Transcriptome analysis of colored calla lily (Zantedeschia rehmannii Engl.) by Illumina sequencing: de novo assembly, annotation and EST-SSR marker development

Zunzheng Wei [1,2] , **Zhenzhen Sun** [3] , **Binbin Cui** [4] , **Qixiang Zhang** [1] , **Min Xiong** [2] , **Xian Wang** [2] , **Di Zhou** [Corresp. 2]

[1] Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular Breeding, National Engineering Research Center for Floriculture and College of Landscape Architecture, Beijing Forestry University, Beijing, China

[2] Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North China), Ministry of Agriculture, Key Laboratory of Urban Agriculture (North), Ministry of Agriculture, Beijing Vegetable Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China

[3] Beijing Key Laboratory of Separation and Analysis in Biomedicine and Pharmaceuticals, Beijing Institute of Technology, Beijing, China

[4] Department of Biology and Chemistry, Baoding University, Baoding, Hebei, China

Corresponding Author: Di Zhou
Email address: zhoudibvrc@163.com

Colored calla lily is the short name for the species or hybrids in section Aestivae of genus Zantedeschia. It is currently one of the most popular flower plants in the world due to its beautiful flower spathe and long postharvest life. However, little genomic information and few molecular markers are available for its genetic improvement. Here, de novo transcriptome sequencing was performed to produce large transcript sequences for Z. rehmannii cv. 'Rehmannii' using an Illumina HiSeq 2000 instrument. More than 59.9 million cDNA sequence reads were obtained and assembled into 39,298 unigenes with an average length of 1,038 bp. Among these, 21,077 unigenes showed significant similarity to protein sequences in the non-redundant protein database (Nr) and in the Swiss-Prot, Gene Ontology (GO), Cluster of Orthologous Group (COG) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. Moreover, a total of 117 unique transcripts were then defined that might regulate the flower spathe development of colored calla lily. Additionally, 9,933 simple sequence repeats (SSRs) and 7,162 single nucleotide polymorphisms (SNPs) were identified as putative molecular markers. High-quality primers for 200 SSR loci were designed and selected, of which 58 amplified reproducible amplicons were polymorphic among 21 accessions of colored calla lily. The sequence information and molecular markers in the present study will provide valuable resources for genetic diversity analysis, germplasm characterization and marker-assisted selection in the genus Zantedeschia.

1 **Transcriptome analysis of colored calla lily (*Zantedeschia rehmannii* Engl.) by**

2 **Illumina sequencing: *de novo* assembly, annotation and EST-SSR marker**

3 **development**

4 Zun-zheng Wei[1,2,†], Zhen-zhen Sun[3,†], Bin-bin Cui[4], Qi-xiang Zhang[1], Min Xiong[2], Xian

5 Wang[2], Di Zhou[2*]

6 [1] Beijing Key Laboratory of Ornamental Plants Germplasm Innovation & Molecular

7 Breeding, National Engineering Research Center for Floriculture and College of

8 Landscape Architecture, Beijing Forestry University, Beijing100083, China;

9 [2] Key Laboratory of Biology and Genetic Improvement of Horticultural Crops (North

10 China), Ministry of Agriculture, Key Laboratory of Urban Agriculture (North), Ministry of

11 Agriculture, Beijing Vegetable Research Center, Beijing Academy of Agriculture and

12 Forestry Sciences, Beijing 100091, China;

13 [3] Beijing Key Laboratory of Separation and Analysis in Biomedicine and Pharmaceuticals,

14 Beijing Institute of Technology, Beijing 10081, China;

15 [4] Department of Biology and Chemistry, Baoding University, Baoding, Hebei 071000,

16 China;

17 [†] These authors contributed equally to this work.

19 Corresponding Author:

20 Di Zhou

21 Zhanghua Road No. 50, Haidian District, Beijing 100097, China

22 Email address: zhoudibvrc@163.com

24 **ABSTRACT**

25 Colored calla lily is the short name for the species or hybrids in section *Aestivae* of genus

26 *Zantedeschia*. It is currently one of the most popular flower plants in the world due to its

27 beautiful flower spathe and long postharvest life. However, little genomic information and

28 few molecular markers are available for its genetic improvement. Here, *de novo*

29 transcriptome sequencing was performed to produce large transcript sequences for *Z.*

30 *rehmannii* cv. 'Rehmannii' using an Illumina HiSeq 2000 instrument. More than 59.9

31 million cDNA sequence reads were obtained and assembled into 39,298 unigenes with an

32 average length of 1,038 bp. Among these, 21,077 unigenes showed significant similarity to

33 protein sequences in the non-redundant protein database (Nr) and in the Swiss-Prot, Gene

34 Ontology (GO), Cluster of Orthologous Group (COG) and Kyoto Encyclopedia of Genes

35 and Genomes (KEGG) databases. Moreover, a total of 117 unique transcripts were then

36 defined that might regulate the flower spathe development of colored calla lily.

37 Additionally, 9,933 simple sequence repeats (SSRs) and 7,162 single nucleotide

38 polymorphisms (SNPs) were identified as putative molecular markers. High-quality

39 primers for 200 SSR loci were designed and selected, of which 58 amplified reproducible

40 amplicons were polymorphic among 21 accessions of colored calla lily. The sequence

41 information and molecular markers in the present study will provide valuable resources for

42 genetic diversity analysis, germplasm characterization and marker-assisted selection in the

43 genus *Zantedeschia*.

44

45 **INTRODUCTION**

46 *Zantedeschia* species, commonly known as calla lily or arum lily, belong to the genus

47 *Zantedeschia* in the family Araceae. These plants are native to central and southern Africa

48 and generally grow in marshy places, on grassy slopes, and even at forest margins. The

49 genus *Zantedeschia* is divided into two distinct sections (*Letty*, *1973*; *Singh*, *1996*): section

50 *Zantedeschia*, with one evergreen species (*Z. aethiopica* Spreng.), and section *Aestivae*,

51 with six deciduous species (*Z. rehmannii* Engl., *Z. jucunda* Letty., *Z. elliottiana* Engl., *Z.*

52 *pentlandii* Wittm., *Z. valida* Singh. and *Z. albomaculata* Baill.). *Z. aethiopica*, also called

53 white calla lily, is characterized by a pure white spathe and leaves that do not die down in

54 the winter. In contrast, the species or hybrids of section *Aestivae*, also called colored calla

55 lily, are characterized by a variety of spathe colors and have leaves that die down in the

56 winter. In addition, a new species, *Z. odorata* Perry., that is dormant in the summer and has

57 a white spathe that can produce an invariably delicate freesia-like scent, was classified into

58 section *Zantedeschia*. There are a number of post-fertilization incompatibility barriers

59 between the species in section *Zantedeschia* and section *Aestivae*, including endosperm

60 degeneration, abnormal embryo development and arrested plastid development (*Yao et al.*,

61 *1994a*; *Yao et al.*, *1995*).

62   As a famous flower plant, colored calla lily is very popular all over the world because of

63 its beautiful flower spathe and long postharvest life. It is an important export flower crop

64 for New Zealand, the Netherlands and the United States (*Tjia*, *1985*; *Funnell*, *1993*;

65 *Funnell & MacKay*, *1999*). The extensive commercial production of colored calla lily for

66 cut flowers and/or planting material occurs in Auckland and Palmerston North (New

67 Zealand), South Holland (the Netherlands), California and Colombia (the United States)

68 and so on. The potential application of colored calla lily for commercial tuber production,

69 cut flowers, pot flowers and even as landscape specimens is also unlimited worldwide

70 (*Tjia*, *1985*). Significant advances have been made, especially for colored calla lily,

71 through traditional breeding practices in the floriculture industry. A multitude of cultivars

72 and hybrids showing great variation in flower color from dark red, pink, orange, yellow to

73 even white are available at present. In fact, most of the cultivars that are hybridized,

74 commercially developed, and used for cut or potted ornamental purposes belong to the

75 species *Z. rehmannii*, *Z. elliottiana*, *Z. pentlandii* and *Z. albomaculata* (*Funnell*, *1993*;

76 *Funnell & MacKay*, *1999*).

77   Molecular markers are powerful genetic tools for gene mapping and molecular marker-

78 assisted selection (MAS) in the breeding of several crop species (*Varshney et al.*, *2007*).

79 Previous studies on colored calla lily also have reported that several dominant markers,

80 including random amplified polymorphic DNA (RAPD) and inter-simple sequence repeats

81 (ISSR), were used for cultivar identification and to assess genetic diversity (*Hamada &*

82 *Hagimori*, *1996*; *Lu et al.*, *2012*; *Zhang et al.*, *2009*; *Chen et al.*, *2013*; *Lu et al.*, *2014*).

83 However, the current conventional breeding of calla lily in section *Aestivae* is still

84 performed without the aid of molecular markers. It is perhaps the dominant inheritance

85 pattern of these markers that hinders the detection of allelic information. In contrast, co-

86   dominant markers, such as simple sequence repeats (SSRs) and single nucleotide

87   polymorphisms (SNPs), are the favored markers for the detection of allelic diversity.

88   Among these markers, SSR markers have gained attention due to their multi-allelic nature,

89   abundance throughout the genome, high reproducibility and polymorphism, adaptability to

90   automation and high-throughput genotyping (*Morgante*, *Hanafey & Powell*, *2002*).

91   Nevertheless, progress in developing SSR markers in section *Aestivae* and the genus

92   *Zantedeschia* is limited compared to that in other crop plants. Only one previous effort has

93   produced 43 polymorphic SSRs from 4,394 EST sequences of *Z. aethiopica*, representing

94   the first dataset of polymorphic EST-SSR markers for the genus *Zantedeschia* (*Wei et al.*,

95   *2012*). This conventional method of EST-SSR development has a high development cost

96   and a low throughput, which may restrict the use of SSR markers in further genetic

97   breeding (*Morgante*, *Hanafey & Powell*, *2002*; *Varshney*, *Graner & Sorrells*, *2005*).

