

1 **Title:** Identification of novel flowering genes using RNA-Seq pipeline employing
2 combinatorial approach in *Arabidopsis thaliana* time-series apical shoot meristem data

3 **Abstract:**

4 Floral transition is a crucial event in the reproductive cycle of a flowering plant during which many
5 genes are expressed that govern the transition phase and regulate the expression and functions
6 of several other genes involved in the process. Identification of additional genes connected to
7 flowering genes is vital since they may regulate flowering genes and vice versa. Through our
8 study, expression values of these additional genes has been found similar to flowering genes *FLC*
9 and *LFY* in the transition phase. The presented approach plays a crucial role in this discovery. An
10 RNA-Seq computational pipeline was developed for identification of novel genes involved in floral
11 transition from *A. thaliana* apical shoot meristem time-series data. By intersecting differentially
12 expressed genes from Cuffdiff, DESeq and edgeR methods, 690 genes were identified. Using
13 FDR cutoff of 0.05, we identified 30 genes involved in glucosinolate and glycosinolate biosynthetic
14 processes as principle regulators in the transition phase which provide protection to plants from
15 herbivores and pathogens during flowering. Additionally, expression profiles of highly connected
16 genes in protein-protein interaction network analysis revealed 76 genes with non-functional
17 association and high correlation to flowering genes *FLC* and *LFY* which suggests their potential
18 and principal role in floral regulation not identified previously in any studies.

19

20 **Keywords:** Apical shoot; Flowering; Pipeline; Cuffdiff; Step Analysis; Differential expression;
21 Enrichment; *Arabidopsis Thaliana*

22

23

24 **1. Background**

25

26 RNA-Seq is now a frequently used method in plant biology. Studies using mutant analysis on
27 *A.thaliana* have uncovered a range of genes involved in flowering (Koornneef, Hanhart and van
28 der Veen, 1991). Analysis of *A. thaliana* apical shoot data suggest that certain known genes are
29 commonly regulated during the transition phase (Klepikova *et al.*, 2015). Recent RNA-Seq studies
30 on fruit plants such as blueberry have used Cuffdiff software (Trapnell *et al.*, 2012) with default
31 parameters (Gupta *et al.*, 2015) to identify candidate genes. Similar studies have also been
32 performed in *A. thaliana* using different computational approaches with default parameters for
33 alternative splicing detection (Liu, Loraine and Dickerson, 2014). Many of these RNA-Seq studies
34 focus on identification of differentially expressed genes (DEGs) using bioinformatics tools.
35 However, very often these comparative studies fail to consider optimal parameters required for
36 upstream processing of the data prior to DEG analysis. Due to this, RNA-Seq studies involving
37 plants sometimes do not generate optimal results.

38 Accurate mapping of raw RNA-Seq reads is essential for identification of DEGs. A recent
39 study by Zhao and Zhang (Zhao and Zhang, 2015) showed that divergent genesets can influence
40 the outcome of the analysis. They characterized the impact of genome annotation choice on read
41 mapping and transcript quantification by analyzing RNA-Seq datasets (Zhao and Zhang, 2015).
42 In addition, certain parameters also affect mapping of reads to the transcriptome. For example, in
43 the popular splice junction mapper TopHat2 (Kim *et al.*, 2013) there are parameters to control
44 'unique mapping' and 'mapping of concordant read pairs'. By changing the value of the 'unique
45 mapping' parameter, reads are only aligned the specified number of times to the genome and
46 therefore the overall mapping rate is affected. Similarly, by switching on the 'mapping of
47 concordant read pairs' parameter, both read pairs map to the same sequence in correct
48 orientation with suitable distance between them. Parameters such as 'maximum fragment length'
49 also affect how the reads are mapped for paired-end reads. Default values of these parameters,
50 which are designed for mammalian genomes, when applied in plants can result in the loss of
51 paired reads that map further than certain default base pairs (bp). Therefore, using default
52 parameters can lead to false positive results. Furthermore, commercial RNA-Seq software such
53 as CLC Genomics fails to consider this aspect, which can lead to generation of unreliable results.
54 Therefore, considering the impact of parameters on read mapping to identification of DEGs, a
55 standardized computational pipeline is required.

56 In this study, we identified novel DEGs which can potentially regulate the function of flowering
57 genes using a bioinformatics pipeline from plant RNA-Seq time-series data processing (Fig. 1).
58 The pipeline which consisted of three components: data processing; Differential Gene Expression
59 (DGE); and Gene Ontology (GO) annotation and enrichment, was applied to *A. thaliana* apical
60 shoot data. We identified several thousand differentially expressed genes during the transition
61 phase (i.e. from vegetative to reproductive developmental phase) by intersection of DGE results
62 obtained from Cuffdiff, DESeq and edgeR tools. Expression profiles of the overlapped genes were
63 studied and GO enrichment analysis was used to identify a list of novel candidate genes involved
64 in flowering. Finally, we identified several genes with crucial protein interactions and potential
65 roles in the regulation of flowering.

66

67 **2. Methods**

68 *2.1 Data analysis pipeline*

69 The pipeline workflow (Fig. 1), starts with conversion of reads in SRA format to FASTQ format as
70 FASTQ files were needed for sequence alignment (Ostell and McEntyre, 2007). A quality metric
71 report was generated using FastQC tool (Andrews, 2010) which briefly outlines metrics of
72 sequence quality, quality scores, sequence content, sequence length distribution, sequence
73 duplication levels, overrepresented sequences, adapter content and Kmer content. Based on the
74 metrics, reads were trimmed to generate trimmed read files for each sample using Cutadapt
75 (Martin, 2011). Following read trimming, samples were again checked for contaminated
76 sequences, adapters, and poor-quality reads so that they could be removed before alignment.
77 Sample reads were then each aligned to the *A. thaliana* genome using TopHat2 (Kim *et al.*, 2013)
78 which is a fast splice junction mapper based on Bowtie2 (Langmead and Salzberg, 2012).
79 Cufflinks and Cuffmerge were used for transcript assembly and transcript merging. DGE was
80 performed using Cuffdiff (Trapnell *et al.*, 2012). For DGE analysis using DESeq and edgeR, BAM

81 files obtained from Tophat2 are converted to raw read counts using HTSeq (Anders, Pyl and
82 Huber, 2015). There are many other transcript quantification tools available, such as RSEM (Li
83 and Dewey, 2011) and StringTie (Pertea *et al.*, 2015), which utilize the BAM file generated from
84 Tophat2 and reference GTF file and produce Reads Per Kilobase of transcript per Million (RPKM)
85 reads. HTSeq utilizes a simpler approach and produces raw read count from the SAM file and
86 reference GTF file. Raw reads were then used for DGE using DESeq (Anders and Huber, 2010)
87 and edgeR (Robinson, McCarthy and Smyth, 2010). There are many other tools available for
88 RNA-Seq DGE analysis such as NPEBseq (Bi and Davuluri, 2013), NOISeq (Tarazona *et al.*,
89 2012), baySeq (Hardcastle and Kelly, 2010), DEGSeq (Wang *et al.*, 2009) and EBSeq (Leng *et al.*,
90 2013). Cuffdiff was chosen as it is specifically designed for DGE analysis from transcripts,
91 spliced regions and promoters, and is best suited to use in conjunction with TopHat2. Since some
92 of the samples in the study did not contain any replicates, DESeq and edgeR were chosen
93 because both the tools are designed to work with and without replicates. Additionally, coupling
94 HTSeq with DESeq and edgeR helps in direct integration of raw read counts from *htseq-count* as
95 input into DESeq and edgeR programs. Post-analysis was performed using the SpliceR (Vitting-
96 Seerup *et al.*, 2014) tool for annotation of transcript features obtained from Cuffdiff (Trapnell *et al.*,
97 2012). Results from Cuffdiff, DESeq and edgeR were merged to obtain collective DEGs in the
98 sample pairs. The final step of the pipeline consisted of gene enrichment, pathway analysis and
99 protein-protein interaction (PPI) network analysis using the Araport portal (Krishnakumar *et al.*,
100 2015), ClueGO (Bindea *et al.*, 2009) and GeneMania (Montojó *et al.*, 2010) for identifying novel
101 gene clusters associated with flower development.

102

103 *2.2 Sample collection and data preparation*

104 Experiments were conducted by Klepikova *et al.* (Klepikova *et al.*, 2015) where a single *A. thaliana*
105 plant (Col-0) was grown in conditions for preventing crossbreeding. Plants were harvested at 7-
106 16 days old to obtain synchronized plants at different developmental stages, denoted by S7 to
107 S16 respectively. Hand-dissected shoot apical meristems were fixed in two biological replicates
108 using tissues from 15 individuals in each sample. 9-14-day old plants were collected in two
109 replicates for a second independent experiment denoted by S9N to S14N respectively. Sequence
110 reads for each sample were obtained from the NCBI Sequence Read Archive under project ID
111 PRJNA268115. A quality report was generated using the FastQC tool to obtain the statistics of
112 reads (Andrews, 2010). The SRA data was converted to FASTQ format using the 'fastq-dump'
113 tool available in the SRA Toolkit, NCBI (Sayers *et al.*, 2009).

114

115 *2.3 Read trimming, reference genome mapping and transcript assembly*

116 Adapter trimming and genome mapping represents the pre-processing step, as seen in Fig. 1.
117 First 15 base pairs of the reads were trimmed using Cutadapt to remove adapter sequences and
118 improve the quality of the reads to keep reads with a Q-score greater than or equal to 30 (Martin,
119 2011). Each sample consists of two reads; therefore, each read was trimmed and a FastQC report
120 was regenerated on the trimmed data to examine the quality and verify that the resultant reads
121 satisfied the criterion. Trimmed reads were mapped to the *A. thaliana* genome (TAIR10) using
122 the TopHat2 aligner (Langmead and Salzberg, 2012; Kim *et al.*, 2013). We ran TopHat2 on both
123 the reads, with the values of the following default parameter changed to suit *A. thaliana*: minimum

124 intron length (-i) was set to 40, maximum intron length (-l) was set to 5000, segment length was
125 set to 20, segment mismatches was set to 2, no multi-hits (-g=1), minimum normalized depth (F)
126 was set to 0 and minimum anchor length was set to 10 (-a=10). The parameter values are
127 summarized in Table 1. Trimmed reads were also aligned using Bowtie2 with minimum (i) and
128 maximum (l) intron length set to 30 and 5000 respectively. Previously, it was reported in some
129 studies that the minimum functional intron length for monocots and dicots was found to lie
130 between 70 to 73 nucleotides (nt) (Goodall and Filipowicz, 1990). However, certain genome-wide
131 studies performed on *A. thaliana* and *O. sativa* datasets provides evidence of introns shorter than
132 60 nt with size range of 20 to 59 nt in length (Deutsch and Long, 1999; Wang and Brendel, 2006).
133 Similarly, the maximum intron length for plant genomes, which is otherwise set to 500bp, is much
134 larger than vertebrates. Since each read is cut into segments that are mapped independently, the
135 segment length for shorter reads in our case should be decreased from its default value. By setting
136 the max multihits option to 1, we are forcing unique mapping of the reads to the genome which
137 will allow best mapping of the read to the genome. By setting the value of minimum anchor length
138 to 10 instead of 8, TopHat2 will report junctions spanned by reads with at least this many bases
139 on each side of the junction. Finally, to eliminate the heuristic filter associated with vertebrate
140 genomes, minimum normalized depth was set to 0 instead of 300.

141 Reads aligned using TopHat2 were then used by Cufflinks (Fig. 1) for assembling
142 individual transcripts with $-i$ 40 and $-l$ 5000 parameters (Trapnell *et al.*, 2012). In plant genomes,
143 the difficulty of estimation of transcript abundance arises due to multi-reads and the genome
144 becomes highly-repetitive. Therefore, to address the uncertainty, an Expectation-Maximization
145 algorithm (EM) has been applied using Cufflinks for estimating transcript abundance which
146 computes fractional distribution of each multi-read after read alignment in the E-step and then
147 estimates relative abundance of transcripts in the M-step until it converges. After obtaining the
148 transcripts for each read, transcripts from two comparable samples were merged using Cuffmerge
149 (Trapnell *et al.*, 2012). For example, for comparing S7 with S8, the transcripts of each read of the
150 two samples which is equal to 4 read transcripts were merged to form an assembled transcript
151 GTF file for further analysis.

152

153 *2.4 Differential gene expression analysis on time-series apical shoot data*

154

155 Differential expression analysis of the reads was carried out by testing the samples at day 8 to 16
156 (S8 to S16) after germination against the sample from day 7 (S7) to obtain DEGs at each
157 consecutive stage. The reason why day 7 was chosen for benchmarking was that plants at this
158 stage after germination had the first and second leaves visible; while on day 16th they had ten
159 visible leaves. Comparisons of two samples from consecutive days were also made (Table 2).
160 These analyses were carried out using Cuffdiff (Trapnell *et al.*, 2012). The multi-read-correct
161 option was enabled to carry out an initial estimation procedure for accurately weighting reads for
162 mapping to multiple locations on the genome. Quartile normalization was used to obtain
163 Fragments Per Kilobase Million (FPKM) and fragment counts via the ratio of 75th quartile fragment
164 counts to 75th quartile value across all samples. False discovery rate (FDR) adjusted p-values
165 (known as q-values) were obtained from the analysis and significantly expressed genes were
166 obtained by filtering q-values less than or equal to 0.05.

