RESEARCH ARTICLE TITLE

Genomic and transcriptomic analysis of sacred fig (*Ficus religiosa*)

KL Ashalatha [1] , KP Arunkumar[2], Malali Gowda[1, 3*]

[*]**CORRESPONDENCE:** malali.gene@gmail.com

[1,3]Malali Gowda

The University of Trans-Disciplinary Health Sciences and Technology

Yelahanka, Bengaluru, 560064, India

Abstract

Background: Peepal/Bodhi tree (*Ficus religiosa* L.) is an important, long-lived keystone ecological species. Communities on the Indian subcontinent have extensively employed the plant in Ayurveda, traditional medicine, and spiritual practices. The Peepal tree is often thought to produce oxygen both during the day and at night by Indian folks. The goal of our research was to produce molecular resources using whole-genome and transcriptome sequencing techniques.

Results: The whole genome of the Peepal tree was sequenced using two next-generation sequencers like Illumina HiSeq1000 and MGISEQ-2000. We assembled the draft genome of 406Mb, using a hybrid assembly workflow. The genome annotation resulted in 35,093 protein-coding genes; 53% of its genome consists of repetitive sequences. To understand the physiological pathways in leaf tissues, we analyzed photosynthetically distinct conditions: bright sunny days and nights. The RNA-seq analysis supported the expression of 26,691 genes. The leaf transcriptomic analysis of the diurnal and nocturnal periods revealed the expression of significant genes involved in the carbon-fixation pathway.

Conclusions: This study presents a draft hybrid genome assembly for *F. religiosa* and its functional annotated genes. The genomic and transcriptomic data-derived pathways are obtained for future studies on the Peepal tree.

Keywords: *Ficus religiosa*, Peepal, Bodhi, Hybrid genome, Transcriptome, Carbon-fixation pathway

Background

The Peepal tree (*Ficus religiosa* L.) is a sacred fig, hemi- epiphyte that belongs to the Moraceae family and has a diploid sporophytic count (2n = 26) [1]. It is known to be a long-lived deciduous species related to the 755 fig species widespread worldwide [2]. The Peepal tree is a cosmopolitan species, having value for cultural and spiritual practices in Buddhism, Hinduism, and Jainism. It is popularly called the Bodhi tree, where Buddha is believed to have meditated and attained spiritual enlightenment underneath this tree. Hence, the culture is spread across Asia and it has been worshipped. Peepal has several vernacular names, like Pippali, Ashwatha, Arali, and so on; it is frequently found together with the Neem tree near Indian Temple [3]. Generally, the Peepal tree has a special significance in communities across India that are believed to produce oxygen day and night. They have a special type of stomata called sunken, giant, or hydathode at the lower leaf epidermis of *F. religiosa*. These are larger than the normal stomata and occur over the veins or are mixed with normal stomata, it indicates that such stomata hold gaseous and water molecules for a longer time [4]. According to our knowledge, there is no scientific evidence to claim oxygen production during a dark period.

In Ayurveda, the Peepal tree has been classified as a Rasayana (a type of drug), whereby rejuvenators and antioxidants aid in relieving the body's stress [5]. Peepal tree alleviates Pitta

50 and Kapha (Ayurvedic classifications), hence prescribed for treatment of the disorders like
51 respiratory and inflammatory disorders, ulcers, stomatitis, hiccup, arthritis, gout, skin diseases,
52 bone fracture, diabetes, etc., [5]. In animal models such as rats, the Peepal tree has been tested
53 for the treatment of neurodegenerative disorders such as Parkinson's disease and Huntington's
54 disease [6], [7], as well as anti-ulcer activity in albino mice [8].
55
56 Next-generation sequencing (NGS) technologies have accelerated the generation of draft
57 genome sequences of Moraceae plant species, the genome size of *Morus notabilis* is 330Mb
58 [9], 333Mb for *Ficus carica* [10], 436Mb for *F. microcarpa*, and 370Mb for *F. hispida* [11].
59 The genome sequencing of non-model plant species Peepal was first mentioned in The Neem
60 Genome book chapter [3]. A recent study has generated the genomic resource of *F.*
61 *religiosa* (332Mb) and *F. benghalensis* (392Mb). However, they generated a limited size of
62 genome assembly and genes (23,929) for the Peepal tree when compared with the present study
63 [12]. Recently, several studies have attempted sequencing non-model plant species like
64 pineapple (*Ananas comosus*) and *Kalanchoë* species revealing the gene expressions of the
65 Crassulacean acid metabolism pathway [13] and [14]. The whole-genome sequencing of
66 common or crystalline ice plants (*Mesembryanthemum crystallinum*) has been shown to switch
67 from Calvin- Benson Cycle (C3) to CAM photosynthesis under a salt stress [15]. Using RNA-
68 seq, the study described by comparing both species with and without the C4 trait and different
69 tissues within a C4 plant suggests the ways of integration into the underlying C3 metabolism
70 [16]. These previous study findings and other physiological features of the Peepal tree enabled
71 us to characterize the pathways in the present work.
72
73 Despite its ecological, medicinal, cultural, and historic importance, the molecular biology and
74 genomics studies on the Peepal tree are scanty. Hence, Peepal is relevant to traditional

75 medicinal practices and Buddha's meditation, we hypothesized to elucidate the genomic
76 sequence and to study the transcriptome of photosynthetic tissue (leaf tissue) in diurnal and
77 nocturnal periods. The objective of the present study was to generate a genome sequence and
78 annotate genes for the Peepal tree. The transcriptomic analysis has been undertaken to identify
79 the expression of genes in the diurnal and nocturnal periods for photosynthetic activity using a
80 molecular approach. In this study, we aim to characterize the genes involved in various
81 physiological, biochemical metabolic, and other pathways. And also, a comparative genomic
82 study has been carried out to analyze the Peepal tree relationship between closely related
83 species of its Moraceae family.
84
85 **Results**
86 ***De novo* hybrid assembly using Illumina and MGI short reads**
87 We used two next-generation technology platforms to sequence the whole genome of the fig
88 species, the Peepal tree. Paired-end reads from Illumina HiSeq1000 and MGISEQ-2000 were
89 266 and 645 million respectively; a total data of 88.44 billion high-quality bases (Quality>20)
90 used for genome assembly. A hybrid assembly was performed using a sequencing depth of
91 65.5X Illumina reads and 158.86X MGI reads. The raw data details are given (Additional file
92 1: Table S1.1). We evaluated the distribution of k-mers in both Illumina and MGI reads to
93 estimate the genome size, resulting in a genome size of 393-Mb and 273-Mb respectively. The
94 combination of Illumina and MGI reads was used for assembling the genome. Hybrid genome
95 assembly yielded a genome of 406-Mb. The contig N50 length is 5,817 bp and the largest
96 contig length is 148Kb. The GC content of the Peepal tree genome is 34.23%. The gap-closing
97 step was performed for the hybrid assembly. There were 35,811 (5.5%) misassembled contigs
98 and 6,04,807 (94.4%) truly assembled contig sequences in the final assembled genome. The
99 workflow of the genome assembly is represented in (Additional file 2: Figure S1). The statistics

100 of assembly contigs and scaffolds are shown (Additional file 3: Table S1.2) and the final
101 scaffold assembly of the genome is given in Table 1.
102
103 The completeness of the Peepal tree genome assembly was assessed with the BUSCO tool. The
104 results showed that 76.5% (232 out of 303 genes) and 84.1% (1210 out of 1440 genes) of genes
105 were conserved as single-copy orthologs in eukaryotic and plant universal data sets,
106 respectively. Out of 232 complete genes in the Eukaryota database, 214 are single-copy
107 orthologs, 18 are duplicates, 57 are fragmented and 14 are missing; out of 1210 complete genes
108 in the Embryophyte database 1173 are single-copy orthologs, 37 are duplicates, 105 are
109 fragmented and 125 are missing (Additional file 4: Figures S2A and S2B). The transcriptome
110 reads were aligned with the assembled genome and found that 99.46% of reads mapping and
111 88.25% of paired reads were mapped.