98   Recent advances in RNA sequencing (RNA-seq) technology and *de novo* assembly

99   provide an excellent strategy for the efficient and cost-effective development of high-

100   throughput EST-SSRs. This strategy can enrich the large amounts of expressed sequence

101   data for non-model organisms for which the reference genome or transcriptome is not

102   available (*Martin & Wang*, *2011*). Several non-model organisms in the Araceae family,

103   including *Z. aethiopica* (*Cândido et al.*, *2014*), *Arum concinnatum* (*Onda et al.*, *2015*),

104   *Anthurium andraeanum* (*Tian et al.*, *2013*; *Yang et al.*, *2015*; *Li et al.*, *2015*), *Pinellia*

105   *ternate* (*Wang et al.*, *2014*), *Amorphophallus konjac* and *A. bulbifer* (*Zheng et al.*, *2013*;

106   *Diao et al.*, 2014), have been recently studied by transcriptome sequencing, which has

107   provided a better understanding of these crops. EST-SSR markers have been successfully

108   characterized and developed in *P. ternate* (Wang et al., 2014), *A. konjac* and *A. bulbifer*

109   (*Zheng et al.*, *2013*). A total of 14,468 and 19,596 EST-SSRs were identified in 12,000 and

110   16,027 non-redundant transcriptome unigenes, respectively, of *P. ternate* and two

111   *Amorphophallus* species (*Wang et al.*, *2014*; *Zheng et al.*, *2013*). Furthermore, 320 primers

112   were synthesized and used to validate the amplification and assessment of polymorphisms

113   in 25 individual *Amorphophallus* spp. plants (*Zheng et al.*, *2013*), of which 275 primer

114   pairs yielded PCR amplification products and 205 were polymorphic. This strongly

115   demonstrated that de novo assembly based on RNA-seq can offer a simple, direct and

116   reliable approach for the identification and development of massive unigene-based

117   microsatellite markers with diverse motifs. In addition to the development of molecular

118   markers, transcriptome sequencing has also been used for the discovery, profiling and

119   quantification of RNA transcripts and novel genes. For example, three lipid transfer

120   proteins (LTP) that are potentially involved in defense against pathogens or predators were

121   identified by the *in silico* screening of the 83,578 transcriptome contigs of *Z. aethiopica*

122   (*Cândido et al.*, *2014*). This represents the first transcriptome dataset for the genus

123   *Zantedeschia*.

124   Here, we presented the generation of a large expressed sequence dataset based on

125   Illumina HiSeq™ 2000 sequencing data from mixed tissues of colored calla lily, *Z.*

126   *rehmannii* cv. 'Rehmannii'. The objectives were as follows: (1) to characterize and

127   annotate the transcriptome information of colored calla lily; (2) to identify all of the

128   candidate genes encoding enzymes or putative transcription factors that are involved in its
129   flower spathe development; (3) to analyze the frequency and distribution of SSRs and
130   SNPs in its transcribed regions; (4) to develop a large set of EST-SSR markers; and (5) to
131   evaluate the transferability and genetic diversity of 21 cultivars or hybrids of colored calla
132   lily using these EST-SSR markers.
133
134   **MATERIALS AND METHODS**
135   **Plant materials**
136   All of the samples used in the present study were maintained at the Bulb and Perennial
137   Flowers Genebank (BPFG), Beijing Academy of Agriculture and Forestry Sciences,
138   Beijing, China. The samples were grown in the experimental greenhouse in Yanqing Farm
139   (latitude 40.46N, longitude 115.91E); this cultivation did not involve endangered or
140   protected species. *Z. rehmannii* cv. Rehmannii, a potted cultivar with a pink spathe, was
141   selected for transcriptome sequencing. An additional twenty accessions (Table S1),
142   including nine individuals from New Zealand (Elmaro, Pink Diamond, Butter Gold, Black
143   Magic, Majestic Red, Sunny Baby, Greta, Goldilocks and Gold affair), five from the
144   Netherlands (Odessa, Captain Reno, Allure, Captain Romance, and Captain Murano), three
145   from the United States (Parfait, Super Gem and Rose Gem) and three from China (Jingcai
146   Yangguang, ZH and Xiangyuan Red), were used to estimate microsatellite variations of the
147   EST-SSR markers developed in our study. All of the accessions are released commercial
148   varieties of New Zealand, the Netherlands, the United States or China; thus, permission for
149   collection is not required.
150   **RNA extraction, cDNA library construction and Illumina sequencing**
151   To achieve a comprehensive collection of expressed transcripts of colored calla lily, ten
152   'Rehmannii' tissue types, including root, tuber, stem, leaf, bud, spathe, pistillate
153   inflorescence and staminate inflorescence, immature seed and mature seed, were separately
154   harvested at the anthesis and wilting stages. All of the tissue samples were immediately
155   frozen in liquid nitrogen and stored at -80 ℃ until RNA extraction. Total RNA Trizol
156   Reagent (Invitrogen, USA) was used to extract RNA from all of the samples, following the
157   manufacturer's instructions. The RNA purity and concentration were determined using a
158   NanoDrop spectrophotometer (Thermo, USA). Equal quantities of total RNA from each
159   sample were then pooled together and used for transcriptome sequencing.
160   cDNA library construction and Illumina-based sequencing were performed by the
161   Shanghai Majorbio Bio-pharm Biotechnology Co. Ltd. (Shanghai, China) according to the
162   manufacturer's instructions. The paired-end cDNA sequencing libraries with approximate
163   average insert lengths of 200 base pairs were prepared from the total RNA, as per the
164   protocol of the Genomic Sample Prep kit (Illumina, San Diego, CA). This process started
165   with mRNA fragmentation, followed by reverse transcription, first- and second-strand
166   synthesis, paired-end adapter ligation and PCR amplification. Library quantification and
167   quality assessment were performed on an Agilent 2100 Bioanalyzer and an ABI
168   StepOnePlus Real-Time PCR System. Finally, the cDNA library was sequenced as 101-
169   mer×2 on an Illumina HiSeq 2000 using paired-end sequencing chemistry.

De novo transcriptome assembly and functional annotation

The raw sequencing image data were transformed into raw reads and stored in FASTQ format. These raw data were then filtered and deposited in the National Center for Biotechnology (NCBI) Sequence Read Archive (SRA) under accession number SRR3310941. All of the reads with adaptor contamination, empty reads, non-coding RNA (such as rRNA, tRNA and miRNA), ambiguous nucleotides comprising more than 5% or low quality value (QV), i.e., an average QV of less than 20 (QV < 20), were discarded or filtered. Then, the clean and high-quality transcriptome sequence data were *de novo* assembled using the short reads assembling program Trinity (*Haas et al.*, *2013*) with default settings. An assembled transcripts database (File S2) was finally achieved and arranged according to the gene family clustering analysis. These final assembled transcripts were used for further bioinformatics analysis.

The assembled sequences were functionally annotated by BLASTx against a series of databases, including the Nr, Swiss-Prot, COG, and KEGG databases, with a common significance threshold cut-off of E-value 1e-10^{-5}. Based on Nr annotation, Blast2GO (https://www.blast2go.com/) and WEGO (http://wego.genomics.org.cn/cgi-bin/wegol) software were used to retrieve and classify GO annotation categories defined by molecular function, cellular component and biological process.

Marker locus detection and SSR primer pair design

MIcroSAtellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/misa.html) was used to identify microsatellites in the assembled transcripts. The minimum number of repeats used to select the SSRs was ten for mononucleotide repeats, six for dinucleotide repeats, and five for tri-, tetra-, penta-, and hexanucleotide repeats. Primer Premier 5.0 (PREMIER Biosoft International, Palo Alto, CA) was then used to manually design 200 pairs of PCR primers to randomly select sequences with SSR loci. The design criteria of the primers were as follows: primer length 18-24 bp; GC content 40-65%; melting temperature 50–65℃; and expected product size 100-300 bp with no secondary structures. All 200 primer pairs were synthesized by Sangon Biological Engineering Technology (Shanghai, China).

Potential SNPs were detected using the programs BWA (http://biobwa.sourceforge.net/) and VarScan (http://varscan.sourceforge.net/). The assembled unigenes were used as references to BLAST the raw sequencing reads.

DNA extraction and EST-SSR marker amplification

The young leaves were collected from 21 accessions (Table S1) grown at BPFG. The total genomic DNA was extracted using the DNeasy Plant Mini Kit (Zexing Biotech, China) following the manufacturer's protocol. The quantity and quality of DNA were evaluated using a Nanodrop ND 1000 spectrophotometer (Thermo Scientific, USA). The DNA was adjusted to a concentration of 20 ng/μl and stored at -20℃ until use.

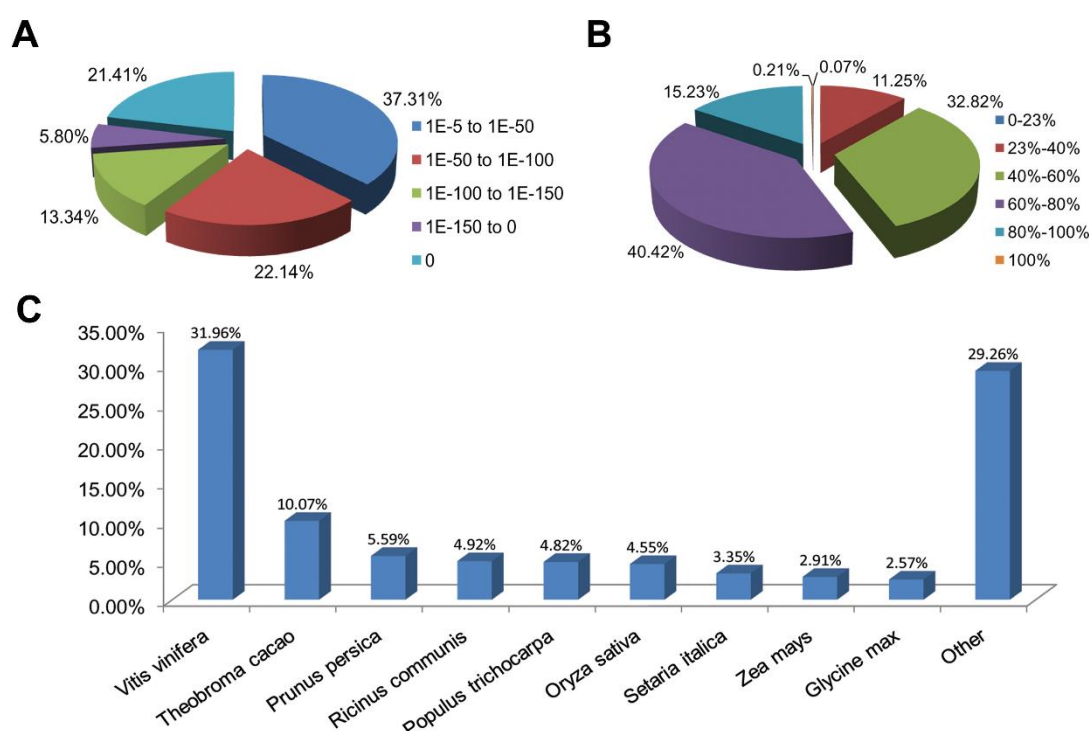
The SSR amplification reactions were conducted as described previously by Wei et al. (*2012*) with little modification. The PCR reaction mixtures (10 μl) contained 1× PCR buffer, 20-30 ng of template DNA, 0.8 mM MgCl$_2$, 15 μM dNTPs, 0.25 μM each primer, and 0.2 U of Taq DNA polymerase (Zexing Biotech, China). PCR amplifications were

