EST ANALYSIS IN *CALAMUS MANAN* MIQ.

Nadarajah, K.[1], Choong, C.Y.[2*], Leong, S.J.[2], Thi, B.K.[2]
Hedley, P.[3] and Waugh, R.[3]

[1]School of BioScience and Biotechnology, Faculty of Science and Technology
[2]School of Environmental and Natural Resources, Faculty of Science and Technology, UKM; and
[3]Scottish Crop Research Institute, Dundee, Scotland
*E-mail: cychoong@pkrisc.cc.ukm.my

ABSTRACT

For the first time, eight floral cDNA libraries (four male viz. ECM01, ECM02, ECM03, ECM04 and four female viz. ECF01, ECF02, EFC03, ECF04) were constructed from different developmental stages of male and female inflorescences, leaf (ECL01) and fruit (ECFR01) of *Calamus manan*. The floral libraries were generated to identify floral ESTs that may be developed into DNA markers for sex determination in *C. manan*. The insert sizes of cDNA clones from ECF02 (18,000 clones), ECM02 (6,359 clones), ECM01 (5,000 clones), ECF01 (1,179 clones), ECF03 (5,000 clones), ECF04 (9,108 clones), ECM03 (5,980 clone), ECM04 (5,318 clones), ECL01 (3,840 clones) and ECFR01 (768 clones) ranged from 0.5 to 2.5 kb with an average insert size of 1.0 kb. A total of 1529 floral ESTs generated randomly from the above floral libraries were analyzed and compared against 371 fruit ESTs and 338 leaf ESTs. The floral ESTs were grouped into 4 categories in term of protein matching: (1) significant with known function (49.8%), (2) significant with unknown function & hypothetical proteins (7.8%), (3) not significant (34.1%) and (4) no match (8.3%). Further analysis on floral ESTs from group 1 yielded 8 categories of functional genes: metabolism (30.4%), protein synthesis (28.0%), organization and cellular function (23.8%), defense and stress regulation (12.6%), flowering (0.4%), development (1.4%) and miscellaneous (3.4%). Meanwhile a major portion of the leaf ESTs (29.7%) were involved in photosynthesis and most of the fruit ESTs had unknown gene function (27%). In addition, 11.3% of the ESTs generated from the fruit, floral and leaf cDNA showed no significant homology to any entry in the Genbank. A total of 3.3% of group 1 and 2 floral ESTs were floral genes. The floral genes identified were the MADS box gene 8, protein stamen specific *fil*1, CONSTANS, FRIGIDA, stigma/stylar cysteine-rich adhesion precursor, flower-specific gamma-thionine precursor, anther-specific proline-rich protein and Early Flowering 5.

Keywords: *Calamus manan*, inflorescence, EST, cDNA library

INTRODUCTION

Calamus manan Miq. or *rotan manau* is a species of large diameter rattan that is of commercial importance in the furniture industry. This rattan variety is found in the lowland dipterocarp forests in Perak, Selangor, Kelantan, Pahang, Terengganu, Negeri Sembilan, Sumatra and south of Thailand (Dransfield 1979). The dioecious *C. manan* possesses female and male inflorescences on separate trees. As is common with dioecious plants, the production of fruit is largely dependent on the sex ratio in a given area.

Since the male and female *C. manan* plants have similar leaf and stem morphology, the identification of male and female plants can only be done when inflorescence are produced. This occurs when the plants are 5-6 years old. Therefore it is impossible to design a seed orchard with the desired male to female ratio. A suitable male to female ratio needs to be maintained in a selection programme to generate planting materials for plantations. According to Aminuddin (1994), the ratio of male to female in a *C. manan* population is 4:1. However, a higher female to male ratio is preferred as it is the female that produces the seeds to maintain the generation.

Thus far very little has been done in studying the molecular biology of *C. manan*. The genetic diversity of *C. manan* has been studied using isozyme markers (Bon et al. 1994; Wickneswari et al. 1995). Information on the molecular biology of flowering in *C. manan* is important in understanding the development of its flowers and to identify the genes that are expressed in the various stages of flower development as well to identify those that are being expressed specifically in either sex. This knowledge is essential for the development of probes and markers that may be used in screening a population. Development of probes or markers especially for sex determination will be beneficial for designing a seed orchard with high productivity and good seeds quality.

Flower development has been studied in the model plant *Arabidopsis thaliana* and *Antirrhinum majus*. The understanding of the molecular architecture of flower development has led to the identification of a group of genes, referred to as the MADS box genes, which are involved in the regulation of the flowering process (Weigel & Nilsson 1995; Theissen et al. 2000). These MADS box genes contain organ and meristem identity genes that are expressed in the inflorescence development (Theissen et al. 2000). Most likely, we believe that similar sets of genes are involved in the regulation of the flowering in *C. manan.* The expression profiles of meristem identity genes and especially organ identity genes that control the development of the male and female flowers will be beneficial in generating probes that can be used in sex determination of *C. manan.*

In this research we have constructed eight cDNA libraries for each inflorescence development stage in the male and female plants. These libraries were used to generate ESTs that will be screened for floral organ and meristem specific

genes. Organ specific genes are genes that regulate development of different floral organs while meristem specific genes are genes expressed during early stage of flower development and regulate meristem development prior to the activity of organ specific genes. These genes may be used in the development of molecular markers for sex determination in the early stage of development in *C. manan*. These markers when generated will provide a valuable tool in designing commercial seed orchard.