167 Sequence read counts were obtained from the reads aligned by Tophat2 using the HTSeq
168 tool to generate raw read counts (Anders, Pyl and Huber, 2015). The read counts were then used
169 to produce a list of differentially expressed genes using DESeq (Anders and Huber, 2010) and
170 edgeR (Robinson, McCarthy and Smyth, 2010). As in the previous step, comparative analysis of
171 S7 against S8 to S16 and step-wise analysis were conducted. Since the dataset contains partial
172 replicates for 5 samples (S9N to S14N), we used blind dispersion estimation for samples with no
173 replicates along with the sharing mode set to ‘fit-only’ and we used pooled empirical dispersion
174 for samples with one or more replicates. The negative binomial method was applied for obtaining
175 DEGs. Results were filtered based on FDR \leq 0.05 and \log_2 fold-change less than -2 and greater
176 than 2. To compare samples involving replicates, the generalized linear model (GLM) was applied
177 for estimating common and tagwise dispersion. To compare samples for which no replicates were
178 found, Fisher’s exact test was applied with the biological coefficient of variation set to 0.2
179 (Benjamini and Hochberg, 1995). For performing DGE analysis using edgeR for samples having
180 no biological replicates, we used common BCV (Biological Coefficient of Variation) with square-
181 root dispersion value which was set to 0.4 for humans and 0.1 for genetically identical organisms.
182

183 *2.5 Alternative splicing classification analysis using SpliceR*

184
185 To obtain statistics of transcript level information, we utilized SpliceR (Vitting-Seerup *et al.*, 2014)
186 to classify isoform transcripts obtained from Cuffdiff. Output files containing FPKM tracking, count
187 tracking and read group tracking files enabled us to detect exon skipping/inclusion (ESI) events,
188 alternative transcription start site (ATSS), alternative transcription termination site (ATTS),
189 alternative 3’ splice site (A3), alternative 5’ splice site (A5) and mutually exclusive exon (MEE)
190 events. Additionally, the average number of transcripts per gene and the average number of ESI
191 events per transcript were computed using the *spliceR* function for each of the “Against S7” and
192 “Step Analysis” sample pairs.
193

194 *2.6 GO enrichment, pathway and protein-protein interaction analysis*

195
196 Results obtained from the overlap of Cuffdiff, DESeq and edgeR were used for functional
197 enrichment to categorize genes and their associated functions. Overlapping DEGs had to express
198 more than once in “Against S7” and “Step analysis” results to be retained for further analysis. GO
199 enrichment functional annotation and clustering of the genes were performed using the Araport
200 portal (Krishnakumar *et al.*, 2015) to identify genes associated with enriched categories. Gene
201 identifiers (e.g. AT1G02335) were used as inputs into the Araport Thalemine tool. These
202 identifiers were then used for enrichment in gene ontologies (biological process, cellular
203 component and molecular function). Pathway analysis was performed using the ClueGo plugin
204 (Bindea *et al.*, 2009) of the Cytoscape software (Shannon *et al.*, 2003). Gene identifiers were
205 used to identify the association and clustering of genes in pathways using KEGG (Ogata *et al.*,
206 1999), REACTOME (Croft *et al.*, 2014) and WikiPathways (Kutmon *et al.*, 2015) databases.
207 Enrichment or depletion of GO categories in ClueGO was performed using the two-sided
208 hypergeometric test and FDR was calculated for the enriched GO categories using the Benjamin
209 and Hochberg approach (Benjamini and Hochberg, 1995). Gene enrichment and clustering
210 results obtained from Araport and Cytoscape were further filtered with FDR \leq 0.05 to identify

211 highly significant enriched clusters. A PPI network was constructed using the GeneMania plugin
212 (Montejo *et al.*, 2010) of the Cytoscape software (Shannon *et al.*, 2003) to obtain prevalent
213 interactors and their degree of interactions from the network.

214

215 *2.7 Calculation of relative expression values*

216

217 To calculate relative expression values, FPKM counts were used in each sample pair and counts
218 were normalized by dividing by the sample pair read count by the maximum read count value
219 from all other sample pairs (i.e. S7-S8 to S7-S16) to obtain relative expression value between 0
220 and 1. Expression profiles of each gene were constructed by comparing expression values from
221 Cuffdiff and DESeq-edgeR.

222

223 *2.8 Calculating correlation of expression values*

224

225 For calculating correlation between the expression profiles, Pearson Correlation Coefficient
226 (PCC) was used. Expression profiles of differentially expressed genes involved in flower
227 development were compared against expression profiles of *FLC* and *LFY* genes to obtain PCC
228 between them. Also, difference in expression using PCC was also evaluated by comparing
229 expression profiles of genes obtained from Cuffdiff, DESeq and edgeR with those obtained from
230 Klepikova *et al.* (Klepikova *et al.*, 2015).

231

232 *2.9 Data availability*

233

234 An apical shoot meristem time-series dataset has been deposited by other researchers in the
235 NCNI SRA database under project ID PRJNA268115 (Klepikova *et al.*, 2015) has been used in
236 this study. Analysis pipelines listing the data processing and differential gene expression steps
237 constructed and revealed through our research has been provided as Linux shell scripts which
238 can be downloaded from GitHub <https://github.com/deshpan4/RNA-Seq-pipeline>.

239

240 **3. Results**

241 *3.1 Differential expression analysis of time-series apical shoot data*

242

243 Results obtained from DGE of five sample pairs were computed in “Against S7” and “Step
244 Analysis” manner as detailed in Table 2. When comparing with S7, 5,266 DEGs were obtained
245 for S7-S10, 2,841 genes for S7-S11, 4,760 for S7-S12, 6,337 for S7-S13 and 2,532 genes for S7-
246 S14 pair. DGE using “Step analysis” was performed to identify genes differentially expressed from
247 the previous day which yielded fewer genes as compared to that obtained from “Against S7”
248 sample pairs.

249 Next, we studied the overlap between Cuffdiff, DESeq and edgeR for sample pairs in
250 “Against S7” and “Step Analysis”. By overlapping DEGs, 418 genes were found for S7-S10 with
251 FDR ≤ 0.05 . Using the same cutoff, S7-S11 generated 277 genes, S7-S12 produced 520 genes,
252 S7-S13 gave 1,534 genes and S7-S14 gave 150 genes (Table 3). On the other hand, 28 genes
253 were found for S9-S10, 3 genes for S10-S11, 7 genes for S11-S12, 38 genes for S12-S13 and

254 74 genes were found for S13-S14. Overlapping genes were also found for Cuffdiff-edgeR,
255 DESeq-edgeR and Cuffdiff-edgeR-DESeq pairs (Table 3). From Cuffdiff-DESeq-edgeR overlap,
256 we identified 690 genes in “Against S7” and 19 genes in “Step analysis” which are significantly
257 expressed in more than one sample pair. This set of common genes is referred to as CGenes in
258 the following analysis.

259 260 3.2 GO enrichment and pathway analysis of differentially expressed genes in the transition phase 261

262 GO enrichment analysis applied to CGenes were classified in three categories: Biological Process
263 (BP), Molecular Function (MF) and Cellular Component (CC). Results from GO enrichment (Fig.
264 2) of common genes obtained from “Against S7” sample pairs show 664 genes were significantly
265 enriched in BP and CC ontologies with p-values < 0.05. Whereas those obtained from “Step
266 Analysis” sample pairs show 18 genes significantly enriched only in the BP ontology with p-value
267 < 0.05. From the pathway analysis of “Against S7” DEGs, 30 genes have been found to be
268 involved in Glucosinolate Biosynthesis, 2-Oxocarboxylic acid metabolism, Sulfur metabolism,
269 Cysteine and methionine metabolism with FDR \leq 0.05 whereas for “Step Analysis” only 4 genes
270 were found to be involved in 2-Oxocarboxylic acid metabolism, C5-Branched dibasic acid
271 metabolism, Valine, leucine and isoleucine biosynthesis.

272 Additionally, pathway analysis using ClueGo revealed 33 genes associated with 8 GO
273 terms which were found specifically in KEGG and REACTOME (Table 4). Pathway enrichment of
274 19 common genes from “Step Analysis” sample pairs shows association of *BCAT4*, *IMD1* and
275 *IPM12* genes in the valine, leucine and isoleucine biosynthesis pathways with significant term p-
276 value < 0.05.

277 278 3.3 Analysis of expression profiles of enriched genes involved in glucosinolate biosynthesis and 279 metabolism 280

281 Expression profiles were constructed from the set of enriched genes by selecting the highly
282 enriched cluster from the CGenes set. Fig. 3 shows the relative expression profiles of the genes
283 expressed in “Against S7” and “Step analysis” sample pairs that play major roles in Glucosinolate
284 Biosynthesis (GluBP), Glycosinolate Biosynthesis (GlyBP), Glucosinolate Metabolic Process
285 (GluMP), Glycosinolate Metabolic Process (GlyMP), Sulfur-Compound Biosynthetic Process
286 (SCBP) and Sulfur-Compound Metabolic Process (SCMP). 21 genes have been found to be
287 associated with GluBP and GlyBP, 27 associated with GluMP and GlyMP, 25 associated with
288 SCBP and 37 have been found to be associated with SCMP. From the expression profiles
289 expression of the genes peaks at S7-S8 and S7-S9 pairs in “Against S7”. The expression
290 decreases to 0.4 in S7-S10. It continues to decrease until it reaches 0 in S7-S13 and continues
291 for the rest of the samples for most of the genes. It is clearly visible that *ACO1*, *ACO2*, *APS1* and
292 *AT4G05090* display different behavior where expression varies between 1 and 0.6 for SCMP. In
293 SCBP, *CYSD1* expression value remains constant between 0.6 and 0.8 whereas for *CYP83B1*
294 value suddenly increases from 0.4 in S7-S12 to 0.9 in S7-S13, drops to 0.4 in S7-S14 and again
295 increases to 0.9 in S7-S15 and S7-S16. In GBP, only *CYP83B1* shows variable expression. Apart
296 from these genes, certain other genes such as *TGG1* and *TGG2* show a “zig-zag” expression
297 pattern which encodes myrosinase enzymes and helps in the breakdown of glucosinolates (Barth

298 and Jander, 2006). As compared to these genes, *CYP83B1* and *CYP83A1* are expressed in the
299 SCMP, SCBP and GluBP. These encode non-redundant enzymes which also metabolize oximes
300 in glucosinolate biosynthesis (Naur *et al.*, 2003). Where the expression of *CYP83A1* follows a
301 general curve of steep decrease in expression from S7-S9, expression of *CYP83B1* is non-
302 identical and shows a “zig-zag” expression pattern like *TGG1* and *TGG2*. Similarly, *ACO1* and
303 *ACO2* in the SCMP also differ in their expression profiles despite being similar in structure and
304 function. *CYP79B3* also encodes a cytochrome protein however its expression is dissimilar from
305 *CYP83B1* which can be clearly distinguished in the GluBP where the value starts to decrease
306 from S7-S8 to S7-S14 and increases from 0.08 to 0.43 in S7-S16.

307

308 3.4 Expression profiles of differentially expressed flowering genes

309 From CGenes, genes responsible for flowering and involved in regulation of flower development
310 were identified. 5 genes were found to be involved in “Flowering”. 18 were found to be associated
311 with “Flower Development”, 8 with “Regulation of Flower Development” and 3 with “Negative
312 Regulation of Flower Development”. Fig. 4 shows expression profiles of genes involved in
313 flowering, flower development, regulation of flower development and negative regulation of flower
314 development. In “Against S7” sample pairs, many experimental genes such as *FLC*, *SOC1*, *EMS1*
315 and *FD* have also been identified by enrichment analysis. Expression profiles of flowering genes
316 shows that *SOC1*, *FCA*, *SAP* and *AGL31* increase in expression as compared to *FLC* which
317 decreases in expression in “Against S7”. All four genes show increase in expression in S7-S10
318 which is followed by decrease in expression in S7-S11. *FCA*, *SAP* and *SOC1* show highest
319 expression in S7-S14 whereas the expression for *AGL31* remains constant between 0.2 and 0.4
320 and finally increases to 1 in S7-S16. In the “Flower Development” process, a large cluster of genes
321 in “Against S7” sample pairs display a “zig-zag” pattern of expression. There are four gene
322 clusters observed in this process. The first cluster consists of *ATX1*, *RDR6*, *SOC1*, *KAN2*, *BPE*,
323 *SRS2*, *FCA*, the expression values of which increase in S7-S9, decrease in S7-S11 and increase
324 again in S7-S12. The second cluster consists of *ATX1*, *NAC054*, *NGA1* and *F-ATMBP* shows a
325 decrease in expression followed by an increase in S7-S15 and S16. The third cluster consists of
326 *EMS1*, *KAN2*, *ABC19*, *SOC1* and *SAP1* shows a peak in expression value from S7-S14. The
327 fourth cluster of genes consists of *SPT*, *SRS2*, *ATX1* and *FCA* in S7-S14 where the expression
328 varies between 0.7 and 0.8. In the “Regulation of Flower Development” process, *POLA*, *FD*,
329 *ATX1*, *SOC1*, *AGL31* and *FCA* show a decrease in expression in S7-S11 whereas *ATX1* shows
330 an increase in expression in S7-S11. In the “Negative Regulation of Flower Development”
331 process, only *FLC*, *AGL31* and *POLA* are expressed. It is important to note that *FLC* has been
332 found to be involved in all the processes of flowering, flower development, regulation and negative
333 regulation of flower development.

