuminata* (PLAZA 3.0), *Oryza sativa* (PLAZA 3.0), *Phalaenopsis equestris*<sup>72</sup>, *Populus trichocarpa* (PLAZA 3.0), *Prunus persica* (PLAZA 3.0), *Setaria italica* (PLAZA 3.0), *Solanum lycopersicum* (PLAZA 3.0), *Solanum tuberosum* (PLAZA 3.0), *Sorghum bicolor* (PLAZA 3.0), *Theobroma cacao* (PLAZA 3.0), *Vitis vinifera* (PLAZA 3.0), *Zea mays* (PLAZA 3.0) were used for positive selection analysis. The longest protein-coding sequence was selected in case of multiple transcripts annotated for one gene locus. The protein-coding sequences were used for discovery of genes under positive selection in each of the CAM species in comparison with the non-CAM species using PosiGene, which is an automated pipeline for genome-wide detection of positively selected genes<sup>8</sup>, with default settings. In this positive selection analysis through the PosiGene pipeline, *K. fedtschenkoi* was used as the reference species for ortholog assignment and each of the three CAM species (i.e., *K. fedtschenkoi*, *A. comosus*, *P. equestris*) was used as anchor species.

### Supplementary Method 2

#### Plant material

The plants were grown in soil under a mixture of fluorescent incandescent lamps providing a photon flux density of 250  $\mu\text{mol m}^{-2} \text{s}^{-1}$  on a 12h light (25°C)/12h dark (18°C) cycle, and day/night relative humidities of 45/75%. In order to capture mRNA abundance changes responsive to diel conditions, samples were collected in triplicate from leaves every 2h over a 24h time course under 12h light/12h dark (L/D; Diel time). Additional tissues were sampled in triplicate including roots, flowers, shoot tips plus young leaves, and stems and frozen in liquid nitrogen until RNA isolation.

### Supplementary Method 3

#### Chromosome counting

*Kalanchoë* plantlets were grown on MS30 media (MS Modified Basal Medium supplemented with Gamborg's B5 vitamins pH 5.8, 3% (w/v) sucrose, 0.7% (w/v) Phytoagar) under a 16 h light/8 h dark cycle at 22°C, resulting in root production for mitotic chromosome preparation. Root tips were excised and treated in ice-cold distilled water for 24 h to induce mitotic arrest. Root tips were fixed in 3:1 ethanol and glacial acetic acid (v/v) for 24 h at room temperature and stored at 4 °C until used. After rinsing root tips in distilled water, root tips were transferred to microfuge tubes containing 50  $\mu\text{l}$  of enzyme solution containing 1% (w/v) Pectolyase Y-23 (MP) and 2% (w/v) Cellulase (BioWorld, Dublin, OH, USA) in citrate buffer (10 mM sodium citrate, 10 mM sodium EDTA, pH 5.5) for 1 h at 37 °C. Tubes were transferred

to ice and washed three times with 70% ethanol. Root tips were macerated in the residual approximately 100  $\mu$ l of liquid using a dissecting probe and centrifuged at 2000 x g for 5 sec or less. The supernatant was decanted and the tubes were inverted on a paper towel to dry. Each cell pellet was resuspended in 35  $\mu$ l glacial acetic acid at room temperature and 10  $\mu$ l of the cell suspension was applied on glass microscope slides and dried in a humidified chamber. Once the solvent had evaporated, one drop of Fluoroshield with DAPI histology mounting medium (Sigma-Aldrich, St. Louis, MO, USA) was applied per slide to dye chromosomes.

#### **Supplementary Method 4**

##### **Illumina sequencing of genome**

To isolate nuclear DNA for genome sequencing, fresh tissues of young leaves were collected, ground in liquid nitrogen, and stored at  $-80^{\circ}\text{C}$  until use. Intact nuclei from ground tissues were recovered in Sucrose-based Extraction Buffer as described<sup>73</sup>. The genomic DNA was extracted by a DNeasy plant maxi-kit (Qiagen, Valencia, CA, USA). TruSeq DNA PCR-Free Library Prep kit (Illumina, CA, USA) was used to generate the paired-end sequencing libraries with an average insert size around 540 bp. The mate-pair libraries with an average insert size of 3, 6, and 11 kb, respectively, were made using the Nextera Mate Pair Sample Prep Kit (Epicentre, WI, USA).

#### **Supplementary Method 5**

##### **Transcriptome sequencing**

A total of 48 samples including 36 mature leaf tissue samples (= 12 time-points  $\times$  3 biological replicates) and 12 other tissue samples (= 4 tissue types  $\times$  3 biological replicates) were collected. The plant samples were frozen in liquid nitrogen, ground using a mortar and pestle, and frozen at  $-80^{\circ}\text{C}$  until use. For total RNA isolation, approximately 100mg of frozen ground tissue was incubated in 850  $\mu$ l of CTAB buffer (1.0%  $\beta$ -Mercaptoethanol) at  $56^{\circ}\text{C}$  for 5 minutes, 600  $\mu$ l chloroform:isoamylalcohol (24:1) was added and samples were spun at full speed for 8 min. The supernatant ( $\sim$ 730  $\mu$ l) was removed from the top layer and applied to a filter column provided in the Spectrum<sup>TM</sup> Plant Total RNA Kit (Sigma, Cat. No. STRN250-1KT). RNA was precipitated in 500  $\mu$ l of 100% ethanol and applied to a Spectrum kit binding column. The protocol provided by the Spectrum kit was followed from that point on. An on-column DNase treatment was done to rid the samples of residual genomic DNA. RNA quantity and quality was quantified using a NanoDrop 1000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA) and 3  $\mu$ g of each sample was aliquoted for transcriptomics analysis. Total RNA samples were sent to HudsonAlpha Institute for Biotechnology (Huntsville, AL, USA) for library prep and sequencing. After passing quality checks, libraries were constructed using the Illumina TruSeq Stranded mRNA library prep kit (Illumina Inc., San Diego, CA). Each library was quantified using a QuBit, average fragment size was determined using an Agilent High Sensitivity Chip and qPCR was used to approximate optimal loading concentrations. Libraries were loaded onto 8 channels of an Illumina HiSeq V4 flowcell and paired end sequencing (2 $\times$ 150 bp) was used to generate reads of  $\sim$ 20 million bp for each sample on an Illumina HiSeq 2500 instrument (Illumina, CA, USA). Raw fastq files were provided for downstream analysis.