112 **Genome and pathways annotation**

113 We identified 35,093 protein-coding genes with the complete structures in the Peepal tree
114 genome (Additional file 5: Text file S1.1 and Additional file 6: Text file S1.2). RNA-seq data
115 from two leaf tissue samples of the Peepal tree and alternative reference ESTs from *Morus*
116 *notabilis* and *Arabidopsis thaliana* protein sequences were used as protein homology evidence
117 during genome annotation. Based on a homology search using BLASTN, out of 35,093 genes
118 predicted, 32,255 genes (91.9%) were having evidence from transcriptome assembly. About
119 76.3% of RNA-seq reads from the day and night leaf tissue samples were mapped to the
120 annotated genes in the Peepal tree.
121
122 Based on the sequence similarities, the complete set of annotated genes and their amino acid
123 sequences were used in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway

124 analysis [17]. This result showed the pathways in metabolism, biosynthesis of secondary
125 metabolites, genetic and environmental information processing, and signal transduction
126 pathway were common as several others. The top 5 highest gene count for pathways like
127 Ribosome (123 genes), Spliceosome (96 genes), Oxidative phosphorylation (86 genes),
128 Thermogenesis (82 genes), and RNA transport (74 genes) were found. As importantly
129 candidate genes were also found for human disease pathways like Huntington's disease (68
130 genes), Parkinson's disease (57 genes), Alzheimer's disease (55 genes), and others listed in
131 (Additional file 7: Table S2).
132
133 The plant model Arabidopsis encodes several orthologues of human proteins that function in
134 mechanisms similar to those in other eukaryotes [18] [19]. Previous findings showed that 70%
135 of oncogenes involved in cancer have Arabidopsis, 67% in *D. melanogaster*, 72% in *C.*
136 *elegans,* and 41% in *S. cerevisiae*. This 'disease gene' similarity is comparable to that observed
137 in other model organisms [20]. Arabidopsis and many other model systems research have led
138 to the discovery or analysis of genes and processes important to human health. Previous studies
139 showed that the Peepal tree has been tested for the treatment of neurodegenerative disorders
140 such as Parkinson's disease and Huntington's disease [6], [7], as well as anti-ulcer activity in
141 animal models. [8]. Another study showed that methanol extract of *F. religiosa* has proven
142 anti-inflammatory properties in LPS-induced activation of BV2 microglial cells, and it might
143 have therapeutic potential for various neurodegenerative diseases [21]. In the present work, we
144 found that Peepal genes show similarities with human disease pathways, which can be utilized
145 to further understand the traditional medicinal practices and Buddha's meditation practices.
146 Thus, plant research opens up new frontiers in terms of drug development and treatment of
147 diseases of great importance to human health. Plants seem to be part of this diverse portfolio
148 of tools necessary to understand fundamental cellular processes.

149 **Protein family and Gene Ontology analysis**

150 The protein family (Pfam) ID and Gene Ontology (GO terms) were assigned to genes using an
151 InterProScan module [22]. Out of 35,093 genes, 24,163 consisted of Pfam IDs that were
152 distributed across 3759 types of Pfam domains, and their gene ontology (GO) terms were also
153 identified. The Pfam domain consisting of proteins that were large in the Peepal tree genome
154 included 3-Deoxy-D-manno-octulosonic-acid transferase, Ring finger domain, PPR repeat
155 family, Helix-loop-helix DNA-binding protein, DYW family of nucleic acid deaminases,
156 Lysine methyltransferase, Putative GTPase activating protein for Arf, Ankyrin repeats and
157 others were listed in (Additional file 8: Table S3).

158 Catalase is an antioxidant enzyme known to catalyze H_2O_2 into water and oxygen. The Catalase
159 gene (FRLM_016351-RA) and its isozyme CAT1 Catalase isozyme 1 (FRLM_016350-RA),
160 (FRLM_012250-RA) structural gene were identified in the genome of the Peepal tree. Two
161 catalase genes were identified in differential expression data, KatE enciphers as a
162 monofunctional catalase and KatG enciphers as a catalase-peroxidase. KatE is known as CatB
163 expression in the day (FPKM – 937.49) and night (FPKM – 1786.02) transcript abundance.
164 KatG expression in the day (FPKM – 162.03) and night (FPKM – 81.53) transcript abundance.
165 KatE gene has been involved in physiological pathways like Glyoxylate and dicarboxylate
166 metabolism, Tryptophan metabolism, MAPK signaling pathway – plant, FoxO signaling
167 pathway, serine-pyruvate transaminase, and KatG Tryptophan metabolism, Tyrosine
168 metabolism, Biosynthesis of secondary metabolism and Drug metabolism.

169 **Identification of homologous, orthologous, and singleton genes**

170 To understand the homologous genes and the orthologous gene evolution, and relationships
171 across the species, we performed the homologous and orthologous gene detection for the

172 Peepal tree with the other 5 species. Homologous gene identification and orthologous
173 clustering of the proteomes of six species, including the model organism *A. thaliana*, *M.*
174 *notabilis* closest relative species of Ficus, and other closely related species of the Moraceae
175 family were selected for the analysis. The number of homologous genes identified for *F.*
176 *religiosa* proteome sequences (35,093) was 29516 in *Arabidopsis thaliana*, 29924 in *Morus*
177 *notabilis*, 29750 in *Cannabis sativa*, 29909 in *Prunus persica*, and 29830 in *Ziziphus jujuba*
178 based on homology analysis. *F. religiosa, A. thaliana*, *M. notabilis, C. sativa, P. persica*, and
179 *Z. jujuba* form a cluster of 24,310 orthologous genes and which are conserved within the
180 species. The number of specific orthologous gene clusters in *F. religiosa, A. thaliana, M.*
181 *notabilis, C. sativa, P. persica*, and *Z. jujuba* were 15016, 16170, 16517, 15921, 16655, and
182 17235 respectively. However, 1184 single-copy gene clusters were found across the six species
183 and specific singletons for *F. religiosa, A. thaliana, M. notabilis, C. sativa, P. persica,* and *Z.*
184 *jujuba* were 10154, 4469, 2284, 1912, 1802, 4209, respectively (Figure 1).

185 The identified single-copy clusters were able to suggest the taxonomic and phylogenetic
186 relationships among a group of species. Based on the similarity of proteomes and single-copy
187 orthologous clustering, we deduced the phylogenetic tree for *F. religiosa* and the other five
188 species. A multiple sequence alignment (MSA) and Neighbour-Joining (NJ) method were used
189 for constructing an evolutionary phylogenetic tree. It was found that *F. religiosa* is closely
190 related to *M. notabilis* by having more similarities in their proteome as they are evolving from
191 the Moraceae family, followed by *C. sativa, Z. jujuba, P. persica,* and *A. thaliana* (Additional
192 file 9: Figure S3).

193 **Comparative analysis of Peepal tree genome**

194 We aligned the genome of the Peepal tree with the three Moraceae family members like *F.*
195 *carica*, *F. microcarpa*, and *M. notabilis*. Comparison of our assembly against these genomes

196 resulted in a mapping of 88.62% to *F. carica*, 89.6% to *F. microcarpa*, and 46.9% to *M.*
197 *notabilis*. The results showed the similarity of the Peepal genome is closer to the genus Ficus
198 (*F. carica* and *F. microcarpa*) and also relatively closer to the genus Morus of the same family.
199 The statistics of genome sequence alignment of *F. religiosa* against *F. carica*, *M. notabilis*, *F.*
200 *microcarpa* genomes (Additional file 10: Text file S2).

201 **Repeats in the genome of the Peepal tree**

202 Repeat library building and repeat identification were performed using the ReapeatModeller
203 and RepeatMasker tools (www.repeatmasker.org) respectively. *De novo* repeat identification
204 resulted in 53.55% (269.62Mb) repetitive sequences in the Peepal tree genome. The RNA
205 elements, long terminal repeats (LTR) high constitute about 5% of repeats and 43.71% of the
206 repeats did not belong to any of the annotated repeats families. The 53.55% of repetitive
207 sequences in the Peepal tree genome are closest to its Moraceae family species, 47% are found
208 in the closest species mulberry (*M. notabilis*), 46.5% in *F. microcarpa*, and 48.9% in *F. hispida*.
209 The repetitive sequences were classified into known categories, such as LINE1 (0.19%), long
210 terminal repeat retrotransposon (5.09%), DNA transposons (1.09%), and simple repeats
211 (3.25%) and unclassified (43.71%) (Additional file 11: Table S4).

212 **Simple sequence repeats (SSRs)**

213 We predicted SSRs from the assembled Peepal tree genome. In total, 7,99,992 SSRs were
214 identified on 2,67,593 sequences, which are composed of mono- (6,06,169), di- (1,43,113),
215 tri- (34,327), tetra- (11,791), penta- (2,911), and hexa- (1,681) type repeats (Additional file 12:
216 Table S5.1). Among mono repeats, the 'A/T' (73.91%) type was the highest followed by 'C/G'
217 (1.87%). Similarly, the 'AT /TA', 'AG/CT', 'AC/GT', and 'CG/CG' types of di repeats were
218 in 9.8%, 2.76%, 1.41%, and 0.09% fractions, respectively. 'AAT/ATT', 'AAG/CTT'
219 'ATA/TAT', 'TTA/TTC', and 'GAA/TAA', were the most abundant tri repeats and 'AAAT'

220 was most predominant in tetra repeats. The detailed distribution of all types of repeats and
221 statistics is shown in (Additional file 13: Table S5.2 and Additional file 14: Table S5.3).

222 **Transcription Factors (TFs)**

223 Transcription factors act in regulating gene expression due to several external and internal
224 signals by activating or suppressing downstream genes. The MAKER annotated protein
225 sequences of Peepal tree genome assembly were used for BLAST analysis with the Plant
226 Transcription Factor Database v5.0 [23] using the *A. thaliana* protein sequence as a reference.
227 A total of 1,264 protein sequences of genome annotated 35,093 protein-coding genes have
228 evidence for 56 families of Transcription factors (TFs) (Additional file 15: Table S6). The TFs
229 families include the abundant ERF, M-type MADS, ARF, DBB, MIKC MADS, WOX, C3H,
230 G2-like, MYB, TALE, B3, HB-other and MYB-related family proteins.

231 **Transcriptome sequencing, assembly, and annotation**

232 *De novo* transcriptome assembly was performed for the mature leaf samples of the Peepal tree
233 collected during the day and night periods. The assembly was performed for each sample and
234 also combined assembly for the reads of both samples. The combined transcriptome assembly
235 was performed and contributed to 152.8-Mb assembled bases with the N50 length of 2076 bp
236 and an average transcript length of 13.16 Kb and 42.17% GC content. The transcriptome
237 assembly and annotation workflow are given in (Additional file 16: Figure S4). The statistics
238 of assembly contigs and sequence assembled contigs are shown in (Additional file 17: Table
239 S7).