212 performed in the GeneAmp PCR System 9700 (Applied Biosystems). The thermal profile
213 included an initial denaturation at 95 °C for 5 min, followed by 20 cycles of 95 °C for 30 s,
214 50-60 °C for 45 s, and 72 °C for 60 s and a final extension at 72 °C for 10 min. The PCR
215 products were separated on 8.0% polyacrylamide non-denaturing gels and visualized by
216 silver staining (*Wei et al.*, *2012*). The product sizes were determined by comparison with a
217 34- to 501-bp pUC19/MspI DNA marker (Zexing Biotech, China)
218 **Data analysis**
219 The number of alleles (*Na*), observed heterozygosity (*Ho*), expected heterozygosity (*He*),
220 and polymorphic information content (*PIC*) were calculated using GenAlEx 6.4 (*Peakall &*
221 *Smous*, *2006*) and Power Marker Version 3.25 (*Liu & Muse*, *2005*). A cluster analysis was
222 conducted and displayed using the neighbor-joining (NJ) algorithm as implemented in
223 Power Marker Version 3.25 (*Liu & Muse*, *2005*) and MEGA version 5.0 (*Tamura et al.*,
224 *2011*).
225
226 **RESUTLTS**
227 **Illumina sequencing and *de novo* assembly**
228
229 Table 1 Summary of transcriptome statistics and functional annotation for colored calla lily
230 'Rehmannii'

	Number	Percentage
Raw reads	59,882,890	
Total sizes (nt)	6,048,171,890	
Clean reads	46,343,613	
Transcripts	62,382	
Unigenes	39,298	
Unigenes (300–500 nt)	13,367	34.02%
Unigenes (500–1,000 nt)	13,088	33.30%
Unigenes (1,000–1,500 nt)	4,975	12.65%
Unigenes (1,500–2,000 nt)	2,988	7.60%
Unigenes (> 2,000 nt)	4,880	12.41%
Mean length (nt)	1,038	
N50 (nt)	1,476	
GC%	45.74%	
Annotated in Nr	21,029	53.51%
Annotated in Swiss-Prot	16,908	43.03%
Annotated in COG	6,731	17.13%
Annotated in GO	15,552	39.57%
Annotated in KEGG	4,532	11.53%
Annotated in at least one database	21,077	53.63%
Total unigenes	39,298	100%

231

232   To provide a comprehensive transcriptome database for calla lily, a mixed cDNA library of
233   cultivar 'Rehmannii' from ten sampled tissues, i.e., root, tuber, stem, leaf, bud, spathe,
234   pistillate inflorescence and staminate inflorescence, immature seed and mature seed, was
235   constructed and sequenced using Illumina paired-end technology. A summary of the
236   sequencing output statistics is shown in Table 1. The sequencing yielded approximately
237   59.9 million raw reads with a total of more than 6.1 billion nucleotides. After removing
238   ambiguous, low-quality reads and reads with adaptors, 46.3 clean reads were obtained for
239   further assembly. The *de novo* assembly yielded 62,382 transcripts and 39,298 unigenes.
240   The sequences of all the assembled unigenes were provided in FASTA format in Table S2.
241   These sequences contained 40.8 Mb of sequence with an average size of 1,038 bp and an
242   N50 of 1,476 bp. The lengths of the unigenes ranged from 351 to 15,521 bp. Of these
243   unigenes, 67.3% (26,455) were shorter than 1,000 bp, 20.3% (7,963) ranged from 1,000 to
244   2,000 bp, and the remaining 12.4% (4, 880) were longer than 2,000 bp (Table 1). These
245   unigenes formed a potential pool for identification of genes and functional molecular
246   markers in colored calla lily.
247   **Functional annotation of the unigenes**
248

249

250   Figure 1 Characteristics of homology search for colored calla lily 'Rehmannii' unigenes
251   against non-redundant protein database (Nr) with an E-Value ≤ 1E-10^{-5}. (A) The E-value
252   distribution of BLASTx hit for each assembled unigene; (B) The similarity distribution of
253   BLASTx hits for each assembled unigenes; (C) Species-based distribution of the top
254   BLASTx hits for each assembled unigenes.
255

256   Functional annotations were performed by a homology-based approach for cultivars
257   'Rehmannii' assembled transcripts. A sequence similarity search was first conducted
258   against the Nr and Swiss-Prot databases using the BLASTx algorithm with an E-value
259   threshold of 10^{-5} (Table 1 and Table S3). As shown in Table 1, the results indicated that
260   21,029 (53.5%) of the 39,298 unigenes showed significant BLASTx matches in the Nr
261   database, while 16,908 (43.0%) were similar to proteins in the Swiss-Prot database. The
262   proportion of sequences showing hits in both the Nr and Swiss-Prot databases was higher
263   among the longer assembled transcripts. More than 55.0% of the unigenes longer than
264   1,000 bp showed homologous matches while fewer than 45.0% of the unigenes shorter
265   than 1,000 bp showed matches (File S2). The E-value, sequence similarity and species
266   distributions of the top hits in the Nr database were also analyzed. The E-value distribution
267   of the top hits showed that 62.7% of the annotated sequences had high homology with E-
268   value $< 10^{-50}$, whereas 37.3% showed a moderate homology with E-values from 10^{-5} to 10^{-50}
269   ^{50} (Fig. 1A). For the sequence similarity distribution analysis, 2,365 (11.3%), 6,902
270   (32.8%), 8,500 (40.4%), 3,202 (15.2%) and 45 (0.21%) sequences were 23-40%, 40-60%,
271   60-80%, 80-100% and 100% similar in the Nr database, respectively (Fig. 1B). In addition,
272   the species distribution showed that *Vitis vinifera* (Vitaceae) was ranked first, with 6,720
273   (32.0%) top BLASTx hits, followed by *Theobroma cacao* (Sterculiaceae), *Prunus persica*
274   (Rosaceae), *Ricinus communis* (Euphorbiaceae) and *Populus trichocarpa* (Salicaceae),
275   with 2,117 (10.1%), 1,176 (5.6%), 1,034 (4.9%), and 1,014 (4.8%) hits, respectively (Fig.
276   1C).

277   **Functional classification of the unigenes**
278

279

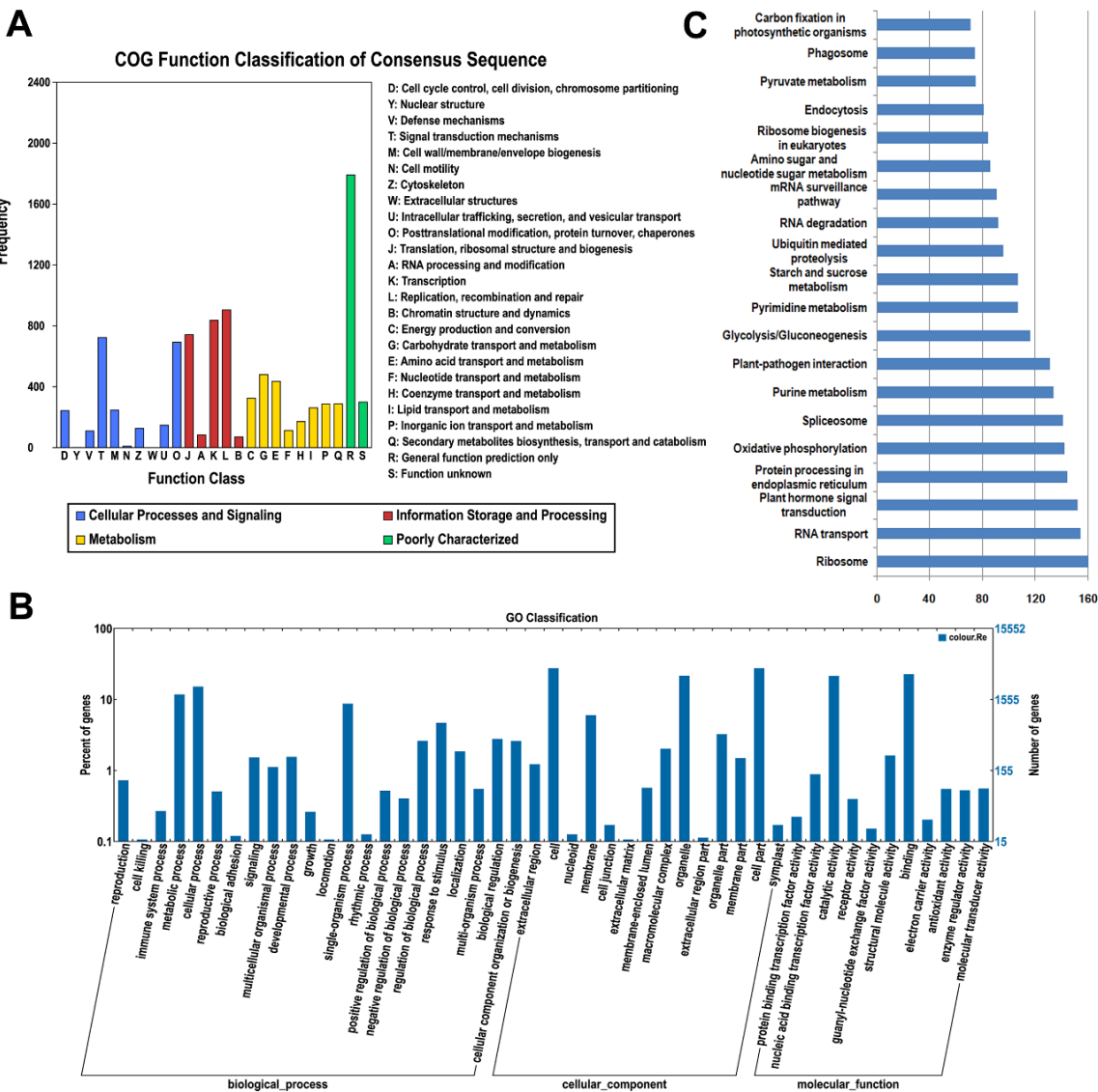



Figure 2 The classification of colored calla lily 'Rehmannii' unigenes. (A) Distribution of
Cluster of Orthologous Group (COG) classification. A total of 6,731 assembled unigenes
were annotated and assigned to 24 functional categories. The x-axis indicates the
subgroups in the COG classification while the y-axis indicates the number of genes in each
main category. (B) Gene ontology (GO) classification of assembled unigenes at level 2. A
total of 15,552 unigenes were grouped into three main GO categories: 'Biological
Processes', 'Cellular Component', and 'Molecular Function'. The x-axis indicates the
subgroups in GO annotation while the y-axis indicates the percentage of specific categories
of genes in each main category. (C) The top 20 KEGG metabolic pathways of assembled
unigenes. The x-axis indicates the number of genes in each metabolic pathway while the y-
axis indicates the subgroups in the top 20 KEGG metabolic pathways.