334 3.5 Identifying important regulators using protein-protein interaction (PPI) network analysis

335 Interactions between DEGs were studied for identifying most prevalent interacting genes and their
336 regulation on neighboring genes. PPI network was constructed for identifying highly connected
337 genes and their most prevalent interactions (Figure 5a). From PPI network analysis, 18 genes
338 were found to have highest interactions with edges ≥ 100 and significantly involved in
339 Glucosinolate Biosynthesis. PPI network analysis revealed that along with 18 genes, 114 genes

340 (Figure 5b) were involved in induced systemic resistance, sulphur compound biosynthetic
341 process, cellular biogenic amine metabolic process, sulphur metabolism and biosynthesis, anion
342 transport, organic acid transport and cellular response to external stimulus.

343

344 *3.6 Identification of floral candidate genes*

345

346 FPKM expression values of *FLC* and *LFY* genes from Day-1 to 10 were used to identify potential
347 novel genes from the CGenes set by selecting those displaying the highest correlation with *FLC*
348 and *LFY* expression profiles and having no ontology information for *A. thaliana*. Results of
349 correlation and GO enrichment analysis using Araport (Krishnakumar *et al.*, 2015) showed that
350 69 and 7 genes which displayed the highest correlation ($PCC \geq 0.9$) in expression to *FLC* and *LFY*
351 respectively did not get enriched in any biological or molecular function (Fig. 6). These genes
352 were labeled as novel genes which can regulate the expression of other known floral regulators
353 during the flowering transition phase. For identification of genes regulated by *FLC* or *LFY*, node
354 connections were studied by filtering out genes connected with *FLC* or *LFY*. Table 5 lists genes
355 regulated and not regulated by *FLC* and *LFY* genes. 69 genes were found to be highly correlated
356 by *FLC* out of which 14 genes were regulated and 55 genes were non-regulated. Similarly, for
357 *LFY*, out of 7 genes 4 were regulated and 3 were non-regulated in the PPI network.

358

359 *3.7 Identification of alternative splicing classes in transcripts*

360

361 From the SpliceR analysis, we identified 1.23 transcripts per gene in the S7 sample and
362 approximately 1.27 transcripts for each sample compared against S7 in the “Against S7” sample
363 pairs, whereas 2.27 and 2.26 transcripts per gene were observed for the “Step analysis” sample
364 pairs. Looking at ESI-AS events in “Against S7”, we observed 0.1 ESI events per gene for S7 and
365 0.07 for each compared sample, whereas in the “Step analysis” sample pairs, 0.08 ESI events
366 per gene were observed for both samples.

367

368 **4. Discussion**

369 Recent progress in determination of DGE in RNA-Seq data using several bioinformatics tools
370 enabled easier identification of genes from samples. A number of tools for processing and
371 analysing RNA-Seq data have been developed. These include Cufflinks, edgeR, DESeq, RSEM
372 and others which claim accurate identification of DEGs. However, the accuracy can only be
373 determined by comparison of results obtained from several computational tools with those
374 obtained from published experimental studies. Using recently published tools for RNA-Seq data,
375 a comparative analysis of results obtained from Cufflinks-Cuffdiff2, DESeq and edgeR was
376 performed and analysis of intersection of DEGs from two or more tools was recommended in
377 order to obtain more robust results (Zhang *et al.*, 2014). In the current study, we have proposed
378 an approach for the identification of DEGs in *A. thaliana* RNA-Seq time-series datasets which
379 includes quality checking, adapter trimming, reference alignment, DEG analysis, alternative
380 splicing classification, DEG merging, GO enrichment and pathway analysis (Fig. 1). The first step
381 in identification of DEGs is to perform accurate genome alignment. Inaccurate parameters often
382 result in the generation of incorrect read counts from the data which could potentially result in

383 erroneous downstream processing. Previous investigations indicated use of default values for
384 processing RNA-Seq data (Klepikova *et al.*, 2015) which included similar minimum intron length
385 values of 70 nt for plants and mammals (Goodall and Filipowicz, 1990). However, mean, medium
386 and minimum intron length in *A. thaliana* and *O. sativa* was found to be much lower (Deutsch and
387 Long, 1999; Wang and Brendel, 2006) than the previously identified and established value of 70
388 nt. Therefore, to correctly identify DEGs from the data, custom parameter values were applied to
389 generate precise alignment of samples against the reference genome. The pipeline was
390 specifically focused on accurate processing and analysis of *A. thaliana* time-series datasets with
391 application in flower development. The analysis was performed by comparing S7 with other
392 samples in a way that S7 was treated as case and the comparing sample was treated as control
393 to identify DEGs. Additionally, analysis was also performed by progressively analyzing the case-
394 control samples in a stepwise manner (Table 2).

395

396 4.1 Known floral transition related genes and their interactions

397 In *A. thaliana* the transition to flowering is controlled through the regulation of certain genes of
398 which FLC and LFY are the most important (Deng *et al.*, 2011; Siriwardana and Lamb, 2012).
399 The transition process involves interaction of FLC with key genes such as *SUPPRESSOR OF*
400 *OVEREXPRESSON OF CONSTANS 1* (SOC1), *FLOWERING LOCUS T* (FT) and
401 *FLOWERING LOCUS D* (FLD) (Deng *et al.*, 2011). However, other repressors such as
402 *TERMINAL FLOWER1* (TFL1), *SHORT VEGETATIVE PHASE* (SVP), *TARGET OF EAT1/2*
403 (*TOE1/2*), *MADS AFFECTING FLOWERING1* (MAF1) to MAF5, *EMBRYONIC FLOWER1*
404 (EMF1) and EMF2 have also been found to be involved for regulation of flowering time (Hartmann
405 *et al.*, 2000; Piñeiro *et al.*, 2003; Ratcliffe *et al.*, 2003; Mathieu *et al.*, 2009; Hanano and Goto,
406 2011; Zhang *et al.*, 2015). *FLC* in particular binds to more than 500 target sites in the Arabidopsis
407 genome and regulates genes which function in developmental pathways (Deng *et al.*, 2011). One
408 of the known interactions of *FLC* is with *FRIGADA* (FRI) where both the genes interact to control
409 flowering time (Caicedo *et al.*, 2004). Expression of *FLC* is also regulated by PHD-polycomb
410 repressive complex 2 (PRC2) consisting of *VRN2*, Su(z)12 homologue and PHD finger proteins
411 *VIN3* and *VRN5*. Repression of *FLC* is mediated by cold-induced epigenetic silencing mechanism
412 during vernalization (De Lucia *et al.*, 2008). Like *FLC*, expression of *SVP* is controlled by the
413 trithorax (TrxG) protein called *BRAHMA* (BRM) (Li *et al.*, 2015). Similarly, the *EARLY BOLTING*
414 *IN SHORT DAYS* (EBS) protein participates in flowering time regulation by repressing the *FT*
415 protein and *CURLY LEAF* (CLF) protein represses the expression of *AGAMOUS LIKE* (AGL)
416 genes (Piñeiro *et al.*, 2003).

417 On the other hand, *LFY* acts as a positive regulator of *APETALA1* (AP1) and the
418 expression of *AP1* was observed late during photo-induction when examined during light
419 treatments (Hempel *et al.*, 1997). During shoot development, members of the *SQUAMOSA*
420 *PROMOTER BINDING PROTEIN-LIKE* (SPL) transcription factor family genes viz, *SPL9* and
421 *SPL15* have been known to control shoot maturation (Schwarz *et al.*, 2008). Like *LFY*, which is
422 a floral meristem identity protein, *CAULIFLOWER* (CAL), *FRUITFUL* (FUL), *AGAMOUS LIKE24*
423 (AGL24), SEP MADS box transcription factors *SEP3/4* and *LATE MERISTEM IDENTITY1/2*
424 (LMI1/2) are also floral meristem identity proteins in Arabidopsis (Kempin, Savidge and Yanofsky,
425 1995; Ferrándiz *et al.*, 2000; Ditta *et al.*, 2004; Castillejo, Romera-Branchat and Pelaz, 2005;

426 Saddic *et al.*, 2006; Gregis *et al.*, 2008; Pastore *et al.*, 2011; Siriwardana and Lamb, 2012).
427 Flowering during long days is also mediated by regulation of certain proteins. The transcription
428 factor CONSTANS (CO) plays a major role in the long-day pathway as this protein is
429 phosphorylated. Therefore it plays a major role in the abundance of the protein (Zhang *et al.*,
430 2015). Phosphorylated CO is preferentially degraded when CONSTITUTIVE
431 PHOTOMORPHOGENIC 1 (COP1) ubiquitin ligase complex is activated (Sarid-Krebs *et al.*,
432 2015). Furthermore, COP1 protein has also been known to interact with the Arabidopsis
433 cryptochromes (CRY1 and CRY2) through C-terminal domains (CCT) (Yang, Tang and
434 Cashmore, 2001). During floral transition certain genes such as *FLOWERING PROMOTIVE*
435 *FACTOR1 (FPF1)* are expressed in apical meristems and are involved in the GA-dependent
436 signaling pathway which regulates the GA response in apical meristem (Kania *et al.*, 1997).

437

438 4.2 Overlapping of Cuffdiff with DESeq and edgeR genes

439 The key step of RNA-Seq data analysis is to identify DEGs using appropriate statistical models.
440 Once the FPKM counts from the sequencing reads were obtained, these were used for finding
441 DEGs using Cuffdiff, DESeq and edgeR. The statistical model in Cuffdiff which is used to evaluate
442 the changes in expression assumes that the number of reads produced by each transcript is
443 proportional to its abundance, although it fluctuates because of biological variability between the
444 replicates and technical variability during sequencing and library preparation. DESeq, on the other
445 hand, allows the user to supply multiple as well as no replicates. DESeq is highly useful when no
446 replicates are present in the datasets. The statistical model in DESeq uses a blind dispersion
447 method that is particularly useful with no replicates where the outlier values cannot be captured
448 during dispersion estimation. On the other hand, edgeR uses both the generalized linear model
449 (glm) and classical empirical Bayes methods, which are used for estimation of gene-specific
450 biological variations even with those datasets having a minimal level of biological variations.
451 Usage of Cuffdiff, DESeq and edgeR methods increase statistical power and help in rationale
452 comparison and thus confirming the suitability of the results. Therefore, DE analysis has been
453 carried out for further comparison in the present investigation.

454 Results show that both Cuffdiff and edgeR displayed significant numbers of differentially
455 expressed genes in S7-S10, S7-S12 and S7-S13 (Table 3). Overlapping of genes can be
456 visualized by Venn diagrams constructed for transition phase samples (Fig. 7). We see in Fig. 7
457 the intersection of Cuffdiff-DESeq-edgeR, Cuffdiff-DESeq and Cuffdiff-edgeR decreases in “Step
458 analysis” sample pairs as compared to “Against S7” sample pairs. edgeR additionally displayed
459 a greater number of DE genes in S7-S14, S7-S15 and S7-S16 which are not notably identified by
460 Cuffdiff or edgeR. On the contrary, Cuffdiff displays the maximum number of DE genes from “Step
461 analysis” results as compared to DESeq and edgeR. By comparing the results of Cuffdiff with
462 DESeq and edgeR, we clearly observed that the overlap from Cuffdiff-edgeR was more significant
463 than Cuffdiff-DESeq or DESeq-edgeR. This difference can be clearly observed in “Step analysis”
464 for S10-S11 (Fig. 7g) where 1347 genes were found to be common for Cuffdiff-edgeR as
465 compared to 4 for Cuffdiff-DESeq. Thus, the total number of common genes was significantly
466 reduced for Cuffdiff-DESeq-edgeR intersection which is primarily due to a smaller gene count in
467 Cuffdiff-DESeq. Thus only 1% of the genes were found to be common for Cuffdiff-DESeq-edgeR

468 confirming that the decrease in the overlap is mostly due to DESeq results. A significant number
469 of genes were found to have an overlap in S7-S13 pairs as shown in Fig. 7d.