240 The *de novo* assembled transcript (1,16,038) sequences were processed for annotation. *De novo*
241 assembled transcripts were clustered to exclude the redundant transcripts and identified 26,479
242 unique transcripts sequences. The statistics of unigenes are given in Table 2. *De novo*

243 assembled transcripts and unigenes were annotated to find the structural and functional genes.
244 The protein families were identified for the uniquely characterized transcripts of RNA data.
245 Out of 26,691 transcripts, Pfam IDs for 19175 were distributed across 3977 types of Pfam
246 domains and their gene ontology terms were also identified (Additional file 18: Table S8).

247 The differential expressed genes (35,182) from the day and night periods leaf tissue samples
248 of the Peepal tree are given in (Additional file 19: Table S9.1) and these transcripts were used
249 for the pathway analysis. The top 272 highly up-regulated differential transcripts were given
250 for diurnal and nocturnal periods (Additional file 20: Table S9.2).

251 The transcription factors (TFs) were identified for the differentially expressed transcripts. From
252 the day sample, the 2 transcripts coded for specific TFs like C3H family protein and nuclear
253 factor Y, subunit A7 (NF-YA7), and in the night period sample, the 6 transcripts coded for
254 specific TFs like ERF family protein, CONSTANS-like 2, MYB-related family protein
255 (Additional file 21: Table S9.3).

256 **Non-coding RNA genes in the Peepal tree genome**

257 Based on a coding potential calculator (CPC), *de novo*-based assembled transcripts (26,691)
258 were further categorized into protein-coding (19,911) and non-coding (6,780). Based on
259 BLASTN analysis, out of 6,780 non-coding transcripts, 4219 transcripts targeted genomic
260 annotated genes and 2,561 remained as non-coding. A total of 30,973 Cufflinks assembled
261 transcripts (reference-based alignment with genome assembly) were further categorized into
262 protein coding (7,163) and non-coding (23,810). Out of 23,810 non-coding transcripts, 14605
263 were having alignment to genome annotated genes using BLASTN. Further, categorization of
264 specific day and night sample transcripts resulted in protein-coding 6,628 and 7,339 and non-
265 coding of 20,528 and 25,494 transcripts respectively. From those non-coding transcripts of day

266 and night, 18,893 and 22,232 transcripts got aligned MAKER-P predicted genes using
267 BLASTN. The remaining transcripts were considered to be non-coding transcripts, as we did
268 not find any match to predicted gene evidence to support them.

269 **miRNAs**: miRNAs are a major class of non-coding RNAs. Based on the homology search, we
270 have identified the microRNA precursors using the miRbase database (http://www.mirbase.org).
271 These microRNAs belong to MIR396, MIR2916, MIR156, MIR164, MIR6236, MIR166,
272 MIR168, and MIR395 families. Among the identified miRNAs, MIR408 was found specific to
273 the night period transcripts of the Peepal tree. On the genome, MIR 408 was found on the genes
274 like TPK5 Two-pore potassium channel 5, prfA peptide chain release factor 1, and also on
275 proteins of unknown function. MIR408 is a highly conserved microRNA in plants. miR408 has
276 enriched photosynthesis by mitigating the efficiency of irradiation utilization and the capacity
277 for carbon dioxide fixation [24].
278
279 In the *de novo*-based transcriptome data, the non-redundant unigenes were used to identify the
280 microRNAs MIR168 and MIR166, and their homologs were found on two of the transcripts.
281 We identified the miRNAs on genomic scaffolds based on mapping the transcriptome data to
282 the genome. This provides information on miRNAs specific to the day and night leaf tissue
283 transcriptome. We identified 23 and 25 pre-miRNA expressions in the day and night period
284 respectively (Additional file 22: Table S10.1). The statistics of transfer RNAs (tRNA) were
285 identified and their details are given in the (Additional file 23: Table S10.2).
286
287 **Elucidation of carbon fixation pathway in Peepal tree**
288 The study was conducted to analyze the gene expression patterns in the leaf tissues of the
289 Peepal tree under two conditions the diurnal (2 PM) and the nocturnal period (2 AM). Through
290 the pathway analysis, the candidate genes for carbon fixation pathways like Crassulacean Acid

291 Metabolism (CAM) pathway, Calvin-Benson cycle (C3) pathway, and C4 - Dicarboxylic
292 pathway were identified and estimated based on their transcript abundance. The transcriptome
293 data contained the 20 putative genes involved in the carbon fixation module of CAM, C3, and
294 C4 including for the key fructose-bisphosphate aldolase class I, fructose-1,6-bisphosphate,
295 phosphoenolpyruvate carboxylase (PEPC/PPC), phosphoenolpyruvate carboxylase kinase
296 (PPCK), NAD+ and NADP+, malate dehydrogenase (MDH), pyruvate orthophosphate
297 dikinase (PPDK) (Additional file 24: Table S11). The complete pathway map was found for
298 CAM and C4 cycles and three genes were missed from the C3 cycle. The three genes fructose-
299 6-phosphate phosphoketolase (EC 4.1.2.22) and phosphoketolase (EC 4.1.2.9) were purified in
300 *Acetobacter xylinum* [25] and sedoheptulokinase (EC 2.7.1.14) was showed in *Bacillus* species
301 [26]. These three genes were not found in the Peepal tree for the C3 cycle.

302 The differentially expressed genes from transcriptomic data were mapped to the reference
303 carbon fixation in photosynthetic organisms pathway on the KEGG database are provided in
304 (Additional file 25: Figure S5) [27], [28], and [29]. The diagrammatic representation of the
305 genes involved in the carbon fixation pathway is shown in (Additional file 26: Figure S6).

306 Ribulose-bisphosphate carboxylase (RuBP carboxylase or *rubisco*) small chain enzyme is
307 enriched in leaf tissue collected during the day (2 PM). Rubisco is the most abundant protein
308 in chloroplasts. The glyceraldehyde-3-phosphate dehydrogenase (NADP+) is enriched in day
309 sample leaf tissue, the enzyme responsible for the reversible conversion of glyceraldehyde 3-
310 phosphate to ribulose bisphosphate using ATP, the acceptor for CO2 and completes the Calvin-
311 Benson Cycle (C3 cycle) [30].
312

313 The signature genes responsible for the CAM cycle were expressed in the Peepal tree during
314 the night. The phosphoenolpyruvate carboxylase kinase (PPCK), NAD(P)-ME (maeB), and

315 Malate dehydrogenase (MDH) genes were highly enriched in the photosynthetic leaf tissue
316 collected during the night period than the day. It indicates that the Peepal tree adapts to the
317 CAM pathway and can fix nocturnal carbon dioxide using the PEP carboxylase (PEPC) enzyme
318 and accumulate malate by the enzyme malate dehydrogenase in the large vacuoles of their cells.
319 The transcriptomic genes from the Peepal tree were mapped completely to the KEGG pathway
320 of the CAM cycle.
321
322 In the C4 Dicarboxylic cycle, the high expression of glutamate-glyoxylate aminotransferase
323 enzyme (GGAT) in the leaf tissue collected during the night period (2 AM) indicates the
324 photorespiration in the Peepal tree. The carbon fixation begins in the mesophyll cells, where
325 CO2 is converted into bicarbonate. It adds the 3-carbon acid phosphoenolpyruvate (PEP) by
326 an enzyme called phosphoenolpyruvate carboxylase. The product of this reaction is the four-
327 carbon acid oxaloacetate, which is reduced to malate another four-carbon acid [31]. The second
328 highest expression is NADP-malate dehydrogenase (MDH), which converts the oxaloacetate
329 generated by PEPC to malate. The genes from Peepal tree differential expression data had a
330 complete map to the C4 cycle. The gene expression pattern of the carbon-fixation pathway in
331 the Peepal tree suggests that the plant switches between the C3, C4, and CAM cycles during
332 the diurnal and nocturnal periods. The Fragments Per Kilobase of transcript per Million
333 mapped reads (FPKM) and Trimmed Mean of M-values (TMM) values of the differential genes
334 expressed for carbon fixation pathway are shown in Figures 2A, B, C, and 3A, B, and C.

335 **Discussion**

336 In this study, the genomics and transcriptomics data were generated and annotated for the
337 keystone species Peepal tree (*F. religiosa*). This plant species is well known for Buddha's
338 enlightenment while he was meditating underneath this tree. This tree is worshipped as birth-

339 giving, regenerative, and medicinally valued for many diseases, and this culture is spread
340 across Asia. We used two Next-Generation Sequencing technologies to sequence the genome
341 of the Peepal tree and described the hybrid whole genome. The photosynthetic tissues of
342 distinct conditions (diurnal and nocturnal) were used for transcriptome sequencing and to
343 understand the genes, proteins, and molecular pathways. The assembled genome resulted in a
344 size of 406-Mb with 35,093 protein-coding genes, based on *ab initio*, homology-based, and
345 mRNA evidence used for annotation. The transcriptome analysis resulted in 26,691 unique
346 transcripts. The completeness of the Peepal tree genome assembly was confirmed based on
347 BUSCO analysis and comparative analysis of transcriptomic data. A total of 35,093 genes were
348 predicted, and 32,255 genes (91.9%) were having evidence from transcriptome assembly.
349 About 76.3% of RNA-seq reads were mapped to the annotated genes. This information has
350 provided confidence for the downstream analyses. The number of protein-coding genes
351 identified in the Peepal tree is related to other plant genomes sequenced over the decades.
352 Similarly, in our laboratory *Santalum album* (Sandalwood) genome was sequenced with the
353 prediction of 38,119 protein-coding genes based on mRNA and peptide evidence [32].