291

All of the assembled unigenes were subjected to a search against the COG, GO and KEGG
databases for further functional prediction and classification (Table S3). The COG
database can provide phylogenetic classification of proteins encoded by several complete

295 genomes of bacteria, archaea and eukaryotes (*Roman et al.*, *2000*). The COG function
296 classification of the 'Rehmannii' sequences is shown in Fig. 2A. In total, 6,731 of the
297 39,298 unigenes showing Nr hits were functionally annotated and classified into 24 COG
298 categories, including cellular structure, biochemistry metabolism, molecular processing,
299 and signal transduction, among others. Given that 9,379 COG-annotated putative proteins
300 were obtained, some of these unigenes were assigned to multiple COG classifications. The
301 cluster for general function prediction only (1,791, 26.6%) represented the largest group,
302 followed by replication, recombination and repair (904, 13.4%) and transcription (836,
303 12.4%). Additionally, only a few unigenes were assigned to cell motility (9, 0.13%) and
304 nuclear structure (1, 0.01%).
305   The GO database, an internationally standardized gene functional classification system,
306 offers dynamic and updated gene ontology that defines gene products in terms of their
307 associated cellular component, molecular function, and biological process in any organism
308 (*Ashburner et al.*, *2007*). Based on Nr annotation, 15,552 (39.6%), unigenes in the present
309 study were assigned to GO classes with 15,964 functional terms. A summary of the
310 'Rehmannii' unigenes classified to each GO Slim term is shown in Fig. 2B. The annotated
311 gene sequences that belong to the cellular component, molecular function, and biological
312 process categories were divided into 47 functional groups. Under the cellular component
313 category, cell (4,254, 27.4%) and cell part (4,253, 27.4%) were the most highly represented
314 group, followed by organelle (3,331, 21.4%). For the molecular function category, the top
315 two groups were binding (3,512, 22.6%) and catalytic activity (3,313, 21.3%). However,
316 the majority of the groups, including antioxidant activity, receptor activity, protein binding
317 transcription factor activity, nutrient reservoir activity, etc., contained only a few unigenes
318 (655, 4.2%). For biological processes, the majority of the unigenes were involved in
319 cellular processes (2,344, 15.1%) and metabolic processes (1,814, 11.7%), indicating that
320 important metabolic and cellular activities occur in 'Rehmannii'. Genes involved in other
321 important biological processes, such as single-organism processes (1,337, 8.6%), stimulus
322 response (712, 4.6%) and biological regulation (416, 2.7%), also accounted for a large
323 fraction of the annotated transcripts. Furthermore, we found that a portion of the unigenes
324 were assigned to developmental processes (226, 1.5%), reproductive processes (63, 0.4%)
325 and growth (25, 0.2%) and may be involved in flower-related biological processes, such as
326 flower-type development and formation, in colored calla lily.
327   The KEGG database (*Kanehisa et al.*, *2004*) determines the biological pathways in
328 which protein and small-molecule interactions occur. A total of 4,532 assembled sequences
329 (11.5%) were consequently annotated and assigned to 117 predicted metabolic pathways.
330 The number of sequences in each pathway ranged from 1 to 160. Fig. 2C shows the top 20
331 KEGG metabolic pathways represented by unique transcripts of calla lily. Ribosome
332 represented the largest number of transcript sequences in our dataset (160), followed by
333 RNA transport (154) and plant hormone signal transduction (152). Among the annotated
334 sequences, more than one-third (1,661) were classified into 46 metabolic groups, such as
335 purine metabolism (134), pyrimidine metabolism (107), starch and sucrose metabolism

336 (75), and so on. This indicates that diverse metabolic processes are active and a variety of

337 metabolites are synthesized in the tissues and organs of colored calla lily.

338 Overall, 21,077 unique sequence-based annotations using the selected Nr, Swiss-Prot,

339 COG, GO and KEGG databases were assigned to assembled transcripts of the cultivar

340 'Rehmannii'. The functional analysis revealed that it is feasible to obtain transcriptome

341 sequences through high-throughput technology, even for non-model plants with large

342 genomes. Furthermore, all of these functional annotation assignments provide valuable

343 information for colored calla lily to investigate specific biochemical and developmental

344 processes and potential gene structures, functions and pathways.

345 **Unigenes related to flower development in colored calla lily**

346 Enzymes functioning in flower development have been well documented in many plants.

347 Based on Nr annotation results of colored calla lily unigenes, we identified a total of 171

348 candidate transcripts encoding enzymes related to pigment biosynthesis (61), floral organ

349 development (26), flowering regulation (63) and flower senescence (31). The EST

350 sequences of all these unigenes are listed in Table S4. Flower development is a complex

351 process controlled by an integrated network of multi-genetic pathways in higher plants.

352 The putative gene homologs identified in the present study were involved in eight

353 pathways, including the anthocyanin biosynthesis pathway (52), carotenoid biosynthesis

354 pathway (9), specification of floral organ identity (26), photoperiod pathway (22),

355 vernalization pathway (11), gibberellic acid pathway (11), autonomous or other pathways

356 (9) and ethylene biosynthesis pathway (31). The identification and analysis of these key

357 genes will provide a foundation for understanding the potential molecular genetic

358 mechanisms controlling different aspects of floral development of colored calla lily in the

359 future.

360 **SSR and SNP loci discovery**

361

362 Table 2 Features of the SSR repeat types identified in colored calla lily 'Rehmannii'

363 unigenes.

Feature	Colored calla lily
Total number of sequences examined	39,298
Total size of examined sequences (Mb)	40.78
Total number of identified SSRs	9,933
Number of SSR-containing sequences	7,997
Number of sequences containing more than one SSR locus	1,556
Number of SSRs present in compound formation	580

364

365 Transcriptome sequencing is important for identification and development of molecular

366 markers, such as SSRs and SNPs. For the development of new markers for colored calla

367 lily, all of the 39,298 unigenes generated in this study were used to mine potential

368 microsatellites using MISA software. A total of 9,933 potential EST-SSRs were identified

369 in 7,997 unigenes, of which 1,556 sequences contained more than one SSR, and 580 SSRs

370 were present in compound form (Table 2 and Table S5). Considering that approximately

371 40,780 kb was analyzed, we detected a frequency of at least one SSR per 4.1 kb in the
372 expressed fraction of the 'Rehmannii' genome.
373
374 Table 3 Summary of EST-SSRs identified from the unigenes of colored calla lily
375 'Rehmannii'.

Repeat motif	Number of repeats							
	5	6	7	8	9	10	> 10	Total
Di- (3,482, 59.78%)								
AG/CT	0	652	568	525	536	381	106	2,768
AT/AT	0	108	68	61	46	49	28	360
AC/GT	0	132	70	54	25	32	15	328
CG/CG	0	14	7	0	3	2	0	26
Tri- (2,261, 38.82%)								
AGG/CCT	316	116	43	4	1	0	0	480
AAG/CTT	258	136	67	6	0	0	1	468
AGC/CTG	252	99	39	2	0	0	0	392
CCG/CGG	245	85	39	3	0	0	0	372
ATC/ATG	96	37	21	3	0	0	0	157
ACC/GGT	84	42	20	3	1	0	0	150
Other	146	49	35	11	1	0	0	242
Tetra- (62, 1.06%)								
AAAG/CTTT	14	2	0	0	0	0	0	16
AGAT/ATCT	12	1	0	0	0	0	0	13
ACAT/ATGT	4	1	1	0	0	0	0	6
AAAT/ATTT	5	0	0	0	0	0	0	5
Others	17	3	1	1	0	0	0	22
Penta- (13,0.22%)	13	0	0	0	0	0	0	13
Hexa- (7, 0.12%)	3	0	2	1	0	0	1	7
Total	1465	1477	981	674	613	464	151	5825
Percentage	25.15%	25.36%	16.84%	11.57%	10.52%	7.97%	2.59%	100%

376
377   The type and frequency of EST-SSRs with different numbers of tandem repeats are
378 summarized in Table 3. Because mononucleotide repeats may not be accurate due to
379 sequencing errors and assembly mistakes, they were excluded from further analyses.
380 Results showed that the identified SSR type was not evenly distributed throughout all the
381 SSR-containing sequences of 'Rehmannii'. The dinucleotide repeat motifs were the most
382 abundant (3,482 or 59.8%), followed by trinucleotide repeat motifs (2, 261 or 38.8%),
383 whereas hexa- (62 or 1.06%), penta- (13 or 0.22%) and tetranucleotide repeat motifs (7 or
384 0.12%) were rare. The number of SSR repeats ranged from 5 to 24, with 6 repeats (1,477,
385 25.4%) being the most common, followed by 5 tandem repeats (1,465, 25.2%) and 7
386 tandem repeats (981, 16.8%). Motifs containing more than 10 repeats were rare (Table S3),

387 with a frequency of only 2.6% (151). Within these SSRs, 47 motif sequence types were
388 identified, of which di-, tri-, tetra-, penta-, and hexanucleotide repeats had 4, 10, 13, 13 and
389 7 types, respectively. AG/CT alone accounted for 79.5% (2,768) of the total dinucleotide
390 repeats, followed by AT/TA (360,10.4%) and AC/GT (328, 9.4%). Among the
391 trinucleotide repeats, GAA/CTT and AAG/CTT were the most abundant (480, 21.2%; 468,
392 20.7%). Other repeats, AGC/CTG, CCG/CGG, ATC/ATG and ACC/GGT, constituted
393 47.4% of the trinucleotide repeats.
394
395 Table 4 Summary of SNPs identified from unigenes of colored calla lily 'Rehmannii'.