## Supplementary Method 6

### Protein-coding gene annotation

Repeat library was generated by RepeatModeler <sup>74</sup> from *K. fedtschenkoi* assembly. 50,414 transcript assemblies (TAs) were made from ~414 M paired-end reads by PERTRAN (Shu, unpublished). The TAs and ESTs were further assembled into PASA transcript assemblies (46,382), and 185,671 TAs and ESTs into sibling PASA transcript assemblies (62,501) by PASA <sup>75</sup>. Loci were determined by PASA transcript assembly alignments and/or EXONERATE alignments of proteins from *Arabidopsis thaliana*, rice, sorghum, mimulus, grape, orchid (*Phalaenopsis equestris*) and Swiss-Prot proteomes to repeat-soft-masked *K. fedtschenkoi* genome using RepeatMasker <sup>76</sup> with up to 2 K bp extension on both ends unless extending into another locus on the same strand. Gene models were predicted by homology-based predictors, FGENESH+ <sup>77</sup>, FGENESH\_EST (similar to FGENESH+, EST as splice site and intron input instead of protein/translated ORF), and GenomeScan <sup>78</sup>. The best scored predictions for each locus are selected using multiple positive factors including EST and protein support, and one negative factor: overlap with repeats. The selected gene predictions were improved by PASA. Improvement includes adding UTRs, splicing correction, and adding alternative transcripts. PASA-improved gene model proteins were subject to protein homology analysis to above mentioned proteomes to obtain Cscore and protein coverage. Cscore is a protein BLASTP score ratio to MBH (mutual best hit) BLASTP score and protein coverage is highest percentage of protein aligned to the best of homologs. PASA-improved transcripts were selected based on Cscore, protein coverage, EST coverage, and its CDS overlapping with repeats. The transcripts were selected if its Cscore is larger than or equal to 0.5 and protein coverage larger than or equal to 0.5, or it has EST coverage, but its CDS overlapping with repeats is less than 20%. For gene models whose CDS overlaps with repeats for more than 20%, its Cscore must be at least 0.9 and homology coverage at least 70% to be selected. The selected gene models were subject to Pfam and Panther analysis and gene models whose protein is more than 30% in Pfam/Panther TE domains were removed.

## Supplementary Method 7

### Construction of orthologous groups

The protein sequences of 26 plant species including *Amborella trichopoda* (PLAZA 3.0 <sup>71</sup>; available at <http://bioinformatics.psb.ugent.be/plaza/>), *Ananas comosus* <sup>72</sup>, *Aquilegia coerulea* (v1.1; available at <https://phytozome.jgi.doe.gov>), *Arabidopsis thaliana* (PLAZA 3.0), *Beta vulgaris* (PLAZA 3.0), *Brachypodium distachyon* (PLAZA 3.0), *Carica papaya* (PLAZA 3.0), *Citrus sinensis* (PLAZA 3.0), *Eucalyptus grandis* (PLAZA 3.0), *Fragaria vesca* (PLAZA 3.0), *Kalanchoë fedtschenkoi* (v1.1; available at <https://phytozome.jgi.doe.gov>), *Medicago truncatula* (PLAZA 3.0), *Mimulus guttatus* (PLAZA 3.0), *Musa acuminata* (PLAZA 3.0), *Oryza sativa* (PLAZA 3.0), *Phalaenopsis equestris* <sup>72</sup>, *Physcomitrella patens* (PLAZA 3.0), *Populus trichocarpa* (PLAZA 3.0), *Prunus persica* (PLAZA 3.0), *Setaria italica* (PLAZA 3.0), *Solanum lycopersicum* (PLAZA 3.0), *Solanum tuberosum* (PLAZA 3.0), *Sorghum bicolor* (PLAZA 3.0), *Theobroma cacao* (PLAZA 3.0), *Vitis vinifera* (PLAZA 3.0), *Zea mays* (PLAZA 3.0) were used for ortholog group construction. The longest protein sequence was selected in case of multiple transcripts annotated for one gene locus. The ortholog groups (OGs) were constructed using FastOrtho (<http://enews.patricbrc.org/fastortho/>), a reimplement of the OrthoMCL program

<sup>79</sup>, with default parameters (except a BLASTp E-value cutoff of 1e-5 and an inflation value of 1.3).

## Supplementary Method 8

### Construction of species phylogeny

Specifically, for each ortholog group of the single-copy genes, the protein sequences were aligned using MAFFT software version 7.221 <sup>80</sup>, using an accurate option (L-INS-i). The coding sequences (CDS) of each ortholog group were aligned using PAL2NAL <sup>81</sup>, guided by their corresponding protein sequence alignments. The species phylogeny was reconstructed from the protein or CDS alignments using the following three strategies:

*Concatenated supermatrix strategy:* The protein sequence alignments of these 210 single-copy-gene ortholog groups were then concatenated into one supermatrix file using FASconCAT <sup>82</sup>. The concatenated supermatrix file was used to infer maximum-likelihood phylogenetic tree using IQ-TREE version 1.4.3 <sup>1</sup>, with automatic model selection (-m TEST) and ultrafast bootstrap (-bb 1000) <sup>83</sup>.