354 We performed the downstream analysis of genomic and transcriptomic data to understand
355 molecular changes in the diurnal and nocturnal leaf tissue samples of the Peepal tree. Analysis
356 of microRNAs, TFs, and carbon fixation pathways has led to the understanding of
357 photosynthesis in the Peepal tree. The long non-coding RNAs have important roles in the
358 metabolic, physiological, and biological processes of the cell. MIR408 has been reported to
359 target various blue copper protein members and phytocyanin family [33], [34]. In Arabidopsis,
360 miR408 responds to copper deficiency and light [35]. In *Oryza sativa* MiR408 plants were
361 efficient at saving and converting light energy into sugars, suggesting that miR408 might
362 promote grain yield by regulating the phytocyanin (PC) content and photosynthesis by down-
363 regulating UCL8 uclacyanin (UCL) gene of the phytocyanin family. So, MIR408 found

364 specific expression in transcripts of night period leaf tissue of the Peepal tree indicating the
365 similar conversion of light energy and accumulation of sugars in the plant during the dark. It
366 also aids in photosynthesis by enhancing carbon fixation.

367 Transcription factors like C3H family proteins were identified by their motif of three cysteines
368 and one histidine residue. They play an important role in regulating growth, developmental
369 processes, and environmental responses in the plant's [36]. In plants, the nuclear factor-YA has
370 a role in drought stress responses. In rice, NF-YA7 is involved in the drought tolerance pathway
371 which is independent of the Abscisic acid manner [37]. Expression of C3H and NF during the
372 day could influence plant growth and development in the Peepal tree. The Ethylene response
373 factor ERF105 showed the cold-regulated transcription factor gene of Arabidopsis and also
374 functions in the freezing tolerance and cold acclimation [38]. In the Peepal tree, ERF, MYB -
375 related family proteins like REVEILLE 1 (RVE1) and late elongated hypocotyl gene (LHY)
376 are expressed during the night period. RVE1 functions in the circadian clock and auxin
377 pathways [39] and LHY maintains the circadian rhythm in Arabidopsis [40]. Both the RVE1
378 and LHY are found expressed in night-specific Peepal tree transcripts indicating the active
379 circadian rhythms and pathways during the dark time.

380 In the current study, Catalase gene expression was found high in the night period transcript
381 abundance of Peepal tree leaf tissue. A previous study on the Peepal tree showed that leaf tissue
382 collected at night time exhibited the scotoactive opening of stomata during the night, which
383 indicates that through the stomatal opening molecular oxygen (O_2) is released by the action of
384 catalase enzyme on hydrogen peroxide (H_2O_2) [41]. Peepal tree plants grown in adverse
385 habitats showed 55% higher H_2O_2 production with about a 30% increase in the peroxidase
386 activity [41]. The physiological interaction between catalase and its substrate H_2O_2 in the plant
387 was determined by quantifying H_2O_2 and assaying the catalase, in which catalase showed a 4-

388 fold increase in activity, especially during the night. Peepal tree has a higher amount of H_2O_2
389 deposition during the night than day, indicating pathway switching between carbon fixation
390 pathways. The transcriptomic sequencing from time-structure diurnal (2 PM) and nocturnal (2
391 AM) leaf samples showed the gene expression patterns of the carbon fixation pathway. The
392 gene expression of mRNA in the C3, C4, and CAM cycles indicated that depending on the
393 carbohydrate, amino acids biosynthesis, metabolism, and environmental conditions the plant
394 switches between these three cycles in a time-structured manner.

395 The enzymatic steps involved in the CAM cycle are similar to the C4 plant's [42]. Plants adapt
396 to the CAM cycle to grow during water constraints and increase the level of carbon dioxide
397 uptake than their C3 and C4 cycles [43]. The photorespiration in C3 plants dissipates more than
398 25% of the carbon fixed through the photosynthesis [44]. The day mRNA data showed that the
399 Peepal tree may carry out the diurnal carbon fixation by the C3 cycle. GGAT plays an important
400 role in the biosynthesis and metabolism of major amino acids. GGAT is involved in the
401 photorespiratory process. It catalyzes the reaction of glutamate and glyoxylate in the 2-
402 oxoglutarate and glycine [45]. High expression of GGAT in the C_4 cycle indicates that there
403 could be photorespiration in the Peepal tree during the night.
404
405 The Peepal tree study provides information on plants using the CAM pathway to fix nocturnal
406 carbon dioxide using the PEP carboxylase (PEPC) enzyme and the accumulation of malate by
407 the enzyme malate dehydrogenase in the large vacuoles of their cells. The essential CAM genes
408 in C3 species expressed in periods suggested that ancestral expression patterns required for
409 CAM may have pre-dated its origin in *Yucca* [46]. In the *Kalanchoë fedtschenkoi* genome
410 study, the convergence in protein sequence and re-scheduling of diel transcript expression of
411 genes was reported to be involved in nocturnal CO_2 fixation, stomatal movement, heat

412 tolerance, circadian clock, and carbohydrate metabolism with the other CAM species in
413 comparison with non-CAM species [14].
414 Some of the previous studies in the pineapple genome revealed the gene lineage transition from
415 C3 photosynthesis to CAM, and CAM-related genes exhibit a diel expression pattern in
416 photosynthetic tissues [47]. The evolution of CAM in *Agave* from C3 photosynthesis shows
417 that the core metabolic components required for CAM have ancient genomic origins which
418 could be traceable to non-vascular plants while regulatory proteins required for diel re-
419 programming of metabolism have shared among the recent origin of C3, C4, and CAM species
420 [48].
421
422 In summary, the genome data and transcript abundance evidence indicate the molecular switch
423 in the carbon fixation pathway of the Peepal tree (*F. religiosa*) during the day and night periods
424 depending on its physiological and environmental conditions. Our study is a foundation for
425 further experiments to determine the underlying mechanisms in C3, C4, and CAM metabolism.

426 **Conclusions**

427 In this study, we generated the genomic and transcriptomic data for Peepal/Bodhi tree.
428 Genomic data pathway analyses identified the genes associated with several physiological,
429 biochemical, metabolic, and disease pathways. Differential expression data from diurnal and
430 nocturnal leaf tissue samples of Peepal revealed gene expression patterns in the carbon fixation
431 pathway during light and dark. The transcript abundance indicates that plants could switch
432 between the three C3, C4, and CAM pathways. The well-annotated genome for the Peepal tree
433 will have broader implications for studies regarding the physiology, evolution, conservation of
434 species, and human neurological diseases.
435

436 **Methods**

437 **Collection of leaf samples and extraction of nucleic acids**

438 The mature leaves were collected from a cultivated Peepal tree (15 years old) at a Private
439 property, Anuganalu village, Hassan District, India (13.0647° N, 76.0363° E). We have
440 followed a non-invasive method for collecting leaf samples. Genomic DNA was extracted from
441 the leaves using the Qiagen DNeasy Plant Mini kit (Catalog #69106), and the quality and
442 quantity of DNA were confirmed using the Nanodrop. From the same Peepal tree, during the
443 day (2 PM) and night (2 AM) periods, the leaf samples were collected and immediately placed
444 on dry ice. Total RNA was isolated from the leaf samples using the QiagenRNeasy Plant Mini
445 kit (Catalog #74904) method and was treated with RNase-free DNase I (Catalog #M0303S)
446 from New England BioLabs for 30 min at 37 °C to remove residual DNA. RNA integrity and
447 quantity were confirmed on Qubit and Tape station using a dsDNA HS (Catalog #32854) kit
448 from Invitrogen and RNA screen tape from Agilent respectively.

449 **DNA and RNA library preparation and sequencing**

450 Whole-genome shotgun DNA library preparation was performed using the Illumina TrueSeq
451 DNA sample preparation kit (FC-121-2001). The paired-end (PE) (2 x 100 nts) sequencing was
452 carried out using Illumina HiSeq 1000 at the Next Generation Genomics Facility at the Centre
453 for Cellular and Molecular Platforms (C-CAMP), Bengaluru. Also, to increase the size of
454 genome data, we sequenced the genome with paired-end (PE) (2 x 100 nts) using the MGISEQ-
455 2000 platform at Bengaluru Genomics Center (BGC) Pvt. Ltd, Bengaluru.

456 The RNA libraries were prepared using "TruSeq RNA Library Prep Kit v2 from Illumina®"
457 with Illumina standardized protocol. The RNA libraries were quantified on Qubit (dsDNA HS
458 kit) and validated on the TapeStation instrument (D1000 screen tape). These RNA libraries

459 were used for sequencing with the Illumina platform at the Bengaluru Genomics Centre (BGC)
460 Pvt. Ltd, Bengaluru.