Transitions	Number	Transversions	Number	
C/T	2,262	A/T	650	396
A/G	2,188	A/C	647	
		T/G	625	
		C/G	790	
Total	4,450	Total	2,712	

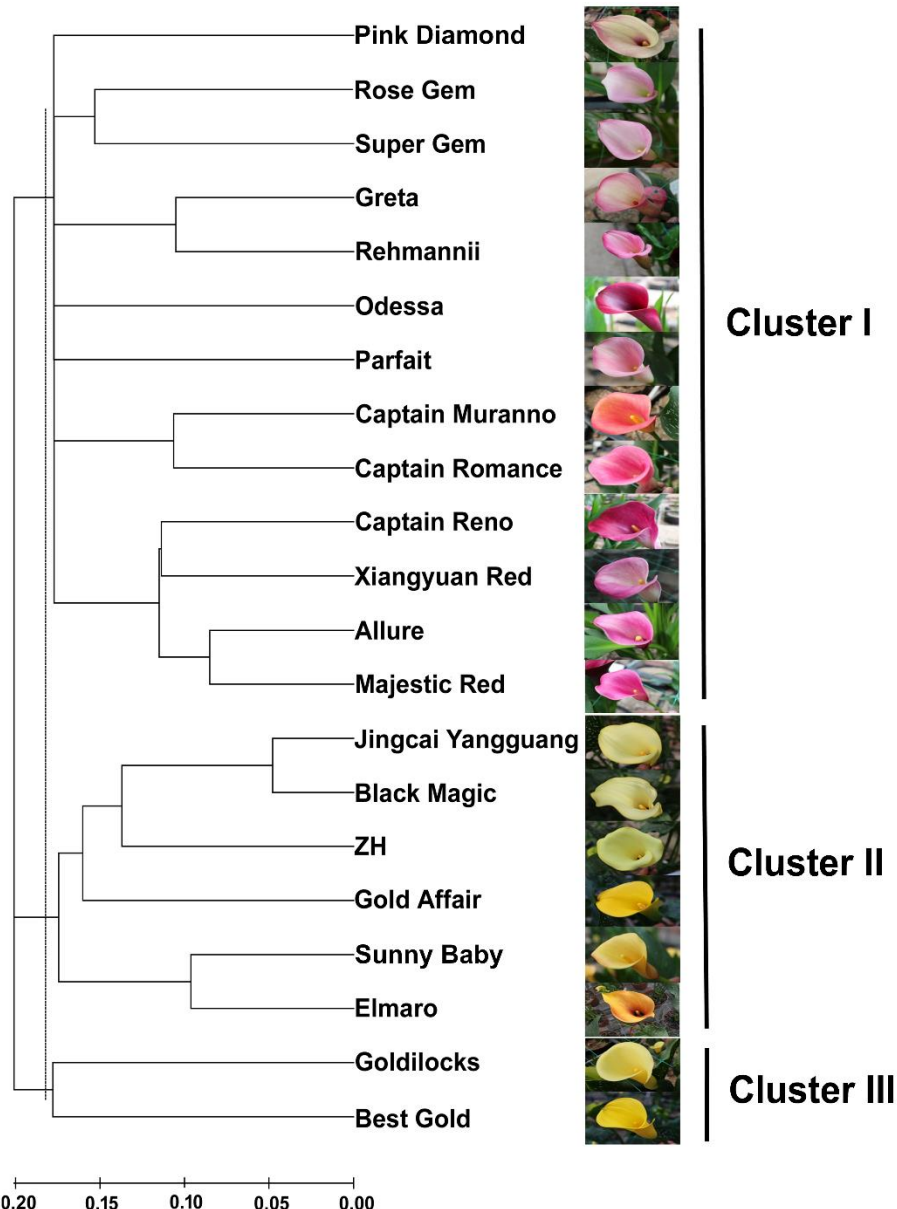
397
398   In addition to EST-SSRs, a total of 7,162 potential high-quality SNPs were identified by
399 mapping against 39,298 reference unigenes. The overall frequency of all types of SNPs
400 was one SNP per 5.69 kb. The predicted SNPs included 4,450 transitions and 2,712
401 transversions (Table 4). The most abundant SNPs detected were C/T (2,262, 31.6%),
402 followed by A/G (2,188, 30.6%) and C/G (790, 11.0%). The numbers of the remaining
403 three SNP types (A/T, A/C, and T/G) were similar, each accounting for less than 10%.
404 Potential SNPs are shown in Table S6.
405 **SSR primer design, polymorphism detection and phylogenetic analysis**
406 A total of 200 EST-SSR loci (repeat motif >1) with appropriate flanking sequences were
407 randomly selected for the design and synthesis of PCR primer pairs. Detailed information
408 of the EST-SSR markers is shown in Table S7. A germplasm panel of five colored calla
409 lily accessions (Rehmannii, Super Gem, Rose Gem, Xiangyuan Red and Allure) was
410 initially used to validate the usefulness of EST-SSR markers in monitoring
411 polymorphisms. A total of 137 (68.5%) of the primer pairs were successfully amplified by
412 PCR, while the remaining failed to generate any clear DNA products. Of the working
413 primer pairs, 77 (56.2%) produced clear PCR amplicons of the expected sizes, whereas 60
414 (43.8%) amplified non-specific products, of which 23 markers generated PCR products
415 larger or smaller than expected and 37 generated more than one band (Table S5).
416
417 Table 5 Characteristics of the 58 polymorphic EST-SSR markers in 21 colored calla
418 lily accessions.

Locus	Na	Ne	Ho	He	PIC	Locus	Na	Ne	Ho	He	PIC
CallaRe015	3	2.057	0.684	0.514	0.425	CallaRe110	2	1.324	0.286	0.245	0.215
CallaRe016	6	3.320	0.600	0.699	0.654	CallaRe117	3	2.410	0.000	0.585	0.513
CallaRe028	4	3.756	1.000	0.734	0.685	CallaRe118	4	2.028	0.667	0.507	0.462
CallaRe030	3	2.095	1.000	0.523	0.409	CallaRe120	4	2.932	0.333	0.659	0.593

CallaRe031	2	1.446	0.286	0.308	0.261	CallaRe128	4	3.556	0.550	0.719	0.670
CallaRe032	2	1.930	0.810	0.482	0.366	CallaRe129	2	1.984	0.545	0.496	0.373
CallaRe036	2	1.296	0.263	0.229	0.202	CallaRe131	4	2.930	0.286	0.659	0.601
CallaRe040	4	2.139	0.278	0.532	0.483	CallaRe135	2	1.220	0.200	0.180	0.164
CallaRe041	3	1.156	0.048	0.135	0.130	CallaRe144	4	2.766	0.588	0.638	0.589
CallaRe042	2	1.265	0.238	0.210	0.188	CallaRe146	3	2.085	0.952	0.520	0.408
CallaRe049	2	1.960	0.857	0.490	0.370	CallaRe147	3	1.841	0.353	0.457	0.411
CallaRe050	6	3.630	0.571	0.724	0.683	CallaRe151	3	2.455	0.667	0.593	0.505
CallaRe055	3	2.057	0.526	0.514	0.425	CallaRe155	2	1.724	0.000	0.420	0.332
CallaRe056	2	1.946	0.833	0.486	0.368	CallaRe156	4	2.309	0.667	0.567	0.486
CallaRe061	4	3.469	0.286	0.712	0.661	CallaRe160	2	1.893	0.000	0.472	0.360
CallaRe066	2	1.265	0.238	0.210	0.188	CallaRe165	2	2.000	1.000	0.412	0.375
CallaRe075	2	1.358	0.313	0.264	0.229	CallaRe166	4	2.520	0.333	0.603	0.541
CallaRe078	4	2.303	0.952	0.566	0.471	CallaRe170	3	2.597	0.737	0.615	0.536
CallaRe080	4	2.285	0.381	0.562	0.519	CallaRe175	6	4.762	0.800	0.790	0.757
CallaRe081	3	1.407	0.333	0.289	0.266	CallaRe178	2	1.835	0.700	0.455	0.351
CallaRe082	3	2.182	0.000	0.542	0.460	CallaRe179	2	1.992	0.188	0.498	0.374
CallaRe089	2	1.995	0.857	0.499	0.374	CallaRe180	3	1.340	0.190	0.254	0.237
CallaRe090	3	2.829	0.500	0.646	0.571	CallaRe185	2	1.637	0.412	0.389	0.314
CallaRe095	3	2.524	1.000	0.604	0.525	CallaRe187	2	1.960	0.000	0.490	0.370
CallaRe097	3	1.956	0.619	0.489	0.407	CallaRe189	3	2.111	0.550	0.526	0.431
CallaRe100	3	2.246	0.500	0.555	0.456	CallaRe190	2	1.498	0.316	0.332	0.277
CallaRe101	3	1.893	0.095	0.472	0.397	CallaRe191	2	1.205	0.188	0.170	0.155
CallaRe106	4	2.431	0.684	0.589	0.506	CallaRe194	2	1.600	0.500	0.375	0.305
CallaRe109	3	2.256	0.619	0.557	0.462	CallaRe198	3	2.492	0.650	0.599	0.514

419

420   Specific EST-SSR markers were used to assess the genetic diversity and relationships
421 among the 21 accessions of colored calla lily from New Zealand, the Netherlands, the
422 United States and China. Of these tested markers, 58 (or 75.3%) were polymorphic, while
423 the others were monomorphic. The polymorphic EST-SSRs consisted of 22 di-, 18 tri-, 6
424 tetra-, 4 penta-motif- and 8 compound-motif-based markers. The raw data and
425 characteristics of the 58 polymorphic EST-SSR markers in 21 accessions are listed in
426 Table S8 and Table 5. The number of alleles per marker (Na) ranged from 2 to 6, with 174
427 alleles in total. The average effective number of alleles per locus (Ne) was 2.163, with a
428 maximum of 4.762 and a minimum of 1.156. The observed heterozygosity (Ho) varied
429 from 0 to 1.000, whereas the expected heterozygosity (He) varied from 0.135 to 0.790. The
430 mean Ho and He values were 0.483 and 0.491, respectively. The polymorphic information
431 content (PIC) values ranged from 0.130 to 0.757 with a mean value of 0.420. Nr
432 annotation results (Table S7) showed that most of the polymorphic SSR-containing
433 sequences shared significant homology to protein-encoding genes from *V. vinifera*
434 (Vitaceae), *T. cacao* (Sterculiaceae), *Z. mays* and *O. sativa* Indica (Gramineae), among
435 others. These positive-hit homologous genes, including nucleic acid binding protein,

436 glycine-rich protein, abscisic acid, zinc-finger transcription factor, had hypothetical or
437 putative functions in defense and stress, transporters and metabolic processes.