*Partitioned analysis of multi-gene alignment strategy:* A concatenated protein sequence alignment with gene partition was generated for the 210 single-copy-gene ortholog groups using FASconCAT-G\_v1.02 <sup>84</sup>, with output in two different formats: the NEXUS format that was imbedded with MrBayes commands for Bayesian inference of species tree) and the FASTA format along with a partition file containing the best-fit models of amino acid replacement selected by using ProtTest version 3.4.2 <sup>85</sup>. The partitioned protein sequence alignment in NEXUS format was used for Bayesian inference of species tree using MrBayes version 3.2.6 <sup>2</sup>, with nst=6, rates=invgamma, mcmc ngen=10,000,000, printfreq=1000, samplefreq=1000. The run was stopped at 4,530,000 MCMC steps when the max standard deviation of split frequencies was 0.004577. Also, the partitioned protein sequence alignment in FASTA format along with a partition file containing the best-fit models of amino acid replacement was used for maximum likelihood inference of species tree using IQ-TREE version 1.4.4 <sup>1</sup>, with ultrafast bootstrap (-bb 1000) <sup>83</sup>.

*Coalescent-based species tree strategy:* The CDS alignments for each of the 210 single-copy-gene ortholog groups were used to reconstruct the multispecies coalescent tree using starBEAST2 <sup>3</sup> based on Bayesian inference of individual gene trees. Specifically, the site, clock and tree substitution models of these gene sets at all loci were unlinked. The uncorrelated lognormal relaxed clock model was applied in the simulations with the clock rate setting as 1.0, gene ploidies of 2.0 for all genes, and a constant population IO model. The population shape was set as 3 with an estimated mean of 1.0. The HKY+ $\Gamma$  site model with four gamma categories and estimated substitute rate and population invariant was chosen. The Markov chain Monte Carlo (MCMC) method was applied for 12 million states, and trees of the simulation were stored every five thousand states, resulting 2,400 trees at each gene locus. The tree evaluation was performed using ASTRAL-II <sup>4</sup> on the combination of the last 1,000 trees at all gene loci with 400 bootstrap replicates.

Also, individual maximum-likelihood gene trees were reconstructed from the CDS alignments for each of the 210 single-copy-gene ortholog groups using RAxML <sup>86</sup> with 500 bootstrap replicates. The gene trees were used to create a species tree using ASTRAL-II <sup>4</sup> analysis. Specifically, a coalescent species tree was first estimated by providing the best trees output from RAxML as well as all the bootstrap replicates. The ASTRAL-II analysis was run for 400 bootstraps, and the final “Astral” tree was treated as the species tree estimate. To estimate

quartet frequency support at each node and posterior probabilities, ASTRAL-II was re-run using a correctly rooted tree (at *Physcomitrella patens*) using the -q and -t 2 options.

## Supplementary Method 9

### Phylogenetic analysis of protein tribes

For phylogenetic analysis, the protein tribes that contain protein sequences from the 13 plant species listed in Supplementary Table 9 were used for analysis of convergent changes in protein sequences in CAM species. Specifically, the protein sequences of the 13 plant species in each tribe were aligned using MAFFT software version 7.221<sup>80</sup>, using an accurate option (L-INS-i). The protein sequence alignment was used to infer maximum-likelihood phylogenetic tree using IQ-TREE<sup>1</sup>, with automatic model selection (-m TEST) and ultrafast bootstrap (-bb 1000)<sup>83</sup>.

## Supplementary Method 10

### Analysis of convergence in protein sequences in CAM species

The tribes that contain at least one “CAM-convergence” clade in phylogenetic tree were used to identify convergent changes in protein sequence. Specifically, representative protein sequence (one sequence per species) was selected for each tribe containing “CAM-convergence” clade. The representative protein sequence for each tribe were aligned using MAFFT software version 7.221<sup>80</sup>, using an accurate option (L-INS-i). The protein sequence alignments and the species tree were used as input for the Grand-Convergence software<sup>87, 88, 89</sup>, which estimates the posterior numbers of convergent and divergent substitutions shared between all pairs of branches in the given phylogeny, to identify the convergent amino acid changes. The convergent sites (with probability > 0.99) predicted by the Grand-Convergence software were manual examined to identify amino acids mutations shared by a *K. fedtschenkoi* gene and its ortholog in at least one of the two monocot CAM species, as compared with C<sub>3</sub> and C<sub>4</sub> species. To check whether the mutations exist in any non-CAM species in the public databases, the *K. fedtschenkoi* protein sequences containing the mutation sites were used to search against the NCBI protein blast database (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>), the Phytozome protein blast database (<https://phytozome.jgi.doe.gov>), and the 1KP project protein blast database (<https://www.bioinfodata.org/Blast4OneKP/>) using a blastp e-value of 1e-5 and a maximum target number cutoff of 20,000. Then the protein sequences of the blast hits were searched for mutation sites using blastp with the *K. fedtschenkoi* mutation-containing sub-sequences (9-19 amino acids in length with the mutation sites in the middle) as queries and a blastp e-value of 10. As such, a *K. fedtschenkoi* gene is defined to be involved in convergent evolution in protein sequence if meeting the following three criteria: 1) the *K. fedtschenkoi* gene is placed together with gene(s) from at least one of the two monocot CAM species (*A. comosus* and *P. equestris*) in a phylogenetic clade that does not contain any genes from C<sub>3</sub> or C<sub>4</sub> species; 2) convergent amino-acid changes were detected (with probability > 0.99) by the Grand-Convergence software<sup>87, 88</sup> between the *K. fedtschenkoi* gene with gene(s) from at least one of the two monocot CAM species (*A. comosus* and *P. equestris*); 3) the *K. fedtschenkoi* gene shares at least one amino acid mutation with its ortholog in at least one of the two monocot CAM species, as compared with C<sub>3</sub> and C<sub>4</sub> species.