461 **Genome size estimation and assembly**

462 Each of the Illumina and MGISEQ-2000 raw reads were processed for a quality check using
463 the FastQC v0.11.6 tool [8]. Then filtering and trimming of raw reads were done to remove the
464 low complexity bases using the TrimGalore-0.4.5
465 (https://www.bioinformatics.babraham.ac.uk/projects/trimgalore/) and reads having quality
466 value Q>20 and length above 20 bases were taken for constructing the assembly. To estimate
467 the genome size, filtered reads were taken for the k-mer distribution (different k-mers from
468 21to 77) and abundance analysis using Jellyfish v1.1.12 [49] and GenomeScope v2.0 [50]. The
469 separate Illumina and MGI Seq generated raw reads were used to construct the assembly using
470 the tools SPAdes-3.13.0 [51] and MaSuRCA-3.2.9 [52] respectively. The parameters were the
471 default k-mer sizes of 21, 33, and 55 for Illumina assembly. The constructed assemblies were
472 used to build the super scaffolds using the tool SSPACE standard v3.0 [53].

473 The combined Illumina HiSeq and MGISEQ raw reads were used to construct the hybrid
474 assembly using the assembler SPAdes-v3.13.0 [51]. The parameters were the default k-mer
475 sizes 21, 33, and 55, with a 77 mer also set. The gaps in the assembly were closed by GMcloser-
476 1.6.2. [54]. The assembly statistics were obtained using the tool Quast v4.6.1 [55]. The
477 completeness and evaluation of the assembly were done by BUSCOv3 tool [56] with the
478 Embryophyte and Eukaryota database and by aligning the RNA-seq reads to the genome.

479 **Structural gene prediction and functional annotation:**

480 Peepal tree assembled scaffolds were processed for structural and functional gene annotation
481 using the MAKER-P v.2.31.10 software [57]. The RNA-sequenced data of *Morus notabilis* [9]

482 consists of expression sequence tags (ESTs) and the GFF (Gene finding format) file which
483 contains the gene features and structures of genes, protein data of *A. thaliana* and RNA-
484 sequence data of Peepal tree were imported as evidence for annotation support. The structural
485 and functional annotation of predicted genes and proteins was performed using BLASTP in the
486 Uniprot database. The protein family, structures, and gene ontology (GO) terms were identified
487 for protein-coding genes using InterProScan-V5.27-66.0 [22].

488 **Gene family construction, identification of homologous and orthologous genes**

489 Protein sequences of *A. thaliana*, *M. notabilis*, *P. persica*, *C. sativa*, *Z. jujuba,* and the protein
490 sequences of the current study *F. religiosa* were taken for the homologous and orthologous
491 gene identification. The homologous genes were identified in *F. religiosa* proteome sequences
492 using the BLASTP analysis against the other 5 proteomes of *A. thaliana*, *M. notabilis*, *P.*
493 *persica*, *C. sativa*, and *Z. jujuba*. OrthoVenn2 [58] was used for the clustering of orthologous
494 genes and the identification of the single-copy orthologous genes in all six proteomes. Further,
495 these single-copy orthologous genes were used for constructing the phylogenetic tree using the
496 tool MAFFT-v7 [59].

497 **Comparative genome analysis**

498 We downloaded the genomes of *F. carica*, *F. microcarpa*, and *M. notabilis*. We aligned these
499 genomes against the Peepal tree genome assembly to understand their relationships using the
500 BWA-V0.7.17 (Burrows-Wheeler Aligner) [60] and Samtools v1.7 [61].

501 **Prediction of repetitive elements: TEs and SSR**

502 The RepeatModeller-open-1.0.11and RepeatMasker-4.0 tools were used for repeat library
503 building and repeat identification in the assembly respectively. The MicroSAtellite

504 identification tool (MISA) [62] was used for the identification of SSRs from assembled genome
505 sequences of *F. religiosa*. The parameters were set to identify perfect di-, tri-, tetra-, penta-,
506 and hexa nucleotide motifs with a minimum threshold of 6, 5, 5, 5, and 5 repeats, respectively.

507 **Prediction of transcription factor families**

508 The families of transcription factors (TFs) were predicted in genome annotations and
509 differentially expressed transcripts of the Peepal tree using Plant Transcription Factor Database
510 v5.0 [23].

511 **Non-coding RNA genes**

512 The transfer RNAs in the Peepal tree genome were found using tRNAscan-SE (v2.0.3) [63]
513 with the 'eukaryotes' option. tRNAscan-SE software deployed with the covariance models
514 identifies the primary sequence and secondary structure information of tRNA and gives the
515 complete tRNA genes for the query genome and transcriptome sequences. tRNAscan-SE
516 software is integrated with Infernal v1.1 to enhance the tRNA search with better covariance
517 and other updated models. Using the isotype-specific covariance model provides the functional
518 classification of tRNAs and in the first pass search cutoff score, 10 is set. The miRbase database
519 (http://www.mirbase.org) was used for the identification of putative miRNAs in the genome
520 and unique identified transcripts sequence data based on the homology search. The long non-
521 coding RNAs (lncRNAs) were identified with the Coding Potential Calculator tools [64].

522 **Transcriptome sequencing, assembly, and annotation**

523 High-quality stranded RNA sequencing (ssRNA-seq) reads were assembled into putative
524 transcripts using Trinity v2.9.0 (Grabherr et al., 2011). Assembled transcripts were passed
525 through Transdecoder v5.02 to predict the coding sequences. The transcripts were clustered to

526 find the unigenes by removing the redundant transcripts using the tool CD-HIT-est v.0.0.1 [65]
527 with a 95% sequence identity threshold. Transcripts assembled from Trinity and CD-HIT-
528 v0.0.1 were used in downstream analyses for gene prediction. Unigenes were used to predict
529 the putative genes using the NCBI non-redundant (nr) database using the BLASTX program
530 and proteins were predicted from the Uniprot database using the BLASTP program. The Trinity
531 assembled transcripts were annotated using Trinotate- V3.11. The raw reads were mapped to
532 scaffold assembled genome using Cufflinks-v2.2.1 [66] and considered as reference assembly.

533 **Transcript quantification and differential gene expression analysis**

534 The estimation of transcripts abundance was determined using RNA-Seq by Expectation-
535 Maximization (RSEM) tool [67], which quantifies transcript level abundance from RNA-seq
536 data. RSEM first generates and pre-processes a set of reference transcript sequences and then
537 aligns reads to reference transcripts followed by an estimation of transcript abundances.
538 Normalized transcripts obtained from the transcript quantification methods were used in the
539 next step for analyzing the differential gene expression of genes. FPKM and Trimmed Mean
540 of M-values (TMM) are calculated to understand the expression levels of genes in day and
541 night samples of the Peepal tree. For further analysis, the gene expression was estimated using
542 FPKM and TMM value minimum ≥1. The TMM value was used to cluster the genes according
543 to their expression pattern using the edgeR package in the R tool. The parameters used in the
544 differential analysis were a probability value P-value of 0.001 and a fold change value of log2.
545 The expression value was also determined for assembled transcripts to verify the expression of
546 genes predicted from gene models. The differentially expressed genes were annotated using
547 BLAST2GO Annotation software [68].

548 **Pathway Analysis**

549 The annotated genes from the assembled genome and the differentially expressed genes from
550 the Peepal tree leaf tissues collected during the day (2.00 PM) and night (2.00 AM) were used
551 for pathway analysis in the KAAS (KEGG Automatic Annotation Server) (KEGG) server [17]
552 using the BBH (bi-directional best hit) method and the search against a default set of 40
553 eukaryotic organisms. It provided the list of pathways where the candidate genes were mapped
554 based on the orthologous homology alignment.

555 **Abbreviations**

556 BUSCO: Benchmarking Universal Single-Copy Orthologous; ESts: Expressed Sequence Tags;
557 PPR repeat: Pentatricopeptide repeat; FPKM: Fragments Per Kilobase of transcript per Million
558 mapped reads; MSA: Multiple Sequence Alignment

559 **Statement:**

560 • **Experimental Plant details:** *Ficus religiosa* a cultivated plant was used in this study.
561 The mature leaves were collected from this plant at Private property, Anuganalu
562 village, Hassan District, India (13.0647° N, 76.0363° E) were kindly provided by
563 Prof. Malali Gowda, DNA Life Foundation, Anuganalu village, Hassan District,
564 India (13.0647° N, 76.0363° E).
565 • The formal identification of the plant was undertaken by Prof. Malali Gowda, who
566 owns the place of plant location and is also a corresponding author of this study.
567 Hence, we had his permission to collect the plant material. The specimen of this
568 material has not been deposited in any publicly available herbarium.