470

471 4.3 Comparison of profiles of differentially expressed genes

472 *FLC* and *LFY* are the two most important genes which regulates transition to flowering and are
473 widely known to play major role in flower development. Using PCC metric, we evaluated similarity
474 in expression profiles of *FLC* and *LFY* obtained in this study and compared with profiles obtained
475 from Klepikova et al (Klepikova *et al.*, 2015). Consistent with the experimental findings, our results
476 as shown in Fig. 8 (a, b, c and d) displays higher mean PCC of 0.86 and 0.88 for *FLC* and *LFY*,
477 respectively which is consistent with published results (Michaels, 1999; Klepikova *et al.*, 2015).
478 From our results, *SVP* has been shown to be highly expressed in all samples during the transition
479 phase which was consistent with the results from Klepikova et al. *SOC1*, on the other hand,
480 showed a five-fold increase in expression during S7 to S10. Contrary to these results from
481 Klepikova et al., our results showed consistently higher expression of *SOC1* in transition phase
482 samples from S7 to S14. Similarly, where the expression of *FLD* showed increase in later stages
483 of floral induction (Klepikova *et al.*, 2015), expression of *FLD* has been found to be consistently
484 higher in all the stages of flower development in our results. However, we found correlation in
485 expression patterns of genes from the *SPL* gene family between our results and those of the
486 results of Klepikova et al. and others. . *SPL3*, *SPL4* and *SPL5* showed decrease in expression
487 whereas *SPL9* and *SPL15* showed increase in expression which was consistent with published
488 results (Schwarz *et al.*, 2008). In “Against S7” sample pairs, expression of *AP1* showed increase
489 in late transition phase stages (i.e S13 and S14) consistent with Hempel et al. (Hempel *et al.*,
490 1997) and Ferrándiz et al. (Ferrándiz *et al.*, 2000) whereas expression of *AP1* was observed in
491 S14 only (Klepikova *et al.*, 2015). Experimental results show *COP1* interact with *CRY1* and *CRY2*
492 proteins in apical shoots (Yang, Tang and Cashmore, 2001; Sarid-Krebs *et al.*, 2015).
493 Additionally, increased expression of *FPF1* (from Cuffdiff and edgeR) has also been observed in
494 our results which is consistent with the published results during the transition phase (Kania *et al.*,
495 1997). Expression profiles of *FPF1*, *SPL9* and *SPL15* have found to be correlated with *LFY*
496 expression (Fig. 8b, d, e and f).

497 Expression of experimental genes were roughly categorized into 4 different types, namely,
498 (1) genes similar in expression to *FLC* displaying decrease in expression from S7-S8 to S7-S16
499 and S15-S16, (2) genes displaying increase in expression from S7-S8 to S7-S16 and S15-S16,
500 (3) genes having variable expression between 1 and 0.6, and (4) genes showing increase in
501 expression followed by a decrease in expression. Genes displaying higher PCC to *FLC* are *SNZ*,
502 *SMZ*, *TOE2* and those having higher PCC to *LFY* are *FPF1*, *CAL*, *AP1*, *SOC1*, *SPL9*, *SPL15*,
503 *EMF2*, *CO*, *AGL71*, *MAF2* and *MAF3*. Genes having lower PCC are *LHP1*, *EMF1*, *EBS*, *VRN2*,
504 *CRY1*, *LD*, *FRI*, *FIE*, *FLD*, *BRM*, *COP1*, *CLF*, *SVP*, *PHYB*, *MAF1*, *CRY2*, *MAF4* and *MAF5*. *TFL1*
505 showed increase in expression in S7-S14 whereas *FT* showed increase in expression in S7-S11
506 which was followed by sudden drop in expression values for both *TFL1* and *FT*.

507

508 4.4 Comparison of expression profiles of cell-cycle related genes

509

510 Expression patterns of cyclin dependent kinases (CDKs) obtained from the pipeline were also
511 compared to obtain degree of variation between those obtained from Klepikova et al. (Klepikova

512 *et al.*, 2015) which included CDKA, CDKB, CDKC, CDKD, CDKP, CDPT and cyclin genes. Results
513 from the comparison shows that most of the CDK genes exhibited moderate correlation with an
514 average ranging between 0.60 – 0.70. While some of the CDKs such as *CYCA3;4*, *CYCB1;1*,
515 *CYCC1;2*, *CYCD1;1*, *CYCD2;1*, *CYCD6;1*, *CYCD4;2*, *CYCP3;1*, *CYCP3;2*, *CDKE;1*, *CDT1A*,
516 *CYCJ18*, *CYL1* and *CYCH;1* displayed particularly higher correlation above 0.90, some displayed
517 particularly poor correlation which included *CKS1*, *CDKB1;1*, *CDKB1;2*, *CDKA;1/CBC2*,
518 *CYCA1;1*, *CYCB3;1*, *CYCB1;5*, *CYCB1;2/CYC1BAT*, *CYCC1;1*, *CYCD4;1*, *CYCP4;1*, *CYCT1;4*,
519 *CDKD;3/CDKD1;3*, *CDKD;1/CDKD1;1* and *CYCL1/RCY1* ranging between 0.20 – 0.60.
520 Correlation analysis shows that some of the cyclin genes exhibit greater similarity, whereas the
521 majority of CDKB and CDKD genes exhibits greater dissimilarity. Expression profiles of poorly
522 correlated genes such as *CKS1* obtained from Klepikova *et al.* shows single peak at M5 (Day-11)
523 whereas two peaks were found on Day-9 and Day-12 in samples S7-S9 and S7-S12.
524 Experimental results clearly show that *CKS1* was constitutively expressed during mitotic and
525 endoreduplication cycles (Jacqmard *et al.*, 1999). This supports the hypothesis that the gene
526 should be highly expressed during the flowering transition phase. Similarly, *CDKB1;1* has also
527 been found to be highly expressed in shoot meristem in *A. thaliana* (Skylar, Matsuwaka and Wu,
528 2013) which supports the evidence of multiple peaks observed in our results.

529

530 4.5 Protein-protein interaction network analysis of differentially expressed genes

531

532 Results of PPI network analysis shows most of the DEGs during the transition phase regulate
533 other DEGs which provide induced resistance and protection against external factors such as
534 stress, pathogens, herbivores, temperature variations, etc. A recent study on the relationship of
535 glucosinolates to flowering in *A. thaliana* suggests that presence of the *MAM1* gene affects
536 glucosinolate accumulation and flowering time in the absence of *APOP2* and *APOP3* genes and
537 leads to production of C3 glucosinolates (Jensen *et al.*, 2015). Results from the PPI network
538 analysis clearly show that *MAM1* regulates several other genes in glucosinolates and displays a
539 high expression profile correlation of 0.75 to *FLC* which supports the hypothesis of glucosinolate
540 production and protection during flowering phase. Glucosinolates are sulphur and nitrogen-rich
541 chemical compounds in plants that provide defense against pathogens and herbivores by forming
542 a toxic compound upon herbivore attack when the cell wall is ruptured (Jensen *et al.*, 2015;
543 Mohammadin *et al.*, 2017). Glucosinolates play a crucial role in flowering time regulation during
544 transition from vegetative to reproductive phase and also provide protection from herbivores and
545 pathogens for the plant's vegetative and generative tissues during the transition phase. Therefore,
546 differential expression of glucosinolates during the transition phase becomes essential.

547 We also identified genes responsible for flowering and involved in flowering and in
548 regulation of flower development from the clustering of 690 expressed genes. Expression profiles
549 of these associated genes were constructed to observe similarities and differences among profiles
550 of experimental genes (Fig. 4). *SPT1*, *RDR6*, *SRS2*, *SAP*, *SOC1* clearly showed increase in
551 expression at S12-S13 in "Flower Development" whereas it showed a "zig-zag" pattern of
552 expression in "Against S7". In contrast, genes such as *NGA1* and *NAC054* displayed a decrease
553 in expression from 0.5 in S7-S13 to 0.1 in S7-S14 and a sudden increase to 1 in S7-S16. Genes
554 such as *SOC1* and *F-ATMBP* showed an increase in expression. Expression profiles of genes in
555 "Step Analysis" showed a distinct peak at S12-S13 which strongly indicates that genes associated

556 with flowering and flower development show identical expression profiles and are expressed only
557 during transition phase.

558

559 **Conclusions**

560

561 In this study, we conducted rigorous investigation and analysis of apical shoot meristem time-
562 series dataset obtained from *A. thaliana*. We constructed a pipeline for identification of
563 differentially expressed genes from overlap of three tools and identified 690 genes. Their
564 functional enrichment was conducted for identification of genes associated with highly enriched
565 biological processes. We also constructed expression profiles of genes enriched in flowering and
566 in the regulation of flower development. We observed that some of these genes displayed distinct
567 expression profiles when compared to those displayed by already known experimental genes
568 which are commonly regulated during floral transition. Additionally, PPI network analysis was
569 conducted to identify prevalent interactors and associated functions during the transition phase.
570 76 novel genes were identified as showing stronger regulation in the network and displaying the
571 highest correlation in expression with *FLC* and *LFY* genes. Further experiments will validate and
572 confirm gene regulation and specific PPIs of the novel genes obtained from the current analysis.

573

574 **References**

575 Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data.',
576 *Genome biology*, 11(10), p. R106. doi: 10.1186/gb-2010-11-10-r106.

577 Anders, S., Pyl, P. T. and Huber, W. (2015) 'HTSeq-A Python framework to work with high-
578 throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. doi:
579 10.1093/bioinformatics/btu638.

580 Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data.*,
581 <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. doi: citeulike-article-id:11583827.

582 Barth, C. and Jander, G. (2006) 'Arabidopsis myrosinases TGG1 and TGG2 have redundant
583 function in glucosinolate breakdown and insect defense', *Plant Journal*, 46(4), pp. 549–562. doi:
584 10.1111/j.1365-313X.2006.02716.x.

585 Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the false discovery rate: a practical and
586 powerful approach to multiple testing', *Journal of the Royal Statistical Society*, pp. 289–300. doi:
587 10.2307/2346101.

588 Bi, Y. and Davuluri, R. V (2013) 'NPEBseq: nonparametric empirical bayesian-based procedure
589 for differential expression analysis of RNA-seq data.', *BMC bioinformatics*, 14(1), p. 262. doi:
590 10.1186/1471-2105-14-262.

591 Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W. H.,
592 Pagès, F., Trajanoski, Z. and Galon, J. (2009) 'ClueGO: A Cytoscape plug-in to decipher
593 functionally grouped gene ontology and pathway annotation networks', *Bioinformatics*, 25(8),
594 pp. 1091–1093. doi: 10.1093/bioinformatics/btp101.

595 Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. and Purugganan, M. D. (2004)
596 'Epistatic interaction between Arabidopsis *FRI* and *FLC* flowering time genes generates a

597 latitudinal cline in a life history trait.’, *Proceedings of the National Academy of Sciences of the*
598 *United States of America*, 101(44), pp. 15670–5. doi: 10.1073/pnas.0406232101.

599 Castillejo, C., Romera-Branchat, M. and Pelaz, S. (2005) ‘A new role of the Arabidopsis
600 SEPALLATA3 gene revealed by its constitutive expression’, *Plant Journal*, 43(4), pp. 586–596.
601 doi: 10.1111/j.1365-313X.2005.02476.x.

602 Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P.,
603 Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels,
604 K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. and D’Eustachio,
605 P. (2014) ‘The Reactome pathway knowledgebase’, *Nucleic Acids Research*, 42(D1). doi:
606 10.1093/nar/gkt1102.

607 Deng, W., Ying, H., Helliwell, C. A., Taylor, J. M., Peacock, W. J. and Dennis, E. S. (2011)
608 ‘FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of
609 Arabidopsis.’, *Proceedings of the National Academy of Sciences of the United States of*
610 *America*, 108(16), pp. 6680–6685. doi: 10.1073/pnas.1103175108.

611 Deutsch, M. and Long, M. (1999) ‘Intron-exon structures of eukaryotic model organisms’,
612 *Nucleic Acids Research*, 27(15), pp. 3219–3228. doi: 10.1093/nar/27.15.3219.

613 Ditta, G., Pinyopich, A., Robles, P., Pelaz, S. and Yanofsky, M. F. (2004) ‘The SEP4 gene of
614 Arabidopsis thaliana functions in floral organ and meristem identity’, *Current Biology*, 14(21), pp.
615 1935–1940. doi: 10.1016/j.cub.2004.10.028.

616 Ferrándiz, C., Gu, Q., Martienssen, R. and Yanofsky, M. F. (2000) ‘Redundant regulation of
617 meristem identity and plant architecture by FRUITFULL, APETALA1 and CAULIFLOWER.’,
618 *Development (Cambridge, England)*, 127(4), pp. 725–734. doi: 10.1046/j.1365-
619 313x.1999.00442.x.

620 Goodall, G. J. and Filipowicz, W. (1990) ‘The minimum functional length of pre-mRNA introns in
621 monocots and dicots’, *Plant Molecular Biology*, 14(5), pp. 727–733. doi: 10.1007/BF00016505.

622 Gregis, V., Sessa, A., Colombo, L. and Kater, M. M. (2008) ‘AGAMOUS-LIKE24 and SHORT
623 VEGETATIVE PHASE determine floral meristem identity in Arabidopsis’, *Plant Journal*, 56(6),
624 pp. 891–902. doi: 10.1111/j.1365-313X.2008.03648.x.

625 Gupta, V., Estrada, A. D., Blakley, I., Reid, R., Patel, K., Meyer, M. D., Andersen, S. U., Brown,
626 A. F., Lila, M. A. and Loraine, A. E. (2015) ‘RNA-Seq analysis and annotation of a draft
627 blueberry genome assembly identifies candidate genes involved in fruit ripening, biosynthesis of
628 bioactive compounds, and stage-specific alternative splicing.’, *GigaScience*, 4(1), p. 5. doi:
629 10.1186/s13742-015-0046-9.

630 Hanano, S. and Goto, K. (2011) ‘Arabidopsis TERMINAL FLOWER1 is involved in the
631 regulation of flowering time and inflorescence development through transcriptional repression.’,
632 *The Plant cell*, 23(9), pp. 3172–84. doi: 10.1105/tpc.111.088641.

633 Hardcastle, T. J. and Kelly, K. A. (2010) ‘baySeq: empirical Bayesian methods for identifying
634 differential expression in sequence count data.’, *BMC bioinformatics*, 11(1), p. 422. doi:
635 10.1186/1471-2105-11-422.