## Supplementary Method 11

### Gene Ontology analysis and pathway annotation

The enrichment of GO biological process was performed using ClueGO<sup>90</sup>, with the right-sided hypergeometric enrichment test at medium network specificity selection and *p*-value (hypergeometric enrichment test) correction was performed using the Holm-Bonferroni step-down method<sup>91</sup>. 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 GO-term group. The GO-terms with corrected  $P < 0.05$  (hypergeometric enrichment test) were considered significantly enriched. Pathway annotation for the protein sequences was performed on the KEGG Automatic Annotation Server KAAS<sup>92</sup>, using the BBH (bi-directional best hit) method to assign orthologs.

## Supplementary Method 12

### Analysis of co-expression network in *Kalanchoë*

The primary transcripts with an average abundance (calculated from three biological replicates) of greater than 1 FPKM in at least one of the 16 samples were utilized to construct a weighted gene co-expression network using the R package WGCNA<sup>93</sup>. The gene expression data were log<sub>2</sub> transformed. Modules were constructed using the following parameters: maxBlockSize=10000, power=9, networkType="unsigned", mergeCutHeight=0.25, minModuleSize=30, corType = "bicor". "bicor" represents the biweight midcorrelation, which is a robust correlation measure for co-expression network analysis<sup>94</sup>.

## Supplementary Method 13

### Cluster analysis of gene expression in *Kalanchoë*

Polynomial regressions of degree=4 were calculated for all transcripts using all 12 time points (i.e., 2, 4, 6, ..., 24 hours after the beginning of the light period) using counts=TRUE flag and a Bonferroni-Holm correction with  $P < 0.05$  (ANOVA of glm models where  $H_0$  = a flat line). Genes marked as influential by the T.fit() function were removed, and regressions were recalculated; only genes that were significantly "non-flat," or genes that could be explained via a polynomial regression, were used for subsequent analyses. Flat genes, or those that did not vary significantly across time points, were ignored. Clustering was done using the "mfuzz" option in maSigPro<sup>5,6</sup>, with the number of clusters being chosen by minimizing the within group sum of squares for  $k=1:20$ . A  $k$  of 11 coincided a maximum reduction of within group sum of squares, with no further reductions for  $k > 11$ . Mfuzz clustering also requires a user-defined  $m$  "fuzzification parameter," which was chosen using the mestimate() function in the Mfuzz R package<sup>95</sup>. An  $m$  of 1.06 was used for the *K. fedtschenkoi* expression data.

Networks for each cluster were constructed via ARACNE<sup>7</sup>. Transcripts were annotated as transcription factors using PlantTFcat<sup>96</sup>, and a list of these TFs were given to ARACNE for each cluster. ARACNE networks are built based on mutual information (MI) indices between genes – how much does one gene explain another. P-values are calculated based on the distribution of MI values in a network. Networks were therefore filtered for only interactions that had a *p*-value of  $< 0.01$  (as calculated by ARACNE). Networks were displayed in Cytoscape<sup>97</sup>, and analyzed using the built-in Network Analysis feature. The size of each node was set to indicate the number of directed edges, with a minimum of 10 edges required for the node to be

visualized. We highlighted the top 1% most connected nodes, as well as canonical CAM genes with at least 10 directed edges.

## Supplementary Method 14

### Comparative analysis of gene expression

To calculate the time-delay between time-course gene expression profiles, the diurnal expression data with two-hour intervals were further adjusted to one-hour interval time series by interpolation using the cubic spline function in MATLAB (Mathworks, Inc.). The expression data were then normalized by Z-score transformation. Pairwise correlation was calculated between each pair of orthologous genes for all possible time delays using the circular cross correlation function (cxcorr) in MATLAB (Mathworks, Inc.). The cxcorr function produces a correlation coefficient between gene 1 and shifted, or lagged, copies of gene 2 as a function of the lag. With each correlation coefficient, a lag value was given. The lag values were then converted into hours, giving an estimate on time delay. The time delay at which the correlation was maximum was selected as the estimated delay between the two genes.

To compare transcript expression between different time-windows, enrichment triangles were constructed to identify cross-species triangles within each ortholog group, in which a *K. fedtschenkoi* gene and a pineapple ortholog had significantly enriched expression in the same time-window, and an *Arabidopsis* ortholog of these genes had significantly enriched expression in the opposite time-window. Specifically, two types of comparison were performed: midday (4, 6 and 8 h from the starting of the light period) vs. midnight (16, 18 and 20 h from the starting of the light period) and dawn (22, 24 and 2 h from the starting of the light period) vs. dusk (10, 12 and 14 h from the starting of the light period), as illustrated in [Figure 4a](#). Any genes that, after preprocessing, had any negative expression values were removed from further analysis. Each gene expression vector  $X$  consisted of gene expression measurements at 12 time points (2h, 4h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 20h, 22h, 24h). Each entry  $x_i$  of each gene expression vector  $X$  was then transformed as follows:

$$x_i^* = \text{int} \left( \left( \frac{x_i}{\sum_j x_j} \right) \times 1000 \right)$$

Essentially, this normalizes each entry of the vector by dividing each entry by the sum of the vector, multiplying the normalized entry by 1000 and taking the integer of the result. This ensures that the values of the transformed vector have whole number values between 0 and 1000. For each species, we then create two-column matrices, namely a midday-midnight matrix and a dawn-dusk matrix for each species. The midday-midnight matrix is constructed such that rows represent genes from a particular species and columns represent light states. One column represents the sum of all transformed midday time points for a given gene and the other column represents the sum of all transformed midnight time points for a given gene. Similarly, we construct a dawn-dusk matrix for each species, in which rows represent genes from the particular species and columns represent either the sum of all dawn time points or the sum of all dusk time points.