438

439 Figure 3 An NJ dendrogram of 21 colored calla lily accessions based on 58 polymorphic
440 EST-SSR markers.

441

442 Based on the shared allele distance, we conducted a cluster analysis to assess the genetic
443 relationships between 21 accessions of colored calla lily (Fig. 3). The dendrogram clearly
444 grouped these accessions into three major clusters (I, II and III) at an approximate genetic
445 distance of 0.18. Cluster I consisted of 13 accessions collected from the Netherlands, the
446 United States, New Zealand and China. Twelve accessions in clusters I (Pink Diamond,
447 Rose Gem, Super Gem, Greta, Parfait, Rehmannii, Odessa, Majestic Red, Xiangyuan Red,
448 Captain Reno, Allure, and Captain Romance) had red, purple and pink spathes, whereas
449 Captain Murano had an orange spathe. Clusters II and III consisted of eight yellow-flower
450 accessions. The accessions in the former cluster (Jingcai Yangguang, Black Magic, ZH,

451 Gold affair, Sunny Baby and Elmaro) originated from New Zealand and China, while those
452 in the latter (Goldilocks and Butter Gold) were collected from New Zealand. The results
453 may indicate a potential association between the genetic relationship of the tested
454 accessions and the spathe color.
455
456 **DISCUSSION**
457 **Illumina transcriptome sequencing and *de novo* assembly**
458 *Zantedeschia* spp. are perennial bulb plants known worldwide for their ornamental uses.
459 These plants are even commonly used in traditional medicine in some African regions
460 (*Letty*, *1973*; *Singh*, *1996*). However, little genomic information is available for the plants
461 of the genus *Zantedeschia*, even with significant advances in DNA sequencing
462 technologies. The genomic researches of *Zantedeschia* spp. may be hindered by the large
463 genome size of these plants. The nuclear DNA contents of colored calla lily *Z. elliottiana*
464 and white calla lily *Z. aethiopica* (*Ghimire et al.*, *2012*) were estimated to be 1.17 ±0.50
465 and 3.72 ±0.10 pg DNA/2C, respectively, equivalent to genome sizes of 1.15 and 3.64
466 Gbp, approximately 2.5 and 7.8 times larger than the rice (*O. sativa*) genome of 0.47 Gb.
467 Moreover, *Zantedeschia* spp. may be a genetically high-heterozygosity bulb-flower
468 species. Since these species having been bred for more than one century, a larger number
469 of outstanding cultivars and hybrids of white and colored calla lilies in a range of sizes and
470 colors have been cross-bred, named and released by European and American breeders.
471 Thus, it is obviously not feasible to consider whole-genome sequencing for this perennial
472 plant because of the high cost and time consumption.
473   For organisms with large, heterozygous and complex genomes containing repetitive
474 sequences, RNA-seq-based *de novo* transcriptome analysis is an attractive alternative to
475 examine the properties of a transcriptome as a proxy for the whole genome (*Martin &*
476 *Wang*, *2011*; *Onda et al.*, *2015*). A transcriptome study of *Z. aethiopica* spathe (*Cândido et*
477 *al.*, *2014*) also demonstrated that transcriptome sequencing is a useful molecular biology
478 tool to rapidly build comprehensive sequence resources of expressed genes for in-depth
479 analysis at high resolution. To date, *de novo* transcriptome analysis has successfully been
480 applied to a wide range of plant species (*Martin & Wang*, *2011*), including crops, fruits,
481 vegetables, forest plants, flowers, and medicinal plants, for various purposes, such as
482 gaining fundamental insight into biological processes, generating different dynamic views
483 of interesting gene expression, discovering novel genes or transcripts, and developing and
484 validating molecular markers. In the present study, Illumina paired-end RNA-seq
485 technology was used to sequence the pooled RNA from ten tissues of 'Rehmannii'. More
486 than 59.9 million 101-bp paired-end reads were yielded, encompassing 6.1 G nucleotides.
487 Trinity software (*Haas et al.*, *2013*) was then used to *de novo* assemble these short reads to
488 generate a total of 39,298 unigenes with an average length of 1,038 bp. The results
489 indicated that *de novo* transcriptome assembly is a feasible strategy to provide genome
490 information for non-model plant species. In sequence assembly, the N50 value is an
491 important statistical measure that is used to evaluate transcriptome assemblies in which a
492 high number corresponds to high quality. The N50 length of the 'Rehmannii' unigenes was

493 1,476 bp, which is comparable to a recent report of the assembled transcriptome sequences

494 of *Z. aethiopica* (N50 = 1,600 bp) using a similar method (*Cândido et al.*, *2014*). However,

495 it should be noted that *Z. aethiopica* has approximately 1.5 times (91.2 million) as many

496 short reads as does 'Rehmannii'. These values are also larger than those reported for

497 related species in the Araceae family, such as *A. andraeanum* (N50 = 1172 bp), *A. konjac*

498 and *A. bulbifer* (N50 = 381 and 534; N50 = 372 and 524, respectively) (*Zheng et al.*, *2013*;

499 *Diao et al.*, *2014*). Previous reports have demonstrated that an accurate and effective

500 assembly tends to have a longer mean length and a larger N50 value (*Chen et al.*, *2015*),

501 which suggests the high quality of our colored calla lily transcriptome sequences.

502 **Functional annotation and classification of unigenes**

503 To predict the biological functions of as many assembled transcripts as possible, various

504 protein databases, including Nr, Swiss-Prot, COG, GO and KEGG, were employed. In

505 total, 21,077 unigenes (Table S3) showed significant hits in the above five databases. The

506 annotation rate of 'Rehmannii' unigenes was 53.6%, which was higher than that of *Z.*

507 *aethiopica* (39.7%) using same method. The higher percentage in this study may be

508 attributed to the higher frequency of unigenes longer than 500 bp in the assembled

509 transcripts. In general, the longer unigenes were more likely to have BLAST matches in

510 the protein databases (*Parchman et al.*, *2010*). It was estimated that the percent of

511 unigenes >500 bp in 'Rehmannii' and *Z. aethiopica* was 65.9% and 42.1%, respectively.

512 Novaes et al. (*2008*) reported that a high number of next-generation sequencing (NGS)

513 short reads often cannot be matched to known genes because the significance of sequence

514 similarity partially depends on the length of the query sequence. Therefore, the non-

515 significantly annotated unigenes that are likely novel transcripts unique to *Z. rehmannii*

516 'Rehmannii' may be too short to allow for statistically meaningful matches.

517  A species-based distribution of the best hits from the BlastX search against the Nr

518 database showed that 32.0% of the annotated sequences of 'Rehmannii' had similarity with

519 dicotyledonous grape *(V. vinifera)*. This finding was confirmed by other transcriptome

520 reports on *Z. aethiopica* (*Cândido et al.*, *2014*), *A. andraeanum* (*Yang et al.*, *2015*; *Li et al.*,

521 *2015*), *A. konjac* and *A. bulbifer* (*Zheng et al.*, *2013*). In fact, calla lily and grape are a

522 monocotyledonous herb and a dicotyledonous woody vine, respectively. So they are distant

523 from each other genetically and evolutionarily. One possible reason for this perceived

524 similarity is the absence of whole-genome sequences in public databases for any species of

525 Araceae. These assembled transcriptome sequences may provide an important data

526 resource for future studies on taxa-specific phenomena in the family Araceae. Consistent

527 with previous studies (*Cândido et al.*, *2014*; *Yang et al.*, *2015*; *Li et al.*, *2015*; *Zheng et al.*,

528 *2013*), the annotated unigenes were classified into 24 COG and 47 GO sub-terms or sub-

529 categories, indicating that our transcriptome data represented a broad diversity of

530 transcripts in colored calla lily. In addition, a total of 4,532 unigenes were annotated and

531 mapped to 117 KEGG pathways. Approximately 80% of top 20 hit pathways were

532 involved in genetic information processing and metabolism, while the others were related

533 to pathways involved in plant hormone signal transduction, plant-pathogen interaction,

534 phagosome, and endocytosis (Fig. 2C). The most highly represented pathway was related

535 to genetic information processing and metabolism, reflecting the fact that calla lily devotes
536 an enormous investment to gene transcription control and capacity, cell maintenance and
537 defense capacity (*Cândido et al., 2014*). Unigene annotation, together with the predicated
538 pathways, also facilitated the discovery of some key genes involved in flower development
539 and function in colored calla lily. In total, we identified 117 homologous sequences
540 involved in eight pathways, including the anthocyanin biosynthesis pathway, carotenoid
541 biosynthesis pathway, photoperiod pathway, vernalization pathway, gibberellic acid
542 pathway, autonomous or others. These captured unigenes again demonstrated that a
543 relatively accurate and high-coverage genomic database can be produced by RNA-seq-
544 based *de novo* transcriptome analysis for non-model plant species.

545 **Identification of EST-SSRarker frequency and type**
546 Markers based on expressed sequences are useful and attractive for the detection of
547 functional variation and gene-based analysis. However, the available EST-SSR markers are
548 insufficient for the genus *Zantedeschia* at the present time. Wei et al. (*2012*) identified 209
549 EST-SSRs from 2,175 non-redundant ESTs derived from cDNA libraries of developing
550 spathe in *Z. aethiopica*. Among these, a total of 166 primer pairs flanking the EST-SSRs
551 could be designed. Note here that these EST-SSRs were identified based on Sanger
552 sequencing. Very limited expressed sequence data can be generally produced via the
553 Sanger sequencing of cDNA libraries compared to RNA-based transcriptome sequencing.
554 In our study, 9,933 potential EST-SSRs were identified in the *de novo* transcriptome
555 sequences obtained by Illumina sequencing (Table 2). These data confirm that
556 transcriptome sequences are excellent resources for the development of numerous SSR
557 markers. In the present study, it was estimated that approximately 20.3% (7,997) of the
558 assembled unigenes possess SSR loci, and the abundance of SSRs was one SSR locus per
559 4.1 kb. The density of SSR-containing sequences in 'Rehmannii' was higher than that in
560 the transcriptome reports for other species in Araceae, such as *Amorphophallus* spp.
561 (11.8%) and *P. ternate* (16.24%) (*Wang et al., 2014*; *Zheng et al., 2013*). The frequency of
562 SSRs was higher than that in *P. ternate* (4.3 kb) (*Zheng et al., 2013*) but lower than that in
563 *Amorphophallus* spp. (3.6 kb) (*Wang et al., 2014*). The difference in the abundance
564 estimation and frequency of SSRs among various species could partially be due to the SSR
565 search criteria, the size of the unigene assembly dataset, the database-mining tools and the
566 sequence redundancy, in addition to actual differences between species (*Wang et al., 2014*;
567 *Zheng et al., 2013*; *Chen et al., 2015*).
568 The SSRs identified in this study were not uniformly distributed in the 'Rehmannii'
569 transcriptome database (Table 3). When mononucleotide repeats were excluded, di-
570 (59.8%) and trinucleotide repeats (38.8%) were the most abundant, whereas hexa- (1.1%),
571 penta- (0.2%) and tetranucleotide repeats (0.1%) were rare. This result is identical to the
572 previous findings of di- and trinucleotide motifs as the most frequent SSR motif types in
573 the transcriptome sequences of many other plants, including *Amorphophallus* spp. and *P.*
574 *ternate* (*Wang et al., 2014*; *Zheng et al., 2013*). As shown in Table 3, the AG/CT motif was
575 the most abundant dinucleotide repeat (27.9%), followed by AT/TA (3.6%) and AC/GT
576 (3.3%). The predominant AG/CT motif repeats were also observed in *Amorphophallus* spp.