636 Hartmann, U., Höhmann, S., Nettesheim, K., Wisman, E., Saedler, H. and Huijser, P. (2000)
637 ‘Molecular cloning of SVP: A negative regulator of the floral transition in Arabidopsis’, *Plant*

638 *Journal*, 21(4), pp. 351–360. doi: 10.1046/j.1365-313X.2000.00682.x.

639 Hempel, F. D., Weigel, D., Mandel, M. a, Ditta, G., Zambryski, P. C., Feldman, L. J. and
640 Yanofsky, M. F. (1997) 'Floral determination and expression of floral regulatory genes in
641 *Arabidopsis*.' *Development (Cambridge, England)*, 124(19), pp. 3845–3853.

642 Jacquard, A., De Veylder, L., Segers, G., De Almeida Engler, J., Bernier, G., Van Montagu, M.
643 and Inze, D. (1999) 'Expression of CKS1At in *Arabidopsis thaliana* indicates a role for the
644 protein in both the mitotic and the endoreduplication cycle', *Planta*, 207(4), pp. 496–504. doi:
645 10.1007/s004250050509.

646 Jensen, L. M., Jepsen, H. S. K., Halkier, B. A., Kliebenstein, D. J. and Burow, M. (2015) 'Natural
647 variation in cross-talk between glucosinolates and onset of flowering in *Arabidopsis*', *Frontiers in*
648 *Plant Science*, 6(SEPTEMBER), p. 697. doi: 10.3389/fpls.2015.00697.

649 Kania, T., Russenberger, D., Peng, S., Apel, K. and Melzer, S. (1997) 'FPF1 promotes flowering
650 in *Arabidopsis*.' *The Plant cell*, 9(8), pp. 1327–38. doi: 10.1105/tpc.9.8.1327.

651 Kempin, S. a, Savidge, B. and Yanofsky, M. F. (1995) 'Molecular basis of the cauliflower
652 phenotype in *Arabidopsis*.' *Science (New York, N.Y.)*, 267(5197), pp. 522–525. doi:
653 10.1126/science.7824951.

654 Kim, D., Perteza, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S. L. (2013) 'TopHat2:
655 accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.',
656 *Genome biology*, 14(4), p. R36. doi: 10.1186/gb-2013-14-4-r36.

657 Klepikova, A. V, Logacheva, M. D., Dmitriev, S. E. and Penin, A. A. (2015) 'RNA-seq analysis of
658 an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation',
659 *BMC Genomics*, 16. doi: 10.1186/s12864-015-1688-9.

660 Koornneef, M., Hanhart, C. J. and van der Veen, J. H. (1991) 'A genetic and physiological
661 analysis of late flowering mutants in *Arabidopsis thaliana*.' *Molecular & general genetics : MGG*,
662 229(1), pp. 57–66. doi: 10.1007/BF00264213.

663 Krishnakumar, V., Hanlon, M. R., Contrino, S., Ferlanti, E. S., Karamycheva, S., Kim, M.,
664 Rosen, B. D., Cheng, C. Y., Moreira, W., Mock, S. A., Stubbs, J., Sullivan, J. M., Krampis, K.,
665 Miller, J. R., Micklem, G., Vaughn, M. and Town, C. D. (2015) 'Araport: The *Arabidopsis*
666 Information Portal', *Nucleic Acids Research*, 43(D1), pp. D1003–D1009. doi:
667 10.1093/nar/gku1200.

668 Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, E. L., Bohler, A., Mélius, J.,
669 Waagmeester, A., Sinha, S. R., Miller, R., Coort, S. L., Cirillo, E., Smeets, B., Evelo, C. T. and
670 Pico, A. R. (2015) 'WikiPathways: capturing the full diversity of pathway knowledge', *Nucleic*
671 *Acids Research*, 44(October 2015), p. gkv1024. doi: 10.1093/nar/gkv1024.

672 Langmead, B. and Salzberg, S. L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nat*
673 *Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

674 Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M. G., Haag, J. D.,
675 Gould, M. N., Stewart, R. M. and Kendziorski, C. (2013) 'EBSeq: An empirical Bayes
676 hierarchical model for inference in RNA-seq experiments', *Bioinformatics*, 29(8), pp. 1035–1043.
677 doi: 10.1093/bioinformatics/btt087.

678 Li, B. and Dewey, C. N. (2011) 'RSEM: accurate transcript quantification from RNA-Seq data
679 with or without a reference genome.', *BMC bioinformatics*, 12(1), p. 323. doi: 10.1186/1471-
680 2105-12-323.

681 Li, C., Chen, C., Gao, L., Yang, S., Nguyen, V., Shi, X., Siminovitch, K., Kohalmi, S. E., Huang,
682 S., Wu, K., Chen, X. and Cui, Y. (2015) 'The Arabidopsis SWI2/SNF2 Chromatin Remodeler
683 BRAHMA Regulates Polycomb Function during Vegetative Development and Directly Activates
684 the Flowering Repressor Gene SVP', *PLoS Genetics*, 11(1). doi: 10.1371/journal.pgen.1004944.

685 Liu, R., Loraine, A. E. and Dickerson, J. A. (2014) 'Comparisons of computational methods for
686 differential alternative splicing detection using RNA-seq in plant systems.', *BMC bioinformatics*,
687 15(1), p. 364. doi: 10.1186/s12859-014-0364-4.

688 De Lucia, F., Crevillen, P., Jones, A. M. E., Greb, T. and Dean, C. (2008) 'A PHD-polycomb
689 repressive complex 2 triggers the epigenetic silencing of FLC during vernalization.',
690 *Proceedings of the National Academy of Sciences of the United States of America*, 105(44), pp.
691 16831–16836. doi: 10.1073/pnas.0808687105.

692 Martin, M. (2011) 'Cutadapt removes adapter sequences from high-throughput sequencing
693 reads', *EMBnet.journal*, 17(1), p. 10. doi: 10.14806/ej.17.1.200.

694 Mathieu, J., Yant, L. J., Mürdter, F., Küttner, F. and Schmid, M. (2009) 'Repression of flowering
695 by the miR172 target SMZ', *PLoS Biology*, 7(7). doi: 10.1371/journal.pbio.1000148.

696 Michaels, S. D. (1999) 'FLOWERING LOCUS C Encodes a Novel MADS Domain Protein That
697 Acts as a Repressor of Flowering', *THE PLANT CELL ONLINE*, 11(5), pp. 949–956. doi:
698 10.1105/tpc.11.5.949.

699 Mohammadin, S., Nguyen, T.-P., van Weij, M. S., Reichelt, M. and Schranz, M. E. (2017)
700 'Flowering Locus C (FLC) Is a Potential Major Regulator of Glucosinolate Content across
701 Developmental Stages of *Aethionema arabicum* (Brassicaceae)', *Frontiers in Plant Science*, 8.
702 doi: 10.3389/fpls.2017.00876.

703 Montojo, J., Zuberi, K., Rodriguez, H., Kazi, F., Wright, G., Donaldson, S. L., Morris, Q. and
704 Bader, G. D. (2010) 'GeneMANIA cytoscape plugin: Fast gene function predictions on the
705 desktop', *Bioinformatics*, 26(22), pp. 2927–2928. doi: 10.1093/bioinformatics/btq562.

706 Naur, P., Petersen, B. L., Mikkelsen, M. D., Bak, S., Rasmussen, H., Olsen, C. E. and Halkier,
707 B. A. (2003) 'CYP83A1 and CYP83B1, two nonredundant cytochrome P450 enzymes
708 metabolizing oximes in the biosynthesis of glucosinolates in *Arabidopsis*.', *Plant physiology*,
709 133(1), pp. 63–72. doi: 10.1104/pp.102.019240.

710 Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. and Kanehisa, M. (1999) 'KEGG: Kyoto
711 encyclopedia of genes and genomes', *Nucleic Acids Research*, pp. 29–34. doi:
712 10.1093/nar/27.1.29.

713 Ostell, J. and McEntyre, J. (2007) 'The NCBI Handbook', *NCBI Bookshelf*, pp. 1–8. doi:
714 10.4016/12837.01.

715 Pastore, J. J., Limpuangthip, A., Yamaguchi, N., Wu, M. F., Sang, Y., Han, S. K., Malaspina, L.,
716 Chavdaroff, N., Yamaguchi, A. and Wagner, D. (2011) 'LATE MERISTEM IDENTITY2 acts
717 together with LEAFY to activate APETALA1', *Development*, 138(15), pp. 3189–3198. doi:
718 10.1242/dev.063073.

- 719 Perteua, M., Perteua, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T. and Salzberg, S. L.
720 (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads',
721 *Nature Biotechnology*, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.
- 722 Piñeiro, M., Gómez-Mena, C., Schaffer, R., Martínez-Zapater, J. M. and Coupland, G. (2003)
723 'EARLY BOLTING IN SHORT DAYS is related to chromatin remodeling factors and regulates
724 flowering in Arabidopsis by repressing FT.', *The Plant cell*, 15(7), pp. 1552–1562. doi:
725 10.1105/tpc.012153.
- 726 Ratcliffe, O. J., Kumimoto, R. W., Wong, B. J. and Riechmann, J. L. (2003) 'Analysis of the
727 Arabidopsis MADS AFFECTING FLOWERING gene family: MAF2 prevents vernalization by
728 short periods of cold.', *The Plant cell*, 15(5), pp. 1159–69. doi: 10.1105/tpc.009506.mous.
- 729 Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) 'edgeR: a Bioconductor package for
730 differential expression analysis of digital gene expression data.', *Bioinformatics (Oxford,*
731 *England)*, 26(1), pp. 139–40. doi: 10.1093/bioinformatics/btp616.
- 732 Saddic, L. A., Huvermann, B., Bezhani, S., Su, Y., Winter, C. M., Kwon, C. S., Collum, R. P. and
733 Wagner, D. (2006) 'The LEAFY target LMI1 is a meristem identity regulator and acts together
734 with LEAFY to regulate expression of CAULIFLOWER', *Development*, 133(9), pp. 1673–1682.
735 doi: 10.1242/dev.02331.
- 736 Sarid-Krebs, L., Panigrahi, K. C. S., Fornara, F., Takahashi, Y., Hayama, R., Jang, S., Tilmes,
737 V., Valverde, F. and Coupland, G. (2015) 'Phosphorylation of CONSTANS and its COP1-
738 dependent degradation during photoperiodic flowering of Arabidopsis', *Plant Journal*, 84(3), pp.
739 451–463. doi: 10.1111/tpj.13022.
- 740 Sayers, E. W., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K., Chetvernin, V., Church, D.
741 M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L. Y., Helmberg, W., Kapustin, Y.,
742 Landsman, D., Lipman, D. J., Madden, T. L., Maglott, D. R., Miller, V., Mizrachi, I., Ostell, J.,
743 Pruitt, K. D., Schuler, G. D., Sequeira, E., Sherry, S. T., Shumway, M., Sirotkin, K., Souvorov,
744 A., Starchenko, G., Tatusova, T. A., Wagner, L., Yaschenko, E. and Ye, J. (2009) 'Database
745 resources of the National Center for Biotechnology Information', *Nucleic Acids Research*,
746 37(SUPPL. 1). doi: 10.1093/nar/gkn741.
- 747 Schwarz, S., Grande, A. V., Bujdoso, N., Saedler, H. and Huijser, P. (2008) 'The microRNA
748 regulated SBP-box genes SPL9 and SPL15 control shoot maturation in Arabidopsis', *Plant*
749 *Molecular Biology*, 67(1–2), pp. 183–195. doi: 10.1007/s11103-008-9310-z.
- 750 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N.,
751 Schwikowski, B. and Ideker, T. (2003) 'Cytoscape: A software Environment for integrated
752 models of biomolecular interaction networks', *Genome Research*, 13(11), pp. 2498–2504. doi:
753 10.1101/gr.1239303.
- 754 Siriwardana, N. S. and Lamb, R. S. (2012) 'The poetry of reproduction: The role of LEAFY in
755 Arabidopsis thaliana flower formation', *International Journal of Developmental Biology*, pp. 207–
756 221. doi: 10.1387/ijdb.113450ns.
- 757 Skylar, A., Matsuwaka, S. and Wu, X. (2013) 'ELONGATA3 is required for shoot meristem cell
758 cycle progression in Arabidopsis thaliana seedlings', *Developmental Biology*, 382(2), pp. 436–
759 445. doi: 10.1016/j.ydbio.2013.08.008.
- 760 Tarazona, S., García, F., Ferrer, A., Dopazo, J. and Conesa, A. (2012) 'NOIseq: a RNA-seq

761 differential expression method robust for sequencing depth biases', *EMBnet.journal*, 17(B), p.
762 18. doi: 10.14806/ej.17.B.265.

763 Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S.
764 L., Rinn, J. L. and Pachter, L. (2012) 'Differential gene and transcript expression analysis of
765 RNA-seq experiments with TopHat and Cufflinks.', *Nature protocols*, 7(3), pp. 562–78. doi:
766 10.1038/nprot.2012.016.

767 Vitting-Seerup, K., Porse, B. T., Sandelin, A. and Waage, J. (2014) 'spliceR: an R package for
768 classification of alternative splicing and prediction of coding potential from RNA-seq data.', *BMC*
769 *bioinformatics*, 15, p. 81. doi: 10.1186/1471-2105-15-81.