For each gene  $i$ , we used the right-tailed Fisher Exact Test to determine if that gene's expression was enriched in midday or midnight (or neither) and dusk or dawn (or neither) according to the contingency tables in [Supplementary Figure 37](#), where  $d_{xy}$  refers to element  $xy$

in the dawn-dusk matrix  $D$ , and  $m_{xy}$  refers to element  $xy$  in the midday-midnight matrix  $M$ . This made use of the Text::NSP::Measures::2D::Fisher Perl module <sup>98</sup> available from the Comprehensive Perl Archive Network (<http://www.cpan.org/>). The False Discovery Rate <sup>99</sup> was controlled per species and light comparison (midday/midnight or dawn/dusk) at a level less than 0.01. For a given comparison (midday vs. midnight or dawn vs. dusk), triangles were identified within families as groups of 3 orthologous genes, (a *Kalanchoë* gene, a pineapple gene and an *Arabidopsis* gene) where the *Kalanchoë* gene and the pineapple genes were enriched in the same light state (e.g., midday) and the *Arabidopsis* gene was enriched in the opposite light state (e.g., midnight). The resulting triangles were visualized as networks in Cytoscape <sup>97</sup>.

## Supplementary Method 15

### Genome synteny analysis

For each pairwise alignments, the coding sequences of predicted gene models are compared to each other using LAST <sup>100</sup>. Our synteny search pipeline defines syntenic blocks by chaining the LAST hits with a distance cutoff of 20 genes apart, also requiring at least four gene pairs per synteny block. For homologs inferred based on the syntenic alignments, we aligned the protein sequences using CLUSTALW <sup>101</sup> and used the protein alignments to guide coding sequence alignments by PAL2NAL <sup>81</sup>. To calculate  $K_s$ , we used the Nei-Gojobori method implemented in *yn00* program in PAML package <sup>102</sup>. In-house Python scripts ([https://github.com/tanghaibao/bio-pipeline/tree/master/synonymous\\_calculation](https://github.com/tanghaibao/bio-pipeline/tree/master/synonymous_calculation)) were used to pipeline all the calculations. MCScanX <sup>103</sup> were used to detect gene synteny and collinearity for seven dicot species including *Arabidopsis thaliana* (PLAZA 3.0), *Carica papaya* (PLAZA 3.0), *K. fedtschenkoi* (v1.1; <https://phytozome.jgi.doe.gov/>), *Populus trichocarpa* (PLAZA 3.0), *Solanum lycopersicum* (PLAZA 3.0), *Theobroma cacao* (PLAZA 3.0), and *Vitis vinifera* (PLAZA 3.0). Four-fold transversion substitutions (4dvt) between syntenic gene pairs were calculated as described previously <sup>104</sup>.

## Supplementary Method 16

### Gas chromatography-mass spectrometry metabolite profiling

The plant samples were frozen in liquid nitrogen and ground using a mortar and pestle, and frozen at  $-80^{\circ}\text{C}$  until analyzed. Approximately 100 mg FW of each sample was 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. Sorbitol was added before 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 silylated for 1 h at  $70^{\circ}\text{C}$  to generate trimethylsilyl (TMS) derivatives, which were analyzed after 2 d with an Agilent Technologies Inc. (Santa Clara, CA, USA) 5975C inert XL gas chromatograph-mass spectrometer as described elsewhere <sup>105</sup>. Metabolite peak extraction, identification, and quantification were as described previously <sup>58, 105</sup>, and unidentified metabolites were denoted by their retention time as well as key mass-to-charge ( $m/z$ ) ratios. TCA cycle organic acids, sugars, and abundant secondary metabolites are known or thought to be under diurnal regulation were the focus of this study.

## Supplementary Method 17

### ***In vitro* protein expression and analysis of enzyme activity**

The full-length coding sequences of phosphoenolpyruvate carboxylase (PEPC) genes were synthesized by IDT (<https://www.idtdna.com/site>), and were then inserted into the bacterial expression vector pGEX6P1 with restriction digestion *BamH I*. The bacterial BL21 strains (Novagen BL21 (DE3) pLysS Singles), which carry the plasmids (including wild-type and mutants), were used for target protein induction with IPTG, and the expressed proteins were then purified using glutathione sepharose 4B beads (GE Healthcare Life Sciences, Pittsburgh, PA, USA). After cleavage of GST with PreScission Protease (GE Healthcare Life Sciences, Pittsburgh, PA, USA), the final target protein concentrations were determined by using Qubit fluorometer 2.0 (Invitrogen, Carlsbad, CA, USA). The protein quality was checked via Western blot. Approximately 1 µg purified proteins were used for running with 10% SDS-Polyacrylamide gel electrophoresis (PAGE) gel and transferred onto PVDF transfer membrane. After blocking with 5% BSA to reduce non-specific binding, the membrane was then incubated with primary anti-PEPC antibody (Agrisera, Sweden) for 1- to 2-h and followed incubation with alkaline phosphatase conjugated secondary anti-IgG antibodies (Sigma-Aldrich, St. Louis, MO, USA) for another 1 h. The signal of target proteins was detected by using 1-Step NBT/BCIP Solution (Thermo Scientific, Wilmington, DE, USA). The PEPC activity was determined using a NADH-mediated assay system according to protocols of Shi *et al.*<sup>106</sup> and Gregory *et al.*<sup>107</sup>. The reactions were performed at 25°C and final PEPC enzyme activity was defined as oxidation of 1 mM NADH (optical density at 340 nm) per minute using SpectraMax Plus 384 (Molecular Devices, Sunnyvale, CA, USA). Approximately 1 µg purified proteins were applied in 1 mL reaction buffer, containing 50 mM HEPES-KOH (pH 8), 15% (v/v) glycerol, 2 mM PEP, 2 mM KHCO<sub>3</sub>, 5 mM MgCl<sub>2</sub>, 2 mM dithiothreitol, 0.15 mM NADH, and 5 units/mL malate dehydrogenase.