569

570 **Declaration**

571 **Ethics approval and consent to participate**

572 The study is compiled with relevant institutional, national, and international guidelines and
573 legislation.

574

575 **Consent for publication**

576 Not Applicable (NA)

577

578 **Availability of data and materials**

579 The raw sequence reads have been deposited under NCBI Sequence Read Archive (SRA)
580 accession numbers SRR7244210 for Illumina sequenced *F. religiosa* WGS, SRR13827064 for
581 MGISEQ sequenced *F. religiosa* WGS –
582 (https://trace.ncbi.nlm.nih.gov/Traces/study1/?acc=PRJNA474013), SRR7343291
583 (https://trace.ncbi.nlm.nih.gov/Traces/study1/?acc=SRR7343291) for *F. religiosa*
584 transcriptome). The Whole Genome Shotgun project has been deposited at
585 DDBJ/ENA/GenBank under the accession JAFMPE000000000
586 (https://www.ncbi.nlm.nih.gov/nuccore/JAFMPE000000000.1/) and Transcriptome Shotgun
587 Assembly project has been deposited at DDBJ/ENA/GenBank under the accession
588 GJAV00000000 (https://www.ncbi.nlm.nih.gov/nuccore/GJAV00000000.1/).
589 Other published data used in the study: The protein sequences of *Arabidopsis thaliana*
590 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/735/GCF_000001735.4_TAIR10.1/G
591 CF_000001735.4_TAIR10.1_protein.faa.gz), *Morus notabilis*
592 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/414/095/GCF_000414095.1_ASM41409v
593 2/GCF_000414095.1_ASM41409v2_protein.faa.gz), *Prunus persica*
594 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/346/465/GCF_000346465.2_Prunus_pers

595 ica_NCBIv2/GCF_000346465.2_Prunus_persica_NCBIv2_protein.faa.gz), *Cannabis sativa*
596 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/900/626/175/GCF_900626175.2_cs10/GCF_9
597 00626175.2_cs10_protein.faa.gz), *Ziziphus jujuba*
598 (https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/020/796/205/GCF_020796205.1_ASM207962
599 0v1/GCF_020796205.1_ASM2079620v1_protein.faa.gz).
600

601 **Competing interests**

602 The authors declare that they have no competing interests.
603

604 **Funding**

605 Not Applicable (NA)
606

607 **Author's contribution**

608 AKL performed the DNA and RNA isolation from leaf tissues, genome assembly and
609 functional annotation, gene prediction, repeat prediction, orthologous gene clustering, Genome
610 and Transcriptome pathway analysis, submitted WGS and RNA-seq data to NCBI, prepared
611 the genomic and transcriptomic study tables and figures, wrote the manuscript; MG, AKL, and
612 AKP designed the genomic and transcriptomic experiments of *Ficus religiosa*; MG conceived
613 and conceptualized the project; reviewed and edited the manuscript.

614 **Acknowledgment**

615 We would like to thank Tata Education and Development Trust fellowship (01/2018-2019)
616 provided for Ashalatha K L for her Ph.D. program. We acknowledge the Next Generation
617 Genomics Facility at the Centre for Cellular and Molecular Platforms (C-CAMP) and

Bengaluru Genomics Center Pvt. Ltd for supporting sequencing and analysis. We are grateful to the MGI team, in Hong Kong for sequencing the whole genome.

Author's Information:

[1]K L Ashalatha: ashalatha708@gmail.com, ashalatha.kl@tdu.edu.in

[2]K P Arunkumar: arunkumar.kallare@gmail.com, arunkallare.csb@gov.in

[*1,3]Malali Gowda: malali.gene@gmail.com

[1] The University of Trans-Disciplinary Health Sciences and Technology, Yelahanka, Bengaluru, 560064, India

[2] Central Muga Eri Research and Training Institute (CMER&TI), Central Silk Board, Ministry of Textiles Lahdoigarh, Jorhat, Assam, 785700, India

[3] DNA Life Organization, Yelahanka, Bengaluru, 560064, India

References

1. IPCN Chromosome Reports, 2014. Index to Plant Chromosome Numbers (IPCN), Tropicos website. St. Louis, Missouri, USA: Missouri Botanical Garden.
2. Van Noort S, Gardiner AJ, Tolley KA. New records of Ficus (Moraceae) species emphasize the conservation significance of inselbergs in Mozambique. South African J Bot. 2007;73:642–9.
3. Ashalatha KL, Gowda M. Heritage of Neem–Peepal Tree Resides a Profound Scientific Facts. 2019. p. 13–9.
4. Chantarasuwan B, Baas P, van Heuven B-J, Baider C, van Welzen PC. Leaf anatomy of Ficus subsection U rostigma (Moraceae). Bot J Linn Soc. 2014;175:259–81.
5. Singh D, Singh B, Goel RK. Traditional uses, phytochemistry and pharmacology of Ficus religiosa: A review. J Ethnopharmacol. 2011;134:565–83.
6. Bhangale JO, Acharya SR. Anti-Parkinson Activity of Petroleum Ether Extract of Ficus religiosa (L.) Leaves. Adv Pharmacol Sci. 2016;2016:1–9.
7. Bhangale JO, Acharya NS, Acharya SR. Protective effect of Ficus religiosa (L.) against 3-nitropropionic acid induced Huntington disease. Orient Pharm Exp Med. 2016;16:165–74.
8. Gregory M, Divya B, Mary RA, Viji MMH, Kalaichelvan VK, Palanivel V. Anti–ulcer activity of Ficus religiosa leaf ethanolic extract. Asian Pac J Trop Biomed. 2013;3:554–6.

9. He N, Zhang C, Qi X, Zhao S, Tao Y, Yang G, et al. Draft genome sequence of the mulberry tree Morus notabilis. Nat Commun. 2013;4:2445.
10. Usai G, Mascagni F, Giordani T, Vangelisti A, Bosi E, Zuccolo A, et al. Epigenetic patterns within the haplotype phased fig (Ficus carica L.) genome. Plant J. 2020;102:600–14.
11. Zhang X, Wang G, Zhang S, Chen S, Wang Y, Wen P, et al. Genomes of the Banyan Tree and Pollinator Wasp Provide Insights into Fig-Wasp Coevolution. Cell. 2020;183:875-889.e17.
12. Abhisek Chakraborty, Shruti Mahajan, Manohar Singh Bisht VKS. Genome sequencing and comparative analysis of Ficus benghalensis and Ficus religiosa trees reveal evolutionary mechanisms of longevity. https://doi.org/doi: https://doi.org/10.1101/2021.10.14.464369.
13. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, et al. The pineapple genome and the evolution of CAM photosynthesis. Nat Genet. 2015;47:1435–42.
14. Yang X, Hu R, Yin H, Jenkins J, Shu S, Tang H, et al. The Kalanchoë genome provides insights into convergent evolution and building blocks of crassulacean acid metabolism. Nat Commun. 2017;8:1899.
15. Guan Q, Tan B, Kelley TM, Tian J, Chen S. Physiological Changes in Mesembryanthemum crystallinum During the C3 to CAM Transition Induced by Salt Stress. Front Plant Sci. 2020;11.
16. Schlüter U, Denton AK, Bräutigam A. Understanding metabolite transport and metabolism in C4 plants through RNA-seq. Curr Opin Plant Biol. 2016;31:83–90.
17. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35 Web Server:W182–5.
18. Hays JB. Arabidopsis thaliana, a versatile model system for study of eukaryotic genome-maintenance functions. DNA Repair (Amst). 2002;1:579–600.
19. Spampinato CP, Gomez-Casati DF. Research on Plants for the Understanding of Diseases of Nuclear and Mitochondrial Origin. J Biomed Biotechnol. 2012;2012:1–12.
20. Jones AM, Chory J, Dangl JL, Estelle M, Jacobsen SE, Meyerowitz EM, et al. The Impact of Arabidopsis on Human Health: Diversifying Our Portfolio. Cell. 2008;133:939–43.
21. Jung HW, Son HY, Minh C Van, Kim YH, Park Y-K. Methanol extract of Ficus leaf inhibits the production of nitric oxide and proinflammatory cytokines in LPS-stimulated microglia via the MAPK pathway. Phyther Res. 2008;22:1064–9.
22. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
23. Jin J, Tian F, Yang D-C, Meng Y-Q, Kong L, Luo J, et al. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res. 2017;45:D1040–5.
24. Pan J, Huang D, Guo Z, Kuang Z, Zhang H, Xie X, et al. Overexpression of microRNA408 enhances photosynthesis, growth, and seed yield in diverse plants. J Integr Plant Biol. 2018;60:323–40.
25. Schramm M, Klybas V, Racker E. Phosphorolytic cleavage of fructose-6-phosphate by fructose-6-phosphate phosphoketolase from Acetobacter xylinum. J Biol Chem. 1958;233:1283–8.
26. Ebata M, Sato R, Bak T. The enzymatic phosphorylation of sedoheptulose*. J Biochem. 1955;42:715–25.
27. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 2000;28:27–30.
28. Kanehisa M. Toward understanding the origin and evolution of cellular organisms. Protein Sci. 2019;28:1947–51.
29. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for