770 Wang, B.-B. and Brendel, V. (2006) 'Genomewide comparative analysis of alternative splicing in
771 plants', *Pnas*, 103(18), p. 602039103. doi: 10.1073/pnas.0602039103.

772 Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2009) 'DEGseq: An R package for
773 identifying differentially expressed genes from RNA-seq data', *Bioinformatics*, 26(1), pp. 136–
774 138. doi: 10.1093/bioinformatics/btp612.

775 Yang, H. Q., Tang, R. H. and Cashmore, A. R. (2001) 'The signaling mechanism of Arabidopsis
776 CRY1 involves direct interaction with COP1.', *The Plant cell*, 13(12), pp. 2573–87. doi:
777 10.1105/tpc.010367.

778 Zhang, B., Wang, L., Zeng, L., Zhang, C. and Ma, H. (2015) 'Arabidopsis TOE proteins convey
779 a photoperiodic signal to antagonize CONSTANS and regulate flowering time', *Genes and*
780 *Development*, 29(9), pp. 975–987. doi: 10.1101/gad.251520.114.

781 Zhang, Z. H., Jhaveri, D. J., Marshall, V. M., Bauer, D. C., Edson, J., Narayanan, R. K.,
782 Robinson, G. J., Lundberg, A. E., Bartlett, P. F., Wray, N. R. and Zhao, Q. Y. (2014) 'A
783 comparative study of techniques for differential expression analysis on RNA-seq data', *PLoS*
784 *ONE*, 9(8). doi: 10.1371/journal.pone.0103207.

785 Zhao, S. and Zhang, B. (2015) 'A comprehensive evaluation of ensembl, RefSeq, and UCSC
786 annotations in the context of RNA-seq read mapping and gene quantification', *BMC Genomics*,
787 16(1), p. 97. doi: 10.1186/s12864-015-1308-8.

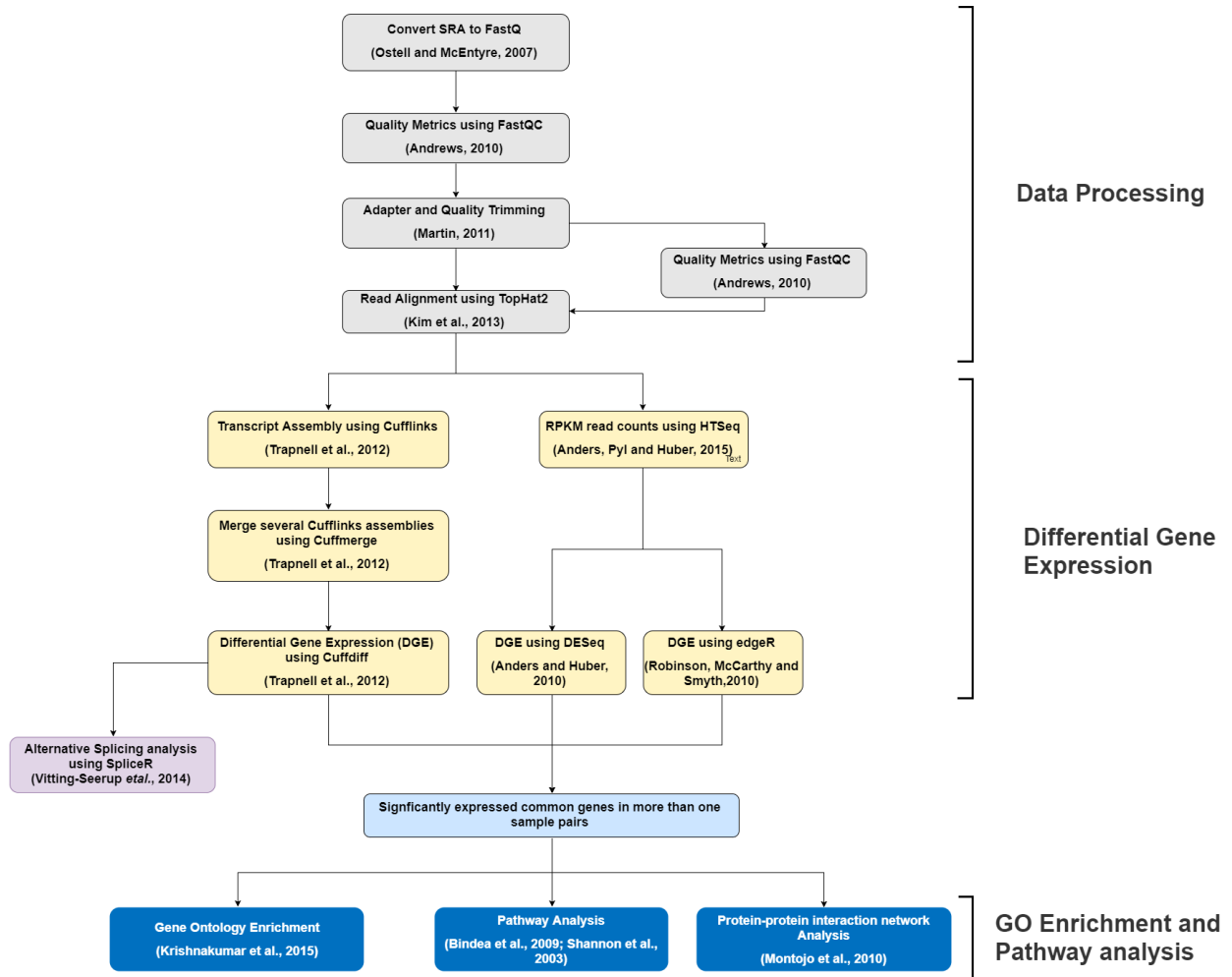
788

789

790

Figure and Tables

791



792

793

794

795

796

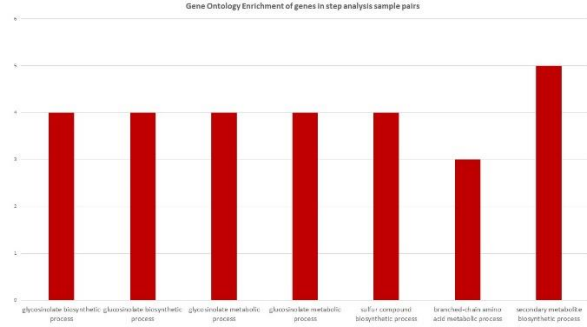
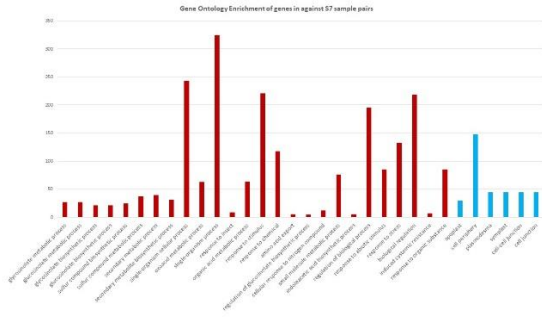
797

798

a

b

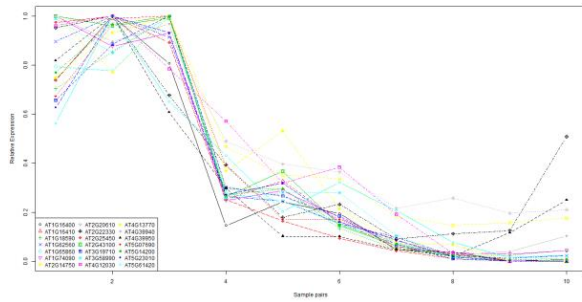
Fig. 1: Flowchart of the proposed RNA-Seq data analysis pipeline. The workflow is divided into three stages namely, data processing, differential gene expression and GO enrichment & network interaction analysis



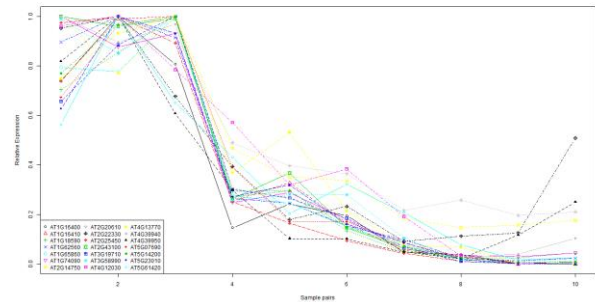
799 Fig. 2: GO enrichment functional classification results. The above figure illustrates classification of genes
 800 into several functional categories obtained using the Araport tool (Krishnakumar *et al.*, 2015). The genes
 801 obtained from Cuffdiff (Trapnell *et al.*, 2012), DESeq (Anders and Huber, 2010) and edgeR (Robinson,
 802 McCarthy and Smyth, 2010) are used for GO classification. (a) Illustrates GO classification results of
 803 common genes obtained from Cuffdiff, DESeq and edgeR for genes obtained from "Against S7" sample
 804 pairs. The bar chart contains two different gene ontologies. Bars colored in red represent genes enriched
 805 in "Molecular Function" whereas bars colored in blue represent genes enriched in "Cellular Component",
 806 (b) GO classification results of common genes obtained from Cuffdiff, DESeq and edgeR for genes
 807 obtained from "Step Analysis" sample pairs. The bar chart only contains genes enriched in "Molecular
 808 Function" which are colored in red. Other two gene ontologies were not observed for the gene set
 809 provided from these sample pairs.