## Supplementary References

1. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* **32**, 268-274 (2015).
2. Ronquist F, *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology* **61**, 539-542 (2012).
3. Ogilvie HA, Bouckaert RR, Drummond AJ. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution* DOI: <https://doi.org/10.1093/molbev/msx126>, (2017).
4. Mirarab S, Warnow T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **31**, 44-52 (2015).
5. Conesa A, Nueda MJ, Ferrer A, Talón M. maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics* **22**, 1096-1102 (2006).
6. Nueda MJ, Tarazona S, Conesa A. Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* **30**, 2598-2602 (2014).
7. Margolin AA, *et al.* ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC bioinformatics* **7 Suppl 1**, S7 (2006).
8. Sahm A, Bens M, Platzer M, Szafranski K. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. *Nucleic Acids Research*, (2017).
9. Ma Y, *et al.* Regulators of PP2C phosphatase activity function as abscisic acid sensors. *Science* **324**, 1064-1068 (2009).
10. Wang Y, Noguchi K, Ono N, Inoue S-i, Terashima I, Kinoshita T. Overexpression of plasma membrane H<sup>+</sup>-ATPase in guard cells promotes light-induced stomatal opening and enhances plant growth. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 533-538 (2014).
11. Nieves-Cordones M, Caballero F, Martínez V, Rubio F. Disruption of the *Arabidopsis thaliana* inward-rectifier K<sup>+</sup> channel AKT1 improves plant responses to water stress. *Plant and Cell Physiology* **53**, 423-432 (2012).
12. De Angeli A, Zhang J, Meyer S, Martinoia E. AtALMT9 is a malate-activated vacuolar chloride channel required for stomatal opening in *Arabidopsis*. *Nature Communications* **4**, 1804 (2013).
13. Shang Y, Dai C, Lee MM, Kwak JM, Nam KH. BRI1-Associated Receptor Kinase 1 regulates guard cell ABA signaling mediated by Open Stomata 1 in *Arabidopsis*. *Molecular Plant* **9**, 447-460 (2016).
14. Takemiya A, *et al.* Phosphorylation of BLUS1 kinase by phototropins is a primary step in stomatal opening. *Nature Communications* **4**, (2013).
15. Geiger D, *et al.* Guard cell anion channel SLAC1 is regulated by CDPK protein kinases with distinct Ca<sup>2+</sup> affinities. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 8023-8028 (2010).
16. Mori IC, *et al.* CDPKs CPK6 and CPK3 function in ABA regulation of guard cell S-type anion- and Ca<sup>2+</sup>-permeable channels and stomatal closure. *PLoS Biology* **4**, e327 (2006).
17. Hashimoto M, Negi J, Young J, Israelsson M, Schroeder JI, Iba K. Arabidopsis HT1 kinase controls stomatal movements in response to CO<sub>2</sub>. *Nature Cell Biology* **8**, 391-397 (2006).
18. Szyroki A, *et al.* KAT1 is not essential for stomatal opening. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 2917-2921 (2001).
19. Lebaudy A, *et al.* Plant adaptation to fluctuating environment and biomass production are strongly dependent on guard cell potassium channels. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5271-5276 (2008).



20. Mustilli A-C, Merlot S, Vavasseur A, Fenzi F, Giraudat J. *Arabidopsis* OST1 protein kinase mediates the regulation of stomatal aperture by abscisic acid and acts upstream of reactive oxygen species production. *Plant Cell* **14**, 3089-3099 (2002).
21. Merlot S, *et al.* Constitutive activation of a plasma membrane H<sup>+</sup> - ATPase prevents abscisic acid - mediated stomatal closure. *EMBO Journal* **26**, 3216-3226 (2007).
22. Kinoshita T, Doi M, Suetsugu N, Kagawa T, Wada M, Shimazaki K-i. Phot1 and phot2 mediate blue light regulation of stomatal opening. *Nature* **414**, 656-660 (2001).
23. Park S-Y, *et al.* Abscisic acid inhibits type 2C protein phosphatases via the PYR/PYL family of START proteins. *Science* **324**, 1068-1071 (2009).
24. Meyer S, *et al.* AtALMT12 represents an R - type anion channel required for stomatal movement in *Arabidopsis* guard cells. *Plant Journal* **63**, 1054-1062 (2010).
25. Vahisalu T, *et al.* SLAC1 is required for plant guard cell S-type anion channel function in stomatal signalling. *Nature* **452**, 487-491 (2008).
26. Exner V, *et al.* A gain-of-function mutation of *Arabidopsis* cryptochrome1 promotes flowering. *Plant Physiology* **154**, 1633-1645 (2010).
27. El-Assal SE-D, Alonso-Blanco C, Peeters AJ, Wagemaker C, Weller JL, Koornneef M. The role of cryptochrome 2 in flowering in *Arabidopsis*. *Plant Physiology* **133**, 1504-1516 (2003).
28. Dixon LE, Knox K, Kozma-Bognar L, Southern MM, Pokhilko A, Millar AJ. Temporal repression of core circadian genes is mediated through EARLY FLOWERING 3 in *Arabidopsis*. *Current Biology* **21**, 120-125 (2011).
29. Baudry A, *et al.* F-box proteins FKF1 and LKP2 act in concert with ZEITLUPE to control *Arabidopsis* clock progression. *Plant Cell* **22**, 606-622 (2010).
30. Shen Y, *et al.* Phytochrome A mediates rapid red light-induced phosphorylation of *Arabidopsis* FAR-RED ELONGATED HYPOCOTYL1 in a low fluence response. *Plant Cell* **21**, 494-506 (2009).
31. Jarillo JA, *et al.* An *Arabidopsis* circadian clock component interacts with both CRY1 and phyB. *Nature* **410**, 487-490 (2001).
32. Alabadí D, Oyama T, Yanovsky MJ, Harmon FG, Más P, Kay SA. Reciprocal regulation between *TOC1* and *LHY/CCA1* within the *Arabidopsis* circadian clock. *Science* **293**, 880-883 (2001).
33. Pruneda-Paz JL, Breton G, Para A, Kay SA. A functional genomics approach reveals CHE as a component of the *Arabidopsis* circadian clock. *Science* **323**, 1481-1485 (2009).
34. Dalchau N, *et al.* The circadian oscillator gene *GIGANTEA* mediates a long-term response of the *Arabidopsis thaliana* circadian clock to sucrose. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 5104-5109 (2011).
35. Nusinow DA, *et al.* The ELF4-ELF3-LUX complex links the circadian clock to diurnal control of hypocotyl growth. *Nature* **475**, 398-402 (2011).
36. Bauer D, *et al.* Constitutive photomorphogenesis 1 and multiple photoreceptors control degradation of phytochrome interacting factor 3, a transcription factor required for light signaling in *Arabidopsis*. *Plant Cell* **16**, 1433-1445 (2004).
37. Malapeira J, Khaitova LC, Mas P. Ordered changes in histone modifications at the core of the *Arabidopsis* circadian clock. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 21540-21545 (2012).
38. Nakamichi N, Kiba T, Henriques R, Mizuno T, Chua N-H, Sakakibara H. PSEUDO-RESPONSE REGULATORS 9, 7, and 5 are transcriptional repressors in the *Arabidopsis* circadian clock. *Plant Cell* **22**, 594-605 (2010).
39. Huang W, *et al.* Mapping the core of the *Arabidopsis* circadian clock defines the network structure of the oscillator. *Science* **336**, 75-79 (2012).
40. Más P, Kim W-Y, Somers DE, Kay SA. Targeted degradation of *TOC1* by *ZTL* modulates circadian function in *Arabidopsis thaliana*. *Nature* **426**, 567-570 (2003).