697 taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2022.
698 https://doi.org/10.1093/nar/gkac963.
699 30. McFarlane CR, Shah NR, Kabasakal B V., Echeverria B, Cotton CAR, Bubeck D, et al.
700 Structural basis of light-induced redox regulation in the Calvin–Benson cycle in
701 cyanobacteria. Proc Natl Acad Sci. 2019;116:20984–90.
702 31. Cousins AB, Baroli I, Badger MR, Ivakov A, Lea PJ, Leegood RC, et al. The Role of
703 Phospho enol pyruvate Carboxylase during C4 Photosynthetic Isotope Exchange and
704 Stomatal Conductance. Plant Physiol. 2007;145:1006–17.
705 32. Mahesh HB, Subba P, Advani J, Shirke MD, Loganathan RM, Chandana SL, et al. Multi-
706 Omics Driven Assembly and Annotation of the Sandalwood (Santalum album) Genome.
707 Plant Physiol. 2018;176:2772–88.
708 33. Jones-Rhoades MW, Bartel DP, Bartel B. MicroRNAs and their regulatory roles in plants.
709 Annu Rev Plant Biol. 2006;57:19–53.
710 34. Sunkar R, Zhu J-K. Novel and Stress-Regulated MicroRNAs and Other Small RNAs
711 from Arabidopsis[W]. Plant Cell. 2004;16:2001–19.
712 35. Abdel-Ghany SE, Pilon M. MicroRNA-mediated Systemic Down-regulation of Copper
713 Protein Expression in Response to Low Copper Availability in Arabidopsis. J Biol Chem.
714 2008;283:15932–45.
715 36. Liu C, Xu X, Kan J, Cheng Z ming, Chang Y, Lin J, et al. Genome-wide analysis of the
716 C3H zinc finger family reveals its functions in salt stress responses of Pyrus betulaefolia.
717 PeerJ. 2020;8:e9328.
718 37. Lee D-K, Kim H Il, Jang G, Chung PJ, Jeong JS, Kim YS, et al. The NF-YA transcription
719 factor OsNF-YA7 confers drought stress tolerance of rice in an abscisic acid independent
720 manner. Plant Sci. 2015;241:199–210.
721 38. Bolt S, Zuther E, Zintl S, Hincha DK, Schmülling T. ERF105 is a transcription factor
722 gene of Arabidopsis thaliana required for freezing tolerance and cold acclimation. Plant Cell
723 Environ. 2017;40:108–20.
724 39. Rawat R, Schwartz J, Jones MA, Sairanen I, Cheng Y, Andersson CR, et al.
725 REVEILLE1, a Myb-like transcription factor, integrates the circadian clock and auxin
726 pathways. Proc Natl Acad Sci. 2009;106:16883–8.
727 40. Mizoguchi T, Wheatley K, Hanzawa Y, Wright L, Mizoguchi M, Song H-R, et al. LHY
728 and CCA1 Are Partially Redundant Genes Required to Maintain Circadian Rhythms in
729 Arabidopsis. Dev Cell. 2002;2:629–41.
730 41. Smitha RB, Bennans T, Mohankumar C, Benjamin S. Oxidative stress enzymes in Ficus
731 religiosa L.: Biochemical, histochemical and anatomical evidences. J Photochem Photobiol B
732 Biol. 2009;95:17–25.
733 42. Christin P-A, Arakaki M, Osborne CP, Bräutigam A, Sage RF, Hibberd JM, et al. Shared
734 origins of a key enzyme during the evolution of C4 and CAM metabolism. J Exp Bot.
735 2014;65:3609–21.
736 43. NOBEL PS. Achievable productivities of certain CAM plants: basis for high values
737 compared with C 3 and C 4 plants. New Phytol. 1991;119:183–205.
738 44. Sharkey TD. Estimating the rate of photorespiration in leaves. Physiol Plant.
739 1988;73:147–52.
740 45. Igarashi D, Miwa T, Seki M, Kobayashi M, Kato T, Tabata S, et al. Identification of
741 photorespiratory glutamate:glyoxylate aminotransferase (GGAT) gene in Arabidopsis. Plant
742 J. 2003;33:975–87.
743 46. Heyduk K, Ray JN, Ayyampalayam S, Moledina N, Borland A, Harding SA, et al. Shared
744 expression of crassulacean acid metabolism (CAM) genes pre-dates the origin of CAM in the
745 genus Yucca. J Exp Bot. 2019;70:6597–609.
746 47. Ming R, VanBuren R, Wai CM, Tang H, Schatz MC, Bowers JE, et al. The pineapple

747 genome and the evolution of CAM photosynthesis. Nat Genet. 2015;47:1435–42.
748 48. Yin H, Guo H-B, Weston DJ, Borland AM, Ranjan P, Abraham PE, et al. Diel rewiring
749 and positive selection of ancient plant proteins enabled evolution of CAM photosynthesis in
750 Agave. BMC Genomics. 2018;19:588.
751 49. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of
752 occurrences of k-mers. Bioinformatics. 2011;27:764–70.
753 50. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al.
754 GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics.
755 2017;33:2202–4.
756 51. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes:
757 A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J
758 Comput Biol. 2012;19:455–77.
759 52. Zimin A V., Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA
760 genome assembler. Bioinformatics. 2013;29:2669–77.
761 53. Boetzer M, Henkel C V., Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled
762 contigs using SSPACE. Bioinformatics. 2011;27:578–9.
763 54. Kosugi S, Hirakawa H, Tabata S. GMcloser: closing gaps in assemblies accurately with a
764 likelihood-based selection of contig or long-read alignments. Bioinformatics. 2015;:btv465.
765 55. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for
766 genome assemblies. Bioinformatics. 2013;29:1072–5.
767 56. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
768 assessing genome assembly and annotation completeness with single-copy orthologs.
769 Bioinformatics. 2015;31:3210–2.
770 57. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: A
771 Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome
772 Annotations. Plant Physiol. 2014;164:513–24.
773 58. Xu L, Dong Z, Fang L, Luo Y, Wei Z, Guo H, et al. OrthoVenn2: a web server for whole-
774 genome comparison and annotation of orthologous clusters across multiple species. Nucleic
775 Acids Res. 2019;47:W52–8.
776 59. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence
777 alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20:1160–6.
778 60. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
779 Bioinformatics. 2009;25:1754–60.
780 61. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years
781 of SAMtools and BCFtools. Gigascience. 2021;10.
782 62. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the
783 development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare
784 L.). Theor Appl Genet. 2003;106:411–22.
785 63. Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA
786 Genes in Genomic Sequence. Nucleic Acids Res. 1997;25:955–64.
787 64. Kong L, Zhang Y, Ye Z-Q, Liu X-Q, Zhao S-Q, Wei L, et al. CPC: assess the protein-
788 coding potential of transcripts using sequence features and support vector machine. Nucleic
789 Acids Res. 2007;35 suppl_2:W345–9.
790 65. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein
791 or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
792 66. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al.
793 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and
794 isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
795 67. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or
796 without a reference genome. BMC Bioinformatics. 2011;12:323.

797 68. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a
798 universal tool for annotation, visualization and analysis in functional genomics research.
799 Bioinformatics. 2005;21:3674–6.
800
801

802 **Tables**

803 **Table 1:** Final assembly and annotation of Peepal genome

804

Features	Final assembly
Total length of assembled sequence (Mb)	406,103,086
Number of scaffolds/contigs	202,258
Minimum scaffold length (bp)	200
Maximum scaffold length (bp)	148483
GC %	34.23
N50 (bp)	5817
L50	17,605
Number of annotated genes	35093

805
806
807
808
809
810
811
812

813 **Table 2:** Statistics of Uni-genes in Peepal Transcripts

Features	Sample (Day - 2 PM)	Sample (Night - 2 AM)	Combined Sample
Number of genes	22597	23360	26479
Number of transcripts	16912	16780	18173
Percent GC	46.37	42.42	46.25
Contig N50 (nts)	1404	2053	1374
Median contig length	813	1011	753
Average Contig (nts)	1060.06	1332.73	1017.44
Total assembled bases	23954145	120608982	27055237

814

815

816

817 **Figure legends**

818 **Figure 1:** Orthologous clustering of 6 species using proteome data deduced 24310 orthologous
819 gene clusters and 1184 single-copy gene clusters across the above 6 species.

820 **Figure 2:** The candidate genes involved in the C3, CAM, and C4 cycles.

821 A) Calvin- Benson (C3) cycle B) Crassulacean acid metabolism (CAM) cycle C) C4 cycle. The
822 X-axis represents the genes involved in pathways, Y-axis is the matrix of normalized
823 expression trimmed mean of M (TMM) values; the Blue graph - leaf tissue collected during
824 the day (2 PM), Red graph - leaf tissue collected during the night period (2 AM).

825 **Figure 3:** Gene expression pattern of *F. religiosa* carbon fixation genes across the diurnal (2
826 PM) and nocturnal (2 AM) expression data. A) C3 cycle, B) CAM cycle C) C4 cycle log2-
827 transformed Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value-
828 based expression profiles are shown.