810
 811

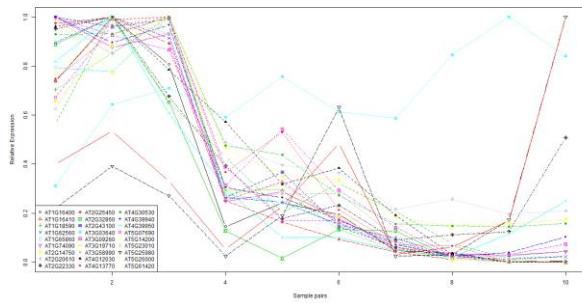
a



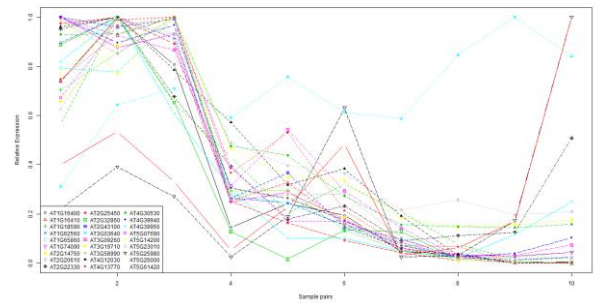
b



c

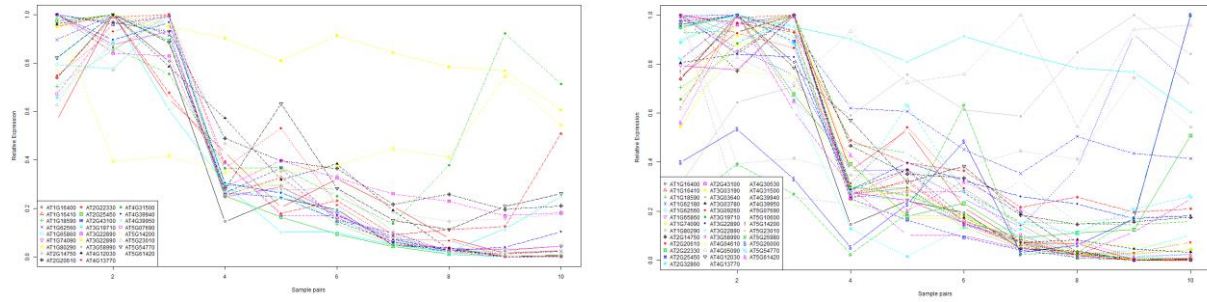


d

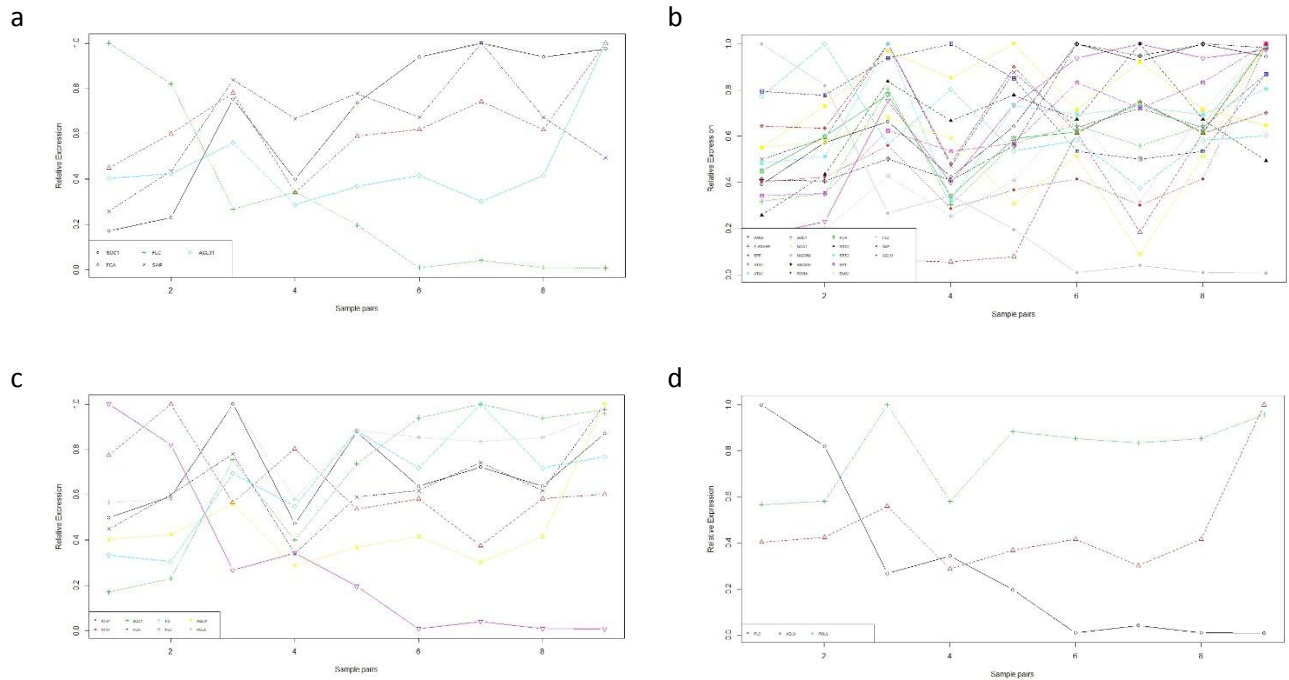


e

f

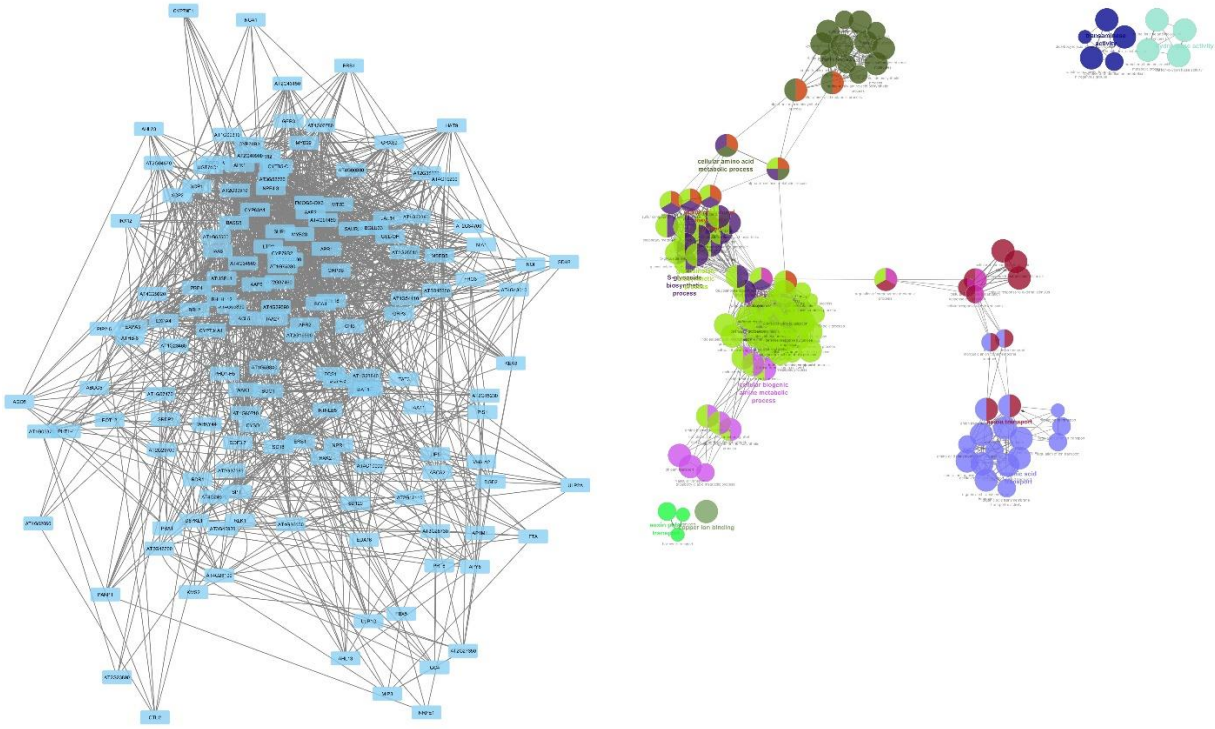


812 Fig. 3: Expression profiles of common genes from Cuffdiff-DESeq-edgeR overlap. The above graphs
 813 show expression profiles of genes enriched in glucosinolate biosynthetic process, glycosinolate
 814 biosynthetic process, glucosinolate metabolic process, glycosinolate metabolic process, sulfur compound
 815 biosynthetic process and sulfur compound metabolic process. (a) to (f) shows expression profiles of gene
 816 clusters in "Against S7" sample pairs. Common genes were obtained by overlapping DEGs from Cuffdiff
 817 (Trapnell *et al.*, 2012), DESeq (Anders and Huber, 2010) and edgeR (Robinson, McCarthy and Smyth,
 818 2010) and expressed in more than one sample pairs.
 819



820 Fig. 4: Expression profiles of flowering genes. The above figure illustrates relative expression profiles of
 821 genes involved in flowering, flower development, regulation of flower development and negative
 822 regulation of flower development. (a), (b), (c) and (d) shows relative expression of genes in "Against S7"
 823 sample pairs
 824

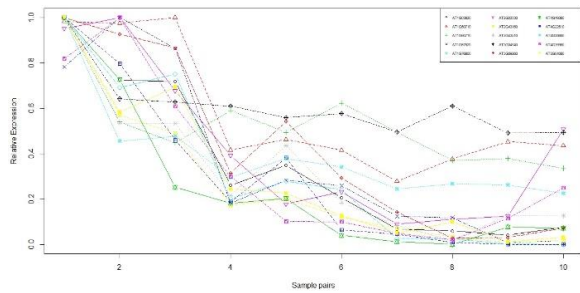
a b



825 Figure 5: Protein-protein interaction network and Functional grouped network (FGN) of CGenes obtained
 826 from Cuffdiff-DESeq-edgeR overlap (Anders and Huber, 2010; Robinson, McCarthy and Smyth, 2010;
 827 Trapnell *et al.*, 2012). (a) PPI network obtained from GeneMania (Montejo *et al.*, 2010) showing
 828 interconnection and regulation of genes displayed by nodes which are colored in blue and edges colored
 829 in grey, (b) FGN obtained from ClueGO (Bindea *et al.*, 2009) with GO Terms as nodes linked based on
 830 kappa score where node size represents enrichment significance.