41. Rugnone ML, *et al.* LNK genes integrate light and clock signaling networks at the core of the *Arabidopsis* oscillator. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 12120-12125 (2013).
42. Rawat R, *et al.* REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin pathways. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 16883-16888 (2009).
43. Hsu PY, Devisetty UK, Harmer SL. Accurate timekeeping is controlled by a cycling activator in *Arabidopsis*. *eLife* **2**, e00473 (2013).
44. Rawat R, *et al.* REVEILLE8 and PSEUDO-REPONSE REGULATOR5 form a negative feedback loop within the *Arabidopsis* circadian clock. *PLoS Genetics* **7**, e1001350 (2011).
45. Perales M, Portolés S, Más P. The proteasome - dependent degradation of CKB4 is regulated by the *Arabidopsis* biological clock. *Plant Journal* **46**, 849-860 (2006).
46. Kim J, Kim Y, Yeom M, Kim J-H, Nam HG. FIONA1 is essential for regulating period length in the *Arabidopsis* circadian clock. *Plant Cell* **20**, 307-319 (2008).
47. Toledo-Ortiz G, *et al.* The HY5-PIF regulatory module coordinates light and temperature control of photosynthetic gene transcription. *PLoS Genetics* **10**, e1004416 (2014).
48. Jones MA, Covington MF, DiTacchio L, Vollmers C, Panda S, Harmer SL. Jumonji domain protein JMJD5 functions in both the plant and human circadian systems. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21623-21628 (2010).
49. Wang Y, Wu J-F, Nakamichi N, Sakakibara H, Nam H-G, Wu S-H. LIGHT-REGULATED WD1 and PSEUDO-RESPONSE REGULATOR9 form a positive feedback regulatory loop in the *Arabidopsis* circadian clock. *Plant Cell* **23**, 486-498 (2011).
50. Hong S, Song H-R, Lutz K, Kerstetter RA, Michael TP, McClung CR. Type II protein arginine methyltransferase 5 (PRMT5) is required for circadian period determination in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 21211-21216 (2010).
51. Wang X, *et al.* SKIP is a component of the spliceosome linking alternative splicing and the circadian clock in *Arabidopsis*. *Plant Cell* **24**, 3278-3295 (2012).
52. Jones MA, Williams BA, McNicol J, Simpson CG, Brown JW, Harmer SL. Mutation of *Arabidopsis* SPLICEOSOMAL TIMEKEEPER LOCUS1 causes circadian clock defects. *Plant Cell* **24**, 4066-4082 (2012).
53. Panda S, Poirier GG, Kay SA. *tej* defines a role for poly (ADP-ribosyl) ation in establishing period length of the *Arabidopsis* circadian oscillator. *Developmental Cell* **3**, 51-61 (2002).
54. Osmond C, Holtum J. Crassulacean acid metabolism. In: *The Biochemistry of Plants, Volume 8: Photosynthesis* (ed Hatch M, Boardman N). Academic Press (1981).
55. Borland AM, *et al.* Engineering crassulacean acid metabolism to improve water-use efficiency. *Trends in Plant Science* **19**, 327-338 (2014).
56. Qian W, Zhang J. Genomic evidence for adaptation by gene duplication. *Genome Research* **24**, 1356-1362 (2014).
57. Ming R, *et al.* The pineapple genome and the evolution of CAM photosynthesis. *Nature Genetics* **47**, 1435-1442 (2015).
58. Abraham PE, *et al.* Transcript, protein and metabolite temporal dynamics in the CAM plant *Agave*. *Nature Plants* **2**, Article number: 16178 (2016).
59. Hartwell J, Nimmo G, Wilkins M, Jenkins G, Nimmo H. Phosphoenolpyruvate carboxylase kinase is a novel protein kinase is a novel protein kinase regulated at the level of gene expression. *Plant Journal* **20**, 333-342 (1999).
60. Hafke JB, Hafke Y, Smith JAC, Lüttge U, Thiel G. Vacuolar malate uptake is mediated by an anion - selective inward rectifier. *Plant Journal* **35**, 116-128 (2003).