829

830

831

832

833

834

835

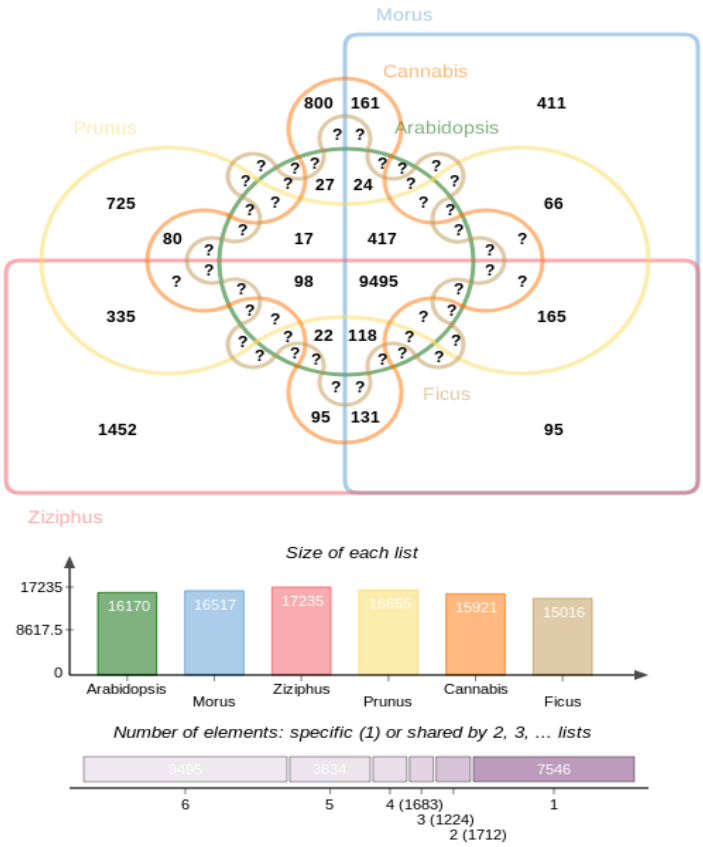

Figure 1: Orthologous clustering of 6 species using proteome data deduced 24310 orthologous gene clusters and 1184 single-copy gene clusters across the above 6 species

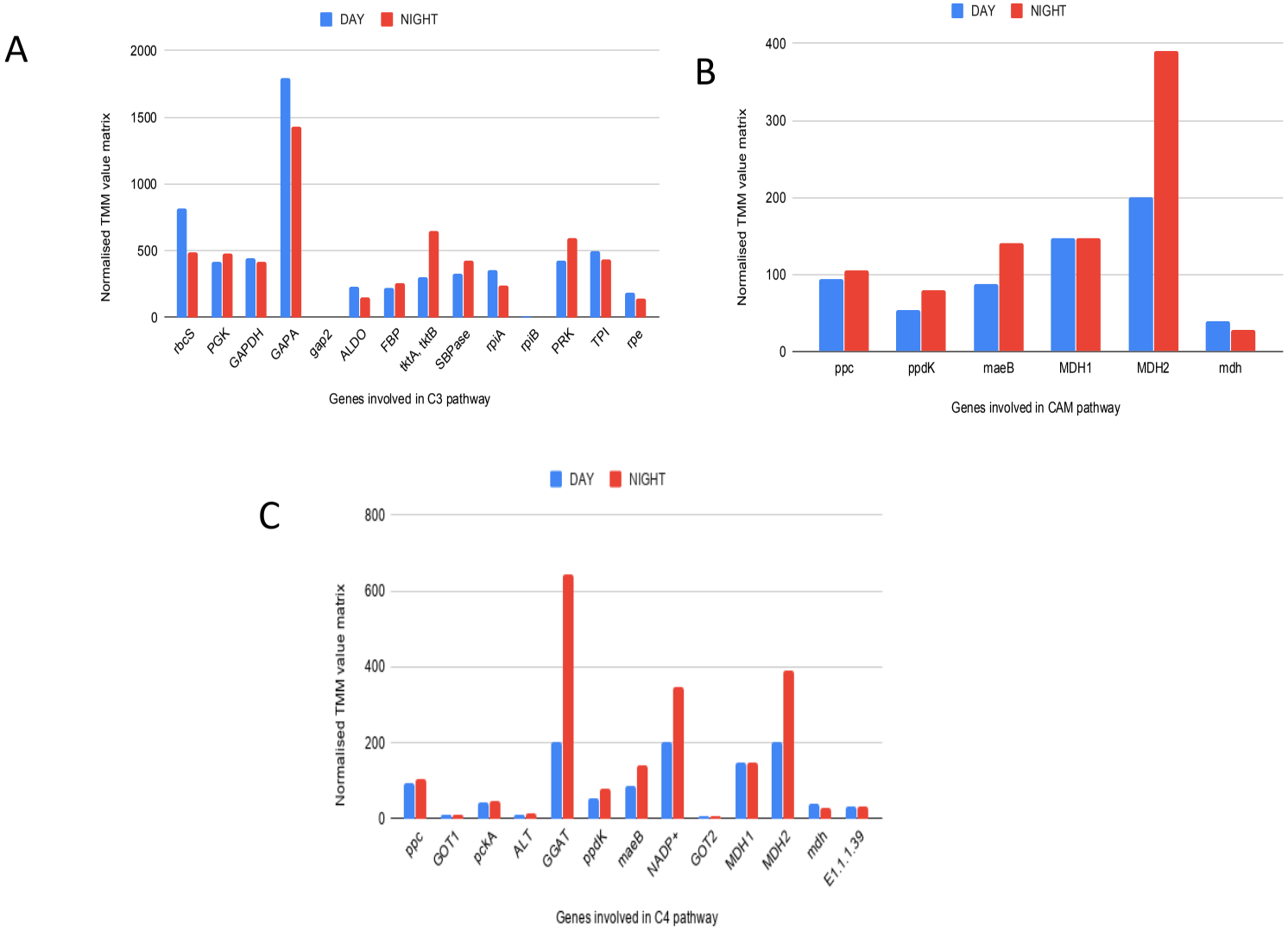

Figure 2: The candidate genes involved in the C3, CAM and C4 cycle.

A) Calvin- Benson (C3) cycle B) Crassulacean acid metabolism (CAM) cycle C) C4 cycle. X – axis represents the genes involved pathways, Y – axis is the matrix of normalised expression trimmmed mean of M (TMM) values; Blue graph – leaf tissue collected during day period (2 PM), Red graph – leaf tissue collected during night period (2 AM)

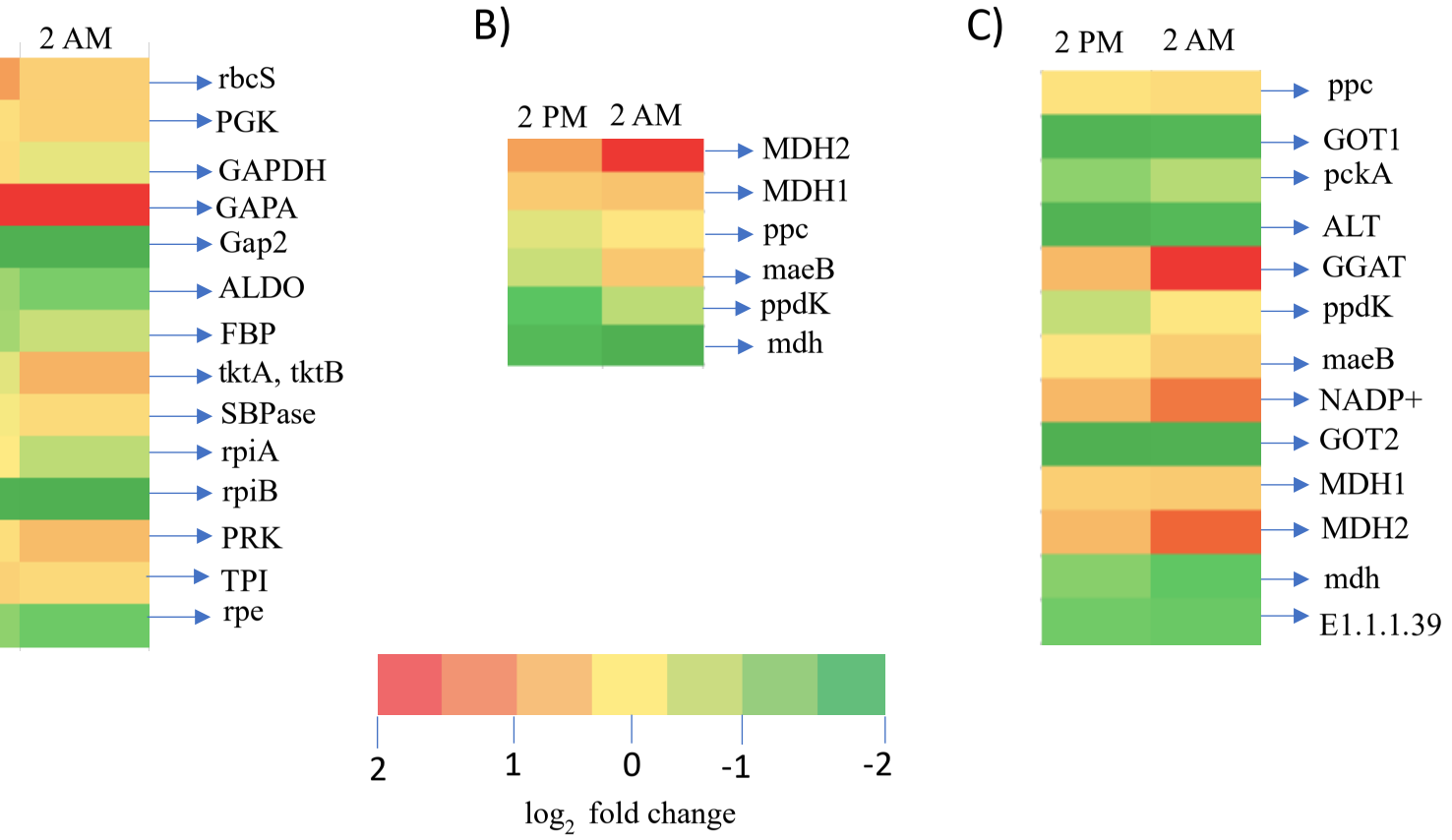

Figure 3: Gene expression pattern of *F. religiosa* carbon fixation genes across the diurnal (2 PM) and nocturnal (2 AM) expression data. A) C3 cycle, B) CAM cycle C) C4 cycle with fold change log2-transformed Fragments Per Kilobase of transcript per Million mapped reads (FPKM) value based expression profiles are shown