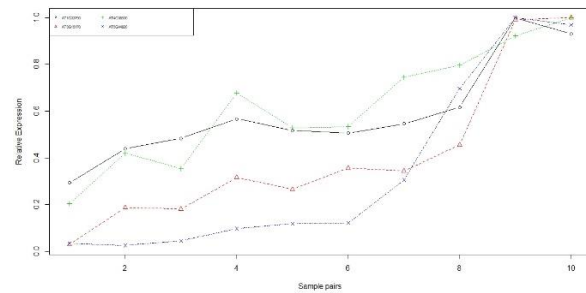
831

a

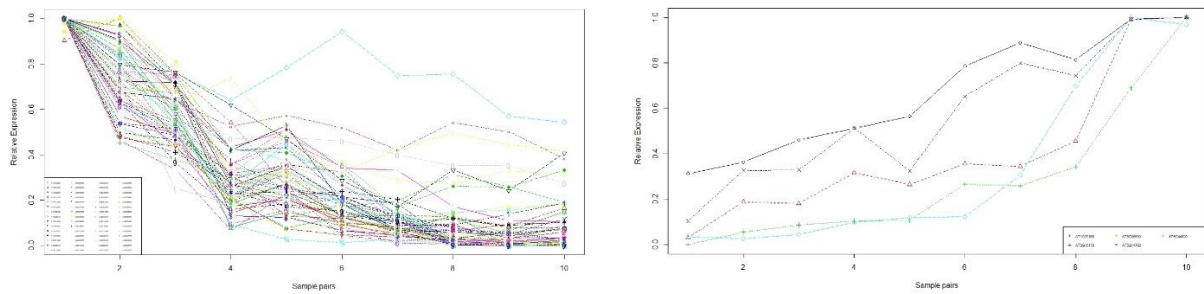


c

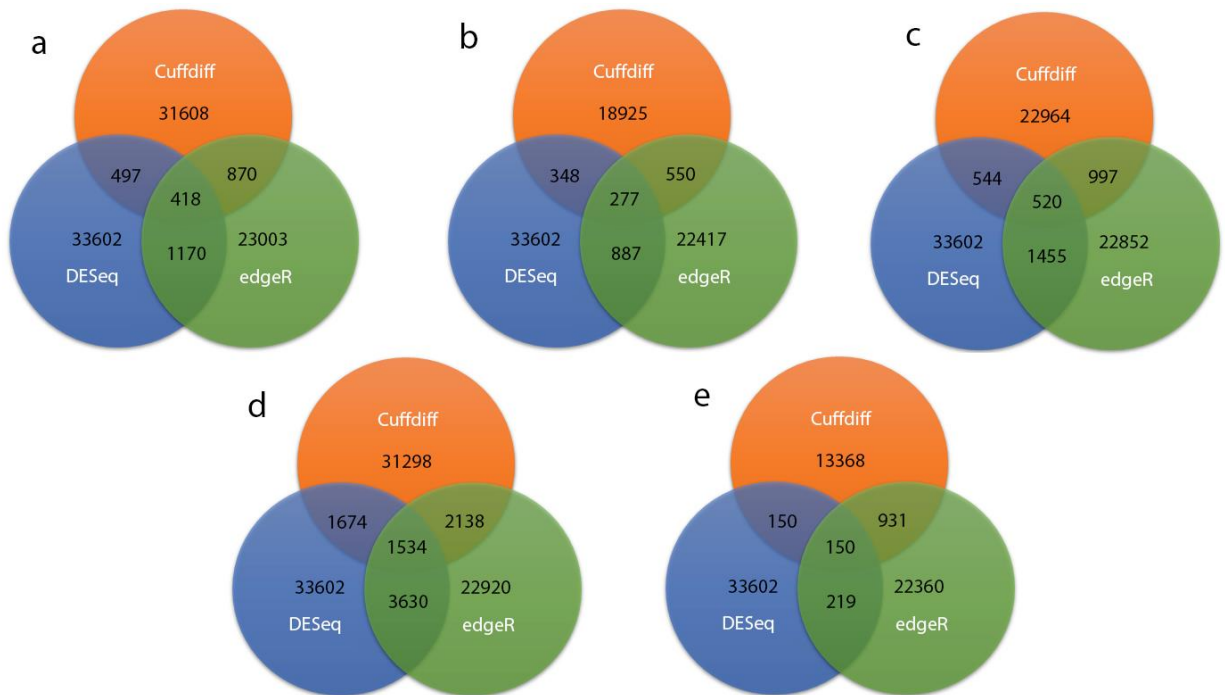
b

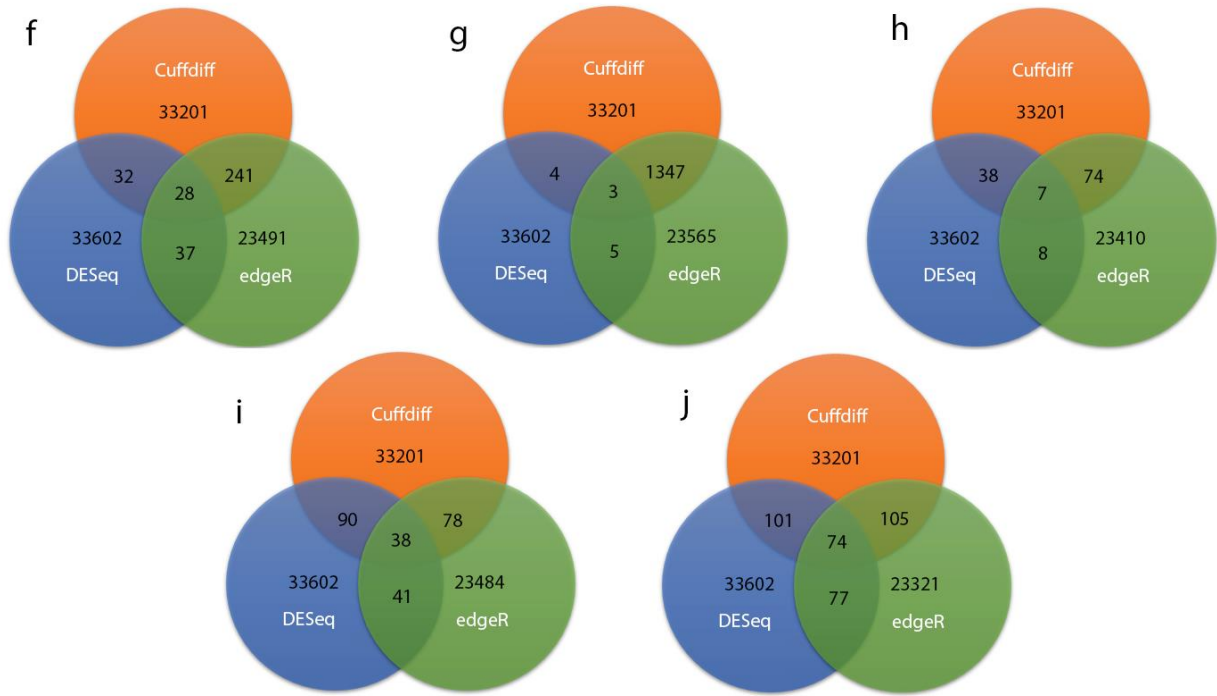


d



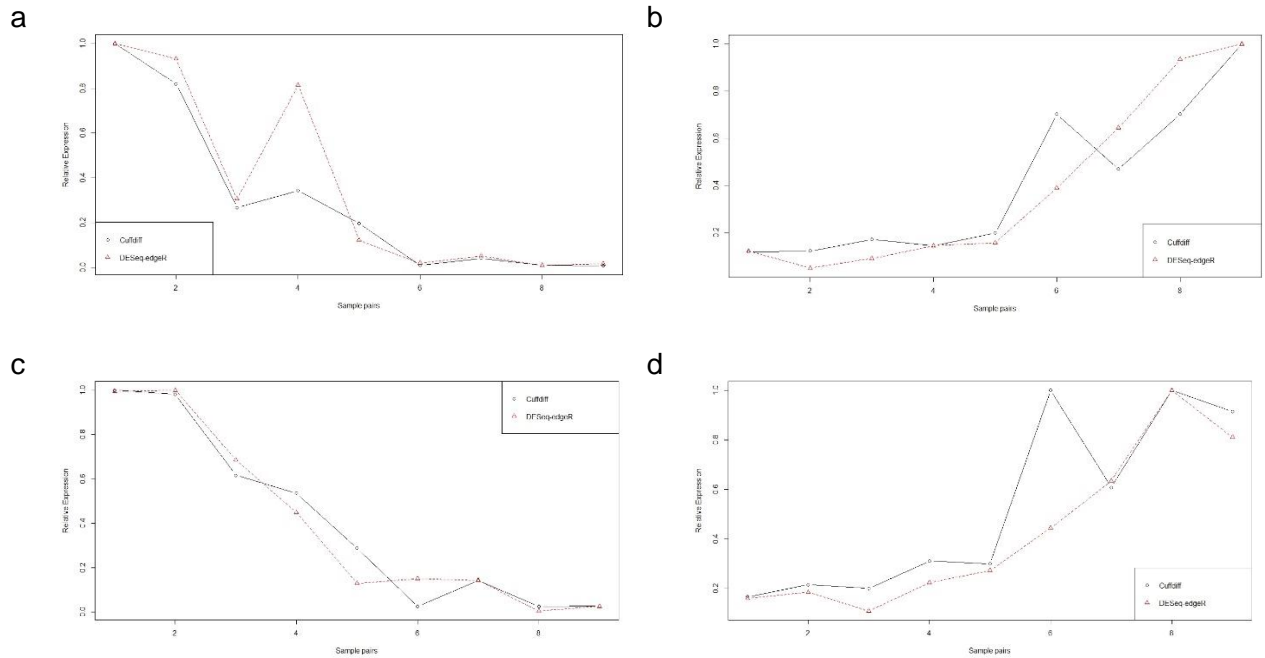
832 Fig. 6: Expression profiles of common genes from Cuffdiff-DESeq-edgeR (Anders and Huber, 2010;
 833 Robinson, McCarthy and Smyth, 2010; Trapnell *et al.*, 2012) overlap showing correlation to *FLC* and *LFY*
 834 genes. (a) shows DEGs showing higher correlation to *FLC* and regulation by *FLC*, (b) shows DEGs
 835 showing higher correlation to *LFY* and regulation by *LFY*, (c) shows DEGs having higher correlations to
 836 *FLC* and may or may not be regulated by *FLC*, (d) shows DEGs having higher correlations to *LFY*
 837 and may or may not be regulated by *LFY*
 838
 839

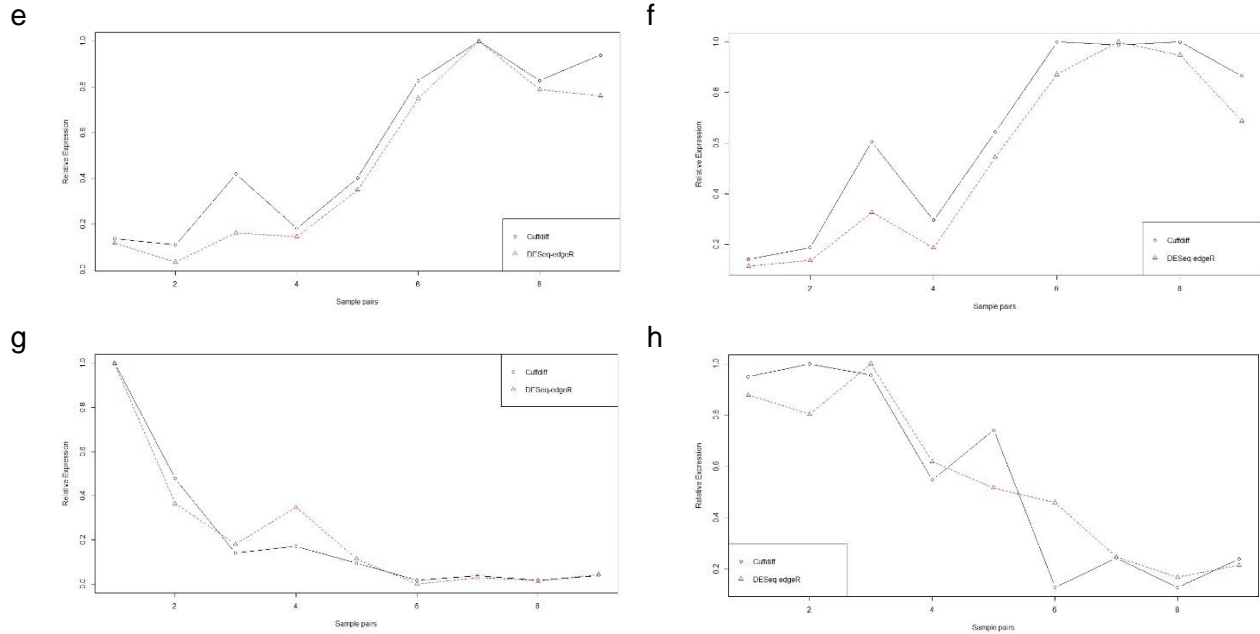




841
842
843
844
845
846
847
848
849

Fig. 7: Venn diagrams summarizing the overlap between Cuffdiff (Trapnell *et al.*, 2012), DESeq (Anders and Huber, 2010) and edgeR (Robinson, McCarthy and Smyth, 2010). (a), (b), (c), (d) and (e) shows overlapping genes from "Against S7" and (f), (g), (h), (i) and (j) shows overlapping genes from Step analysis. The overlapping genes are the DEGs in (a) S7-S10, (b) S7-S11, (c) S7-S12, (d) S7-S13, (e) S7-S14, (f) S9-S10, (g) S10-S11, (h) S11-S12, (i) S12-S13 and (j) S13-S14





850
 851 Fig. 8: Expression profiles of specific experimental genes in S7-S8, S7-S9, S7-S10, S7-S11, S7-S12, S7-
 852 S13, S7-S14, S7-S15 and S7-S16 using Cuffdiff (Trapnell *et al.*, 2012) and DESeq (Anders and Huber,
 853 2010) and edgeR (Robinson, McCarthy and Smyth, 2010) for (a) FLC, (b) LFY, (c) SMZ, (d) PPF1, (e)
 854 SPL9, (f) SPL15, (g) SNZ and (h) TOE2 genes
 855

856 Table 1: List of some parameters used for reference alignment of reads using Tophat2. Each parameter
 857 contains their description, default value and the changed value for the analysis.

Flag	Meaning	Default Value	Changed Value
-i	The minimum intron length.	70 nt	40 nt
-l	The maximum intron length.	500000 nt	5000 nt
--segment-length	Each read is cut up into segments, each at least this long. These segments are mapped independently.	25 segments	20 segments
-g	Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number.	20 alignments	1 alignment
-a	The "anchor length".	8 bp	10 bp
-F	Minimum normalized depth	300 bp	0 bp

858
859
860
861

Table 2: Comparison chart for differential expression analysis. Two analyses were carried out: first, all samples were compared to day 7 (S7) when plants had two leaves visible; second, a step-wise analysis was done between two successive days.

Against S7	Step analysis
S7 vs S10	S9 vs S10
S7 vs S11	S10 vs S11
S7 vs S12	S11 vs S12
S7 vs S13	S12 vs S13
S7 vs S14	S13 vs S14

862
863

Table 3: Number of overlapping DEGs found in Cuffdiff, DESeq and edgeR results with FDR <= 0.05

Against S7					Step analysis				
Sample pairs	Cuffdiff - DESeq-edgeR	Cuffdiff -DESeq	Cuffdiff -edgeR	edgeR-DESeq	Sample pairs	Cuffdiff-DESeq-edgeR	Cuffdiff -DESeq	Cuffdiff -edgeR	edgeR-DESeq
S7-S10	418	497	870	1170	S9-S10	28	32	241	37
S7-S11	277	348	550	887	S10-S11	3	4	1347	5
S7-S12	520	544	997	1455	S11-S12	7	38	74	8
S7-S13	1534	1674	2138	3630	S12-S13	38	90	78	41
S7-S14	150	150	931	219	S13-S14	74	101	105	77

864
865
866
867
868

Table 4: List of filtered GO terms along with associated genes found in KEGG (Ogata *et al.*, 1999) and REACTOME (Croft *et al.*, 2014) databases obtained from ClueGO analysis (Bindea *et al.*, 2009) having Term PValue corrected with Benjamini-Hochberg < 0.05 for common genes obtained from "Against S7" sample pairs.

GOID	GO Term	Ontology Source	Term PValue Corrected with Benjamini-Hochberg	Group PValue Corrected with Benjamini-Hochberg	% Associated Genes	Nr. Genes	Associated Genes Found
GO:0000261	Monobactam biosynthesis	KEGG	0.03	0.03	21.43	3.00	AK3, APS1, AT2G44040
GO:0000270	Cysteine and methionine metabolism	KEGG	0.01	0.00	10.10	10.00	ACO1, ACO2, AK3, ASP5, AT3G05430, CYSD1, HMT3, MS2, SERAT2;1, TAT3
GO:0000920	Sulfur metabolism	KEGG	0.00	0.00	19.51	8.00	AKN2, APK, APR1, APR2, APS1, AT4G05090, CYSD1, SERAT2;1
GO:0000290	Valine, leucine and	KEGG	0.02	0.02	17.39	4.00	BCAT4, IMD1, IPMI1, IPMI2

	isoleucine biosynthesis						
GO:0000660	C5-Branched dibasic acid metabolism	KEGG	0.02	0.02	30.00	3.00	IMD1, IPMI1, IPMI2
GO:0000380	Tryptophan metabolism	KEGG	0.02	0.00	13.04	6.00	CYP79B2, CYP79B3, CYP83B1, SUR1, TGG1, TGG2
GO:0000966	Glucosinolate biosynthesis	KEGG	0.00	0.00	57.89	11.00	BCAT4, CYP79B2, CYP79B3, CYP79F1, CYP79F2, CYP83A1, CYP83B1, MAM1, SOT17, SOT18, SUR1
GO:7438889	Cytosolic sulfonation of small molecules	REACTOME	0.02	0.16	16.00	4.00	AKN2, APK, SOT17, SOT18

869
870 Table 5: List of DEGs from Cuffdiff-DESeq-edgeR overlap (Anders and Huber, 2010; Robinson, McCarthy
871 and Smyth, 2010; Trapnell *et al.*, 2012) showing higher correlations to *FLC* and *LFY* and status of regulation

Regulated by FLC and having higher correlation to FLC	Not regulated by FLC but having higher correlation to FLC	Regulated by LFY and having higher correlation to LFY	Not regulated by LFY but having higher correlation to LFY
AT1G03820, AT1G60710, AT1G67870, AT1G76800, AT2G22330, AT2G43150, AT2G43510, AT3G04940, AT3G09260, AT4G19380, AT4G22510, AT4G28660, AT4G39950, AT5G04080	AT1G04240, AT1G06090, AT1G14700, AT1G20850, AT1G28400, AT1G28710, AT1G51680, AT1G65500, AT1G67910, AT1G70940, AT2G01950, AT2G02010, AT2G14580, AT2G22800, AT2G23600, AT2G30080, AT2G37040, AT2G37170, AT2G37180, AT2G37460, AT2G37640, AT2G38080, AT2G38800, AT3G02910, AT3G05727, AT3G10120, AT3G21550, AT3G22740, AT3G25190, AT3G49780, AT3G53560, AT3G61210, AT3G62930, AT4G01390, AT4G04610, AT4G14465, AT4G22485, AT4G22513, AT4G22517, AT4G22520, AT4G22530, AT4G24060, AT4G31990, AT4G32880, AT4G34560, AT4G36570, AT5G43580, AT5G50200, AT5G51890, AT5G52050, AT5G59330, AT5G60780, AT5G63180, AT5G63710, AT5G64110	AT3G15170, AT5G44620, AT1G33790, AT4G36930	AT1G70160, AT5G06530, AT5G14700