61. Meyer S, *et al.* Malate transport by the vacuolar AtALMT6 channel in guard cells is subject to multiple regulation. *Plant Journal* **67**, 247-257 (2011).
62. Smith JAC, *et al.* Transport across the vacuolar membrane in CAM plants. In: *Crassulacean Acid Metabolism* (ed Winter K, Smith J). Springer-Verlag (1996).
63. Emmerlich V, *et al.* The plant homolog to the human sodium/dicarboxylic cotransporter is the vacuolar malate carrier. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 11122-11126 (2003).
64. Dittrich P, Campbell WH, Black CC. Phosphoenolpyruvate carboxykinase in plants exhibiting crassulacean acid metabolism. *Plant Physiology* **52**, 357-361 (1973).
65. Dever LV, Boxall SF, Kneřová J, Hartwell J. Transgenic perturbation of the decarboxylation phase of crassulacean acid metabolism alters physiology and metabolism but has only a small effect on growth. *Plant Physiology* **167**, 44-59 (2015).
66. Hartwell J, Dever LV, Boxall SF. Emerging model systems for functional genomics analysis of crassulacean acid metabolism. *Current Opinion in Plant Biology* **31**, 100-108 (2016).
67. Kliebenstein D. Secondary metabolites and plant/environment interactions: a view through *Arabidopsis thaliana* tinged glasses. *Plant, Cell & Environment* **27**, 675-684 (2004).
68. Maeda H, Dudareva N. The shikimate pathway and aromatic amino acid biosynthesis in plants. *Annual Review of Plant Biology* **63**, 73-105 (2012).
69. Delker C, Zolman BK, Miersch O, Wasternack C. Jasmonate biosynthesis in *Arabidopsis thaliana* requires peroxisomal  $\beta$ -oxidation enzymes—Additional proof by properties of *pex6* and *aim1*. *Phytochemistry* **68**, 1642-1650 (2007).
70. Munemasa S, Hossain MA, Nakamura Y, Mori IC, Murata Y. The *Arabidopsis* calcium-dependent protein kinase, CPK6, functions as a positive regulator of methyl jasmonate signaling in guard cells. *Plant Physiology* **155**, 553-561 (2011).
71. Proost S, *et al.* PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Research* **43**, D974-981 (2015).
72. Cai J, *et al.* The genome sequence of the orchid *Phalaenopsis equestris*. *Nature Genetics* **47**, 65-72 (2015).
73. Carrier G, *et al.* An efficient and rapid protocol for plant nuclear DNA preparation suitable for next generation sequencing methods. *American Journal of Botany* **98**, e13-15 (2011).
74. Smit AFA, Hubley R. RepeatModeler Open-1.0. <http://www.repeatmasker.org>, (2008-2015).
75. Haas BJ, *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666 (2003).
76. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. <http://www.repeatmasker.org>, (1996-2012).
77. Salamov AA, Solovyev VV. Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research* **10**, 516-522 (2000).
78. Yeh RF, Lim LP, Burge CB. Computational inference of homologous gene structures in the human genome. *Genome Research* **11**, 803-816 (2001).
79. Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13**, 2178-2189 (2003).
80. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**, 772-780 (2013).
81. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research* **34**, W609-612 (2006).
82. Kuck P, Meusemann K. FASconCAT: Convenient handling of data matrices. *Molecular Phylogenetics and Evolution* **56**, 1115-1118 (2010).

83. Minh BQ, Nguyen MA, von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* **30**, 1188-1195 (2013).
84. Kück P, Longo GC. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* **11**, (2014).
85. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164-1165 (2011).
86. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688-2690 (2006).
87. Castoe TA, *et al.* Evidence for an ancient adaptive episode of convergent molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 8986-8991 (2009).
88. Qian C, Bryans N, Kravkov I, de Koning APJ. Visualization and analysis of statistical signatures of convergent molecular evolution. *University of Calgary* <http://labjasondkio>, (2015).
89. Qian C, de Koning APJ. Rapid discovery of convergent molecular evolution across entire phylogenies. *University of Calgary* <http://labjasondkio>, (2015).
90. Bindea G, *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091-1093 (2009).
91. Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**, 65-70 (1979).
92. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**, W182-W185 (2007).
93. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
94. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software* **46**, i11 (2012).
95. Kumar L, E. Futschik M. Mfuzz: A software package for soft clustering of microarray data. *Bioinformatics* **2**, 5-7 (2007).
96. Dai X, Sinharoy S, Udvardi M, Zhao PX. PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC bioinformatics* **14**, 321 (2013).
97. Shannon P, *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**, 2498-2504 (2003).
98. Banerjee S, Pedersen T. The design, implementation, and use of the Ngram statistics package. In: *Computational Linguistics and Intelligent Text Processing. CICLing 2003. Lecture Notes in Computer Science, vol 2588* (ed Gelbukh A). Springer (2003).
99. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **57**, 289-300 (1995).
100. Kielbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. *Genome Research* **21**, 487-493 (2011).
101. Larkin MA, *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
102. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
103. Wang Y, *et al.* MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**, e49 (2012).
104. Tuskan GA, *et al.* The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**, 1596-1604 (2006).
105. Tschaplinski TJ, *et al.* Down-regulation of the caffeic acid O-methyltransferase gene in switchgrass reveals a novel monolignol analog. *Biotechnology for Biofuels* **5**, 71 (2012).

106. Shi J, *et al.* Phosphoenolpyruvate carboxylase in *Arabidopsis* leaves plays a crucial role in carbon and nitrogen metabolism. *Plant Physiology* **167**, 671-681 (2015).
107. Gregory AL, *et al.* *In vivo* regulatory phosphorylation of the phosphoenolpyruvate carboxylase AtPPC1 in phosphate-starved *Arabidopsis thaliana*. *Biochemical Journal* **420**, 57-65 (2009).