577   and *P. ternate* (*Wang et al.*, *2014*; *Zheng et al.*, *2013*). The most abundant trinucleotide
578   repeat motif in calla lily was AGG/CCT, closely followed by AAG/CTT, similar to reports
579   in *Amorphophallus* spp. but different from those in *P. ternate* (*Wang et al.*, *2014*; *Zheng et
580   al.*, *2013*). This difference possibly arises from the SSR search parameters and search
581   algorithms. CCG/CGG was the most frequent trinucleotide motif in *P. ternate* but the third
582   and fourth most common repeat type in *Amorphophallus* spp. and colored calla lily
583   'Rehmannii', respectively. Anyhow, these abundance results corroborated with the
584   suggestion that the trinucleotide motif CCG/CGG is common in monocots. In addition, we
585   also noticed that GC-rich trinucleotide motifs (ACC/GGT, ACG/CGT, AGC/CTG,
586   AGG/CCT and CCG/CGG comprised > 65%) were more abundant than AT-rich
587   trinucleotide motifs (AAG/CTT, AAT/ATT, ACT/AGT, and ATC/ATG comprised < 35%)
588   in these three Araceae species. These results strongly support the fact that the high GC
589   content and consequent codon usage bias are specific features of monocot genomes
590   (*Morgante*, *Hanafey & Powell*, *2002*).
591   **Evaluation of genetic diversity and relationships among colored calla lily accessions**
592   Cultivar 'Rehmannii' in the present study is a hybrid of *Z. rehmannii*. It has the desired
593   attributes for potted flower purposes, such as pink spathes, perfect trumpet-shaped
594   inflorescence, lanceolate and semi-erect leaves, and high productivity. Similar
595   characteristics were also observed in another four cultivars, including Super Gem, Rose
596   Gem, Xiangyuan Red and Allure. Hence, all five varieties here were initially used to
597   evaluate EST-SSR marker usefulness in monitoring polymorphisms. A total of 200 primer
598   pairs were synthesized and tested, of which 137 (68.5%) successfully yielded amplicons in
599   these five cultivars. Of the abovementioned working primer pairs, 77 (56.2%) produced
600   PCR products of the expected fragment size used to screen for polymorphisms among 21
601   individual accessions. Finally, 58 polymorphic EST-SSR markers were obtained with a
602   polymorphic proportion of 75.3%. EST-SSR markers are advantageous to SSRs in non-
603   transcribed regions due to their higher amplification rates and cross-species transferability
604   (*Varshney*, *Graner & Sorrells*, *2005*). The results here indicate that these transcript-based
605   SSRs are conserved in the germplasms of colored calla lily, suggesting that they will have
606   a broad utilization in taxonomic and cultivar identification, as well as
607   comparative mapping. The amplification and polymorphic rate of the EST-SSRs developed
608   in our study (59.3%) is much higher than that obtained in *Z. aethiopica* using Sanger
609   sequencing (*Wei et al.*, *2012*), suggesting that the *de novo* transcriptome sequence based on
610   Illumina RNA-seq was accurate and of high quality. In *Z. aethiopica* (*Wei et al.*, *2012*), 68
611   (40.9%) of the EST-SSR primer pairs yielded PCR amplification products in 24
612   accessions, of which 43 (63.2%) exhibited polymorphisms. However, the result is lower
613   than that reported in *Amorphophallus* spp. (*Zheng et al.*, *2013*). In *A. konjac* and *A.
614   bulbifer*, 270 (84.4%) primer pairs produced amplicons in two wild accessions, and 205
615   (89.1%) EST-SSR markers were polymorphic between 25 wild and cultivated accessions.
616   The low amplification and polymorphism rate in colored calla lily may be attributed to
617   various factors (*Varshney et al.*, *2007*; *Varshney*, *Graner & Sorrells*, *2005*), including the
618   presence of introns in the corresponding cDNA, SNPs or InDels (insertion-deletions) in the

619   primers, assembly errors in the *de novo* transcriptome sequences, and the high
620   heterozygosity of the calla lily genome. Nonetheless, the failed amplification may be
621   remedied if the PCR amplification conditions are re-optimized, such as by applying a
622   lower annealing temperature and/or using gradient PCR.
623   Most calla lily cultivars have been bred following the intra/inter-specific hybridization
624   of the species *Z. rehmannii*, *Z. albomaculata*, *Z. elliottiana* and *Z. pentlandii* within the
625   section *Aestivae*, producing plants and flowers with a broad range of shapes and colors.
626   The species *Z. rehmannii*, *Z. albomaculata*, *Z. elliottiana* and *Z. pentlandii* are closely
627   related to each other based on cytogenetic karyotypes (*Yao et al.*, *1994b*). However, the
628   genetic relationship among commercial hybrids or varieties is still not well defined.
629   Characters based on which species or even varieties have been previously separated, for
630   example, the degree of spotting on the leaf, the presence of bristles on the petioles and
631   peduncles and, to some extent, the leaf shape, have been unreliable when a wide range of
632   hybrids is examined (*Letty*, *1973*; *Singh*, *1996*). Molecular markers, such as RAPD and
633   ISSR (*Hamada & Hagimori*, *1996*; *Zhang et al.*, *2009*; *Lu et al.*, *2012*; *Chen et al.*, *2013*;
634   *Lu et al.*, *2014*), therefore, were used for cultivar identification and to evaluate the genetic
635   diversity and relationship of germplasm resources of colored calla lily. However, very
636   limited and uncertain information has been obtained until now. EST-SSR markers facilitate
637   better cross-genome comparisons and genetic diversity and relationship evaluation because
638   their target coding domains are more likely to be conserved between relatives (*Morgante*,
639   *Hanafey & Powell*, *2002*; *Varshney*, *Graner & Sorrells*, *2005*). In the present study, 174
640   alleles were identified in 21 accessions using 58 polymorphic EST-SSR markers with an
641   average of 3.0 alleles per locus. The discriminating power, as determined by the PIC value,
642   ranged from 0.130 to 0.757 with a mean value of 0.420. Wei et al. (*2012*) reported 43 EST-
643   SSR markers with an average gene diversity of 0.446 (*PIC*) in 24 *Z. aethiopica* individual
644   plants. The polymorphism level here was comparable to that of the EST-SSR based study
645   in *Z. aethiopica*.
646   A cluster analysis based on 58 polymorphic EST-SSR markers was then used to assess
647   the genetic relationships among 21 accessions. Three major groups were identified at a cut-
648   off genetic distance index of 0.18. Cluster I consisted of thirteen accessions with a complex
649   set of colored spathes, including red, purple, and pink, among others. The cultivated
650   species *Z. rehmannii* cv. Rehmannii and another four cultivars (Super Gem, Rose Gem,
651   Xiangyuan Red and Allure) used in the initial validation of EST-SSR markers were
652   grouped into several sub-clusters. These sub-clusters may reflect that *Z. rehmannii* is easily
653   crossed with other species (*Letty*, *1973*; *Funnell*, *1993*; *Singh*, *1996*), giving rise to hybrids
654   with lobed leaves and spathes that vary in color and shape. Surprisingly, eight accessions
655   with yellow flowers were separated into Clusters II and III. The cultivars Black Magic in
656   Cluster II and Best Gold in Cluster III are intra-species hybrids of *Z. elliottiana* and *Z.*
657   *pentlandii*, respectively, representing the close relationship between Clusters II and III, as
658   well as *Z. elliottiana* and *Z. pentlandii*. The species *Z. elliottiana* and *Z. pentlandii*, also
659   known a**s** Yellow Arum (*Letty*, *1973*; *Singh*, *1996*), have leaves that are hastate to cordate
660   at the base and differ in their consistency of yellow spathes. These species are distant from

661 the other species *Z. albomaculata* and *Z. rehmannii*. Ronald et al. (*2007*) determined the

662 plastome composition of *Z. aethiopica*, *Z. rehmannii*, *Z. albomaculata* subsp.

663 *albomaculata*, *Z. albomaculata* subsp. *macrocarpa*, *Z. elliottiana* and *Z. pentlandii* with

664 species-specific CAPS markers developed from the plastidial intergenic region of *trn*D and

665 *trn*C (DC). They revealed that *Z. elliottiana* and *Z. pentlandii* showed a DC-*Alu*I and a DC-

666 *Hae*III restriction pattern that differed from that of *Z. rehmannii* and *Z. albomaculata*. The

667 results presented here also support the above findings and suggest that EST-SSR makers

668 developed from *de novo* transcriptome analysis are potential tools for taxonomy and

669 cultivar identification in the genus *Zantedeschia*. Interestingly, the genetic relationships

670 among colored calla lily accessions seemly related to the spathe color. However, we should

671 note that this conclusion was drawn from limited numbers of *Zantedeschia* accessions in

672 section *Aestivae*. A higher number of accessions and more individual cultivated plants will

673 be essential for verifying the abovementioned relationship in future studies.

674

675 **CONCLUSIOINS**

676 This study was an attempt to present the transcriptome of colored calla lily using Illumina

677 next-generation sequencing and *de novo* assembly. A total of 39,298 unigenes with an

678 average length of 1,038 bp were generated, of which 53.6% (21,077) were annotated using

679 the Nr, Swiss-Prot, COG, GO and KEGG databases. Based on the transcriptome dataset,

680 we identified a total of 117 unique transcripts related to flower development, including

681 pigment biosynthesis, floral organ development, flowering regulation and flower

682 senescence. Moreover, a large number of SSRs and SNPs were mined and identified, and

683 high-quality primers of 200 SSR loci were designed and demonstrated for their

684 amplification and cross-species transferability in germplasm resources of colored calla lily.

685 Finally, a relatively distinct genetic relationship among 21 accessions in section *Aestivae*

686 was elucidated via 58 EST-SSR markers. The enrichment results highlight the potential of

687 a *de novo* transcriptome dataset for functional genomics studies and molecular marker

688 development. The EST-SSR markers generated in this study will enhance the current

689 repository for the genus *Zantedeschia* and will be useful for taxonomic study and crop

690 improvement programs.

691

692 **REFERENCES**

693 **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP,**

694  **Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis**

695  **A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G**.

696  2000. Gene ontology: tool for the unification of biology. Nature Genetics 25:25–29

697  DOI 10.1038/75556.

698 **Cândido EDS, Fernandes GDR, Alencar SAD, Cardoso MHES, Lima SMDF,**

699  **Miranda VDJ, Porto WF, Nolasco DO, Oliveira-Júnior NGD, Barbosa AEAD,**

700  **Pogue RE, Rezende TMB, Dias SC, Franco OL. 2014**. Shedding some light over

701  the floral metabolism by arum lily (*Zantedeschia aethiopica*) spathe *de novo*

702  transcriptome assembly. *PLoS ONE* **9**:e90487 DOI 10.1371/journal.pone.0090487.

703   **Chen H, Liu L, Wang L, Wang S, Somta P, Cheng X. 2015.** Development and
704   Validation of EST-SSR Markers from the Transcriptome of Adzuki Bean (*Vigna*
705   *angularis*). *PLoS ONE* **10**:e0131939 DOI 10.1371/journal.pone.0131939.
706   **Chen Z, Xu BL, Pu CJ, Xu Q, Tian G. 2013.** Physiological and Biochemical and
707   resistance changes and ISSR polymorphic analysis exposed to $^{12}C^{6+}$ heavy ion
708   radiation on calla lily. *Journal of Nuclear Agricultural Sciences* **27**:0552–0556 DOI
709   10.11869/hnxb.2013.05.0552.
710   **Diao Y, Yang CZ, Yan M, Zheng XF, Jin SR, Wang YW, Hu ZL. 2014.** *De novo*
711   transcriptome and small RNA analyses of two *Amorphophallus* species. *PLoS One* **9**:
712   e95428 DOI org/10.1371/journal.pone.0095428.
713   **Funnell KA, MacKay BR. 1999.** Directions and challenges of the New Zealand calla
714   industry, and the use of calcium to control soft rot. In *International Symposium on*
715   *Development of Bulbous Flower Industry*; Sheen TF, Chen JJ, Yang TC, Eds.; Taiwan
716   Seed Improvement and Propagation Station: Taichung, Taiwan, 30–31.
717   **Funnell KA. Zantedeschia. 1993.** In *The physiology of flower bulbs*. Hertogh AD, Nard
718   ML, Eds.; Elsevier: Amsterdam, Netherlands**,** 683–703.
719   **Ghimire BK, Yu CY, Kim HJ, Chung IM. 2012.** Karyotype and nucleic acid content in
720   *Zantedeschia aethiopica* Spr. and *Zantedeschia elliottiana* Engl.. *African Journal of*
721   *Biotechnology* **11**:11604–11609 DOI 10.5897//AJB12.06.
722   **Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger**
723   **MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi**
724   **F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD,**
725   **Friedman N, Regev A. 2013.** *De novo* transcript sequence reconstruction from RNA-
726   seq using the Trinity platform for reference generation and analysis. *Nature Protocols*
727   **8**:1494–1512 DOI 10.1038/nprot.2013.084.
728   **Hamada K, Hagimori M. 1996.** RAPD-based method for cultivar-identification of calla
729   lily (*Zantedeschia* spp.). *Scientia Horticulturae* **65:**215–218 DOI 10.1016/0304-
730   4238(95)00869-1.
731   **Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. 2004**. The KEGG resource
732   for deciphering the genome. *Nucleic Acids Research* **32:**277–280.
733   **Letty C. 1973.** The genus Zantedeschia. *Bothalia* **11**:5–26 DOI
734   10.4102/abc.v11i1&2.1963.
735   **Li Z, Wang JB, Zhang XQ, Xu L. 2015.** Comparative transcriptome analysis of
736   Anthurium "Albama" and its anthocyanin-loss mutant. *PLoS One* **10:**e0119027 DOI
737   10.1371/journal.pone.0119027.
738   **Liu K, Muse SV. 2005.** PowerMarker: an integrated analysis environment for genetic
739   marker analysis. *Bioinformatics* **21**:2128–2129.
740   **Lu B, Zheng YH, Peng F, Shu XC, Chen XX. 2012.** Optimization of RAPD reaction
741   system by uniform design on *Zantedeschica* hybrid. *Northern Horticulture* **11**:123–
742   126.
743   **Lu Bo, Zheng YH, Chen M, Peng F, Shu XC, Gao F. 2014.** Effects of 60 Co-γ ray
744   radiation on color Lily Parfait and their physiological responses to high temperature
745   and humidity. *Journal of Nuclear Agricultural Sciences* **28**:1353–1357 DOI
746   10.11869/j.issn.100-8551.2014.08.1353.

747   **Martin JA, Wang Z. 2011.** Next-generation transcriptome assembly. *Nature Reviews*
748   *Genetics* **12**:671–682 DOI 10.1038/nrg3068.
749   **Morgante M, Hanafey M, Powell W. 2002.** Microsatellites are preferentially associated
750   with non-repetitive DNA in plant genomes. *Nature Genetic* **30**:194–200.
751   **Novaes E, Drost D, Farmerie WG, Pappas GJ, Grattapaglia D, Ronald R, Kirst M.**
752   **2008.** High-throughput gene and SNP discovery in *Eucalyptus grandis*, an
753   uncharacterized genome. *BMC Genomics* **9**:312 DOI 10.1186/1471-2164-9-312.
754   **Onda Y, Mochida K, Yoshida T, Sakurai T, Seymour RS, Umekawa Y, Pirintsos SA,**
755   **Shinozaki K. 2015.** Transcriptome analysis of thermogenic *Arum concinnatum*
756   reveals the molecular components of floral scent production. *Scientific Reports* **5**:8753
757   DOI 10.1038/srep08753.
758   **Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA. 2010.**
759   Transcriptome sequencing in an ecologically important tree species: assembly,
760   annotation, and marker discovery. *BMC Genomics* **11**:180 DOI 10.1186/1471-2164-
761   11-180.
762   **Peakall R, Smous PE. 2006.** GenAlEx 6: Genetic analysis in Excel. Population genetic
763   software for teaching and research. *Molecular Ecology Notes* **6**:288–295 DOI
764   10.1111/j.1471-8286.2005.01155.x.
765   **Roman LT, Michael YG, Darren AN, Eugene VK. 2000.** The COG database: a tool for
766   genome-scale analysis of protein functions and evolution. *Nucleic Acids*
767   *Research* **28**:33–36.
768   **Singh Y. 1996.** *Contributions to the systematic of the genus Zantedeschia*
769   *Spreng.(Araceae)*; University of Pretoria Press: Pretoria, South Africa, 169.
770   **Snijder RC, Santiago BF, Tuyl VJM. 2007.** The role of plastome-genome
771   incompatibility and biparental plastid inheritance in interspecific hybridization in the
772   genus *Zantedeschia* (Araceae). *Floriculture and Ornamental Biotechnology* **1**:150–
773   157 DOI 10.1007/BF00225906.
774   **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011.** MEGA5:
775   Molecular evolutionary genetic analysis using maximum likelihood, evolutionary
776   distance, and maximum parsimony methods. *Molecular Biology and Evolution*
777   **28**:2731–2739 DOI 10.1093/molbev/msr121.
778   **Tian DQ, Pan XY, Yu YM, Wang WY, Zhang F, Ge YY, Shen XL, Shen FQ, Liu XJ.**
779   **2013.** *De novo* characterization of the Anthurium transcriptome and analysis of its
780   digital gene expression under cold stress. *BMC Genomics* **14**:1–14 DOI
781   10.1186/1471-2164-14-827.
782   **Tjia B. 1985.** Hybrid calla lilies: A potential new crop for Florida. Proceedings of
783   the *Florida State Horticultural Society* **98**:127–130.
784   **Varshney RK, Graner A, Sorrells ME. 2005**. Genic microsatellite markers in plants:
785   features and applications. *Trends in Biotechnology* **23**:48–55.
786   **Varshney RK, Mahendar T, Aggarwal RK, B örner A. 2007.** Genic molecular markers
787   in plants: development and applications. In *Genomics-assisted crop improvement*;
788   Varshney, R.K., Tuberosa, R. Eds.; Springer Netherlands: Belgium, Germany, 13–29.
789   **Wang S, Zhang Z, Jiang NH, Zhang GH, Sha BC, Yang SC, Chen JW. 2014.** SSR
790   information in transcriptome of *Pinellia ternata*. *Zhong Yao Cai* **37:**1566–1569 DOI
791   10.13863/j. issn1001-4454.2014.09.015.
792   **Wei ZZ, Luo LB, Zhang HL, Xiong M, Wang X, Zhou D. 2012.** Identification and
793   characterization of 43 novel polymorphic EST-SSR markers for arum lily,
794   *Zantedeschia aethiopica* (Araceae). *American Journal of Botany* **99:**e493–497 DOI
795   10.3732/ajb.1200228.

796 **Yang YX, Chen XX, Xu B, Li YX, Ma YH, Wang GD. 2015.** Phenotype and
797   transcriptome analysis reveals chloroplast development and pigment biosynthesis
798   together influenced the leaf color formation in mutants of *Anthurium andraeanum*
799   'Sonate'. *Frontiers in Plant Science* **6:**139 DOI 10.3389/fpls.2015.00139.
800 **Yao JL, Cohen D, Rowland RE. 1994.** Karyotype studies in the genus *Zantedeschia*.
801   *South African Journal of Botany* **60:**4–7.
802 **Yao JL, Cohen D, Rowland RE. 1995.** Interspecific albino and variegated hybrids in the
803   genus *Zantedeschia*. *Plant Science* **109**:199–206 DOI 10.1016/0168-9452(95) 04163-
804   O.
805 **Yao JL, Cohen D, Rowland RE. 1994.** Plastid DNA inheritance and plastome-genome
806   incompatibility in interspecific hybrids of *Zantedeschia* (Araceae). *Theoretical*
807   *and Applied Genetics* **88**:255–260 DOI 10.1007/BF00225906.
808 **Zhang YC, Tang GG, Chu YX, Wu AZ. 2009.** Optimization of ISSR reaction system
809   and preliminary study on *Zantedeschia*. *Molecular Plant Breeding* **7:**827–832 DOI
810   10.3969/mpb.007.000827.
811 **Zheng XF, Pan C, DiaoY, You YN, Yang CZ, Hu ZL. 2013.** Development of
812   microsatellite markers by transcriptome sequencing in two species of *Amorphophallus*
813   (Araceae). *BMC Genomics* **14:**490 DOI 10.1186/1471-2164-14-490.