In addition, one cDNA library each from fruit and leaf tissue was constructed. The fruit and leaf ESTs generated were compared with floral ESTs to look into the expression profiles of genes in these tissues. The fruit and leaf ESTs are expected to have the same sets of genes that are found in the floral ESTs but there will be differences in the expression levels of the genes. These findings may lead to a better understanding of the biological processes which occurs in various tissues as well as the relationship between these biological activities in different organs.

MATERIALS AND METHODS

Plant materials

The male and female inflorescences, fruit and leaf tissues were obtained from compartments F28 and F41, Bukit Lagong Forest Reserve, Forest Research Institute Malaysia (FRIM), Kepong, Kuala Lumpur.

Construction of cDNA library

Total RNA was extracted using a modified RNeasy Midi Kit (Qiagen, Germany) method. mRNA was isolated from total RNA using Dynabeads (Dynal, Norway). The first strand of the cDNA was synthesized using the *Ready to Go You Prime 1st Strand Bead* (Amersham, USA) and an adaptor for the *Not*I primer (Life Technologies, USA). The second strand of the cDNA was synthesized using the *Superscript™ Plasmid System with Gateway™ Technology for cDNA Synthesis and Cloning Kit* (Life Technologies, USA). The double stranded cDNA obtained was purified using the *MinElute PCR Purification Kit* (Qiagen). A *Sal*I adaptor (Life Technologies) was added to the 5' cDNA to enable ligation into cloning vector. The cDNA was purified once again with the *MinElute PCR Purification Kit*. The cDNA was then digested with *Not*I and cDNA above 500 bp was purified via *MinElute Gel Extraction Kit* (Qiagen) and cloned into pSPORT1 vector. This construct was then electroporated into ElectroMAX™ DH10B™ competent cells (Invitrogen, USA) and plated on a LB/ampicilin (0.05 mg/ml)/X-gal (40 mg/ml) plates and incubated overnight at 37°C .

cDNA sequencing and analysis

From the cDNA libraries a total of 3456 clones were randomly selected and used in large scale plasmid preparation using *Montage Plasmid Miniprep96* and *MultiScreen Separation System* (Millipore, USA). The sequencing was performed from the 5' end using the ABI PRISM 3700 System (Applied Biosystems) with the M13R primer (Invitrogen, USA). The sequencing and sequence analysis of ESTs generated were conducted at the Computational Biology Laboratory, Scottish Crop Research Institute (SCRI, Dundee Scotland) and the Genomics Laboratory, BioValley UKM-MTDC. The ESTs generated were edited automatically in a batch by Phred20 (Bouck et al. 1998). The vector and adaptor sequences were edited out automatically via stackPACK. The cDNA sequences (ESTs) were matched against the databases in Genbank via the BLAST technique (Altschul et al. 1994). The BLASTX and BLASTN programmes were used to obtain a match against proteins and nucleic acids respectively. The EST sequences were also compared against the available dbEST at NCBI database (http://www.ncbi.nlm.nih.com). Following alignment to available databases in Genbank, the ESTs were classified into taxonomical groups based on protein function via the modified MIPS Functional Catalogue Database (http://mips.gsf.de/proj/ funcatDB/search_main_frame.html). Each EST was certified as significant when there was more than 50% sequence identity, an Expected value of $\leq e^{-10}$ and a score of ≥ 100.

RESULTS AND DISCUSSION

EST analysis

cDNA libraries were constructed using the RNA obtained from tissues of various inflorescence stages, fruit and leaf in *C. manan.* Ten cDNA libraries were synthesized and details of each library is given in Table 1. The ten cDNA libraries have a total of 60,552 clones. The average insert size for the libraries is 1 kb.

Two thousand six hundred and eighty eight clones (192 clones were selected from ECM01, 96 clones were selected from ECF01, another 96 clones from ECF02, and 1152 clones from ECM04 and ECF03 respectively) from the male and female inflorescence libraries and another 384 clones each from the fruit (ECFR01) and leaf (ECL01) library were randomly selected for sequencing for use in the characterization of expression and regulation of the flowering genes. A total of 3456 clones from the above libraries were subjected to a 5' end single pass sequencing which produced 2063 readable EST sequences. A 5' end single pass sequencing was conducted to maximize information and to avoid the poly A tails. A total of 74% (1529 ESTs) of the inflorescence, 88% (338 ESTs) of leaf and 96% (371 ESTs) of fruit ESTs generated were of good quality and the EST size was estimated at 400 bp once the low quality sequences and vector sequences were trimmed out. The generated EST sizes were compared to works done with *Brassica campestris* (~320 bp) and *Lotus japonicus* (~380 bp) (Lim et al. 1996; Endo et al. 2000) and this therefore indicated that the qualilty of sequences obtained were

comparable to other EST works. ESTs larger than 150 bp reduced the data analysis ambiguity to ~4%. The larger the EST, the more information the sequence will produce (Matsubara & Okubo 1993).

Table 1. Details of *C. manan* floral, fruit and leaf cDNA libraries.

Tissue type	Stage of development	Library	Insert size (kb)	Number of Clones
Female inflorescence	Main inflorescence branch	ECF03	0.5 - 1.5	5,000
	First inflorescence branch	ECF04	0.5 -1.5	9,108
	Second branch early development	ECF02	0.5 - 2.5	18,000
	Second branch late development	ECF01	0.5 - 1.5	1,179
Male inflorescence	Main branch inflorescence	ECM04	0.5 - 1.5	5,318
	Second branch inflorescence	ECM02	0.5 - 2.5	6,359
	Third branch early development	ECM03	0.5 - 1.5	5,980
	Third branch late development	ECM01	0.5 - 1.8	5,000
Fruit	Immature fruit of female plant	ECFR01	0.5-1.0	768
Leaf	Young leaf of female plant	ECL01	0.5-1.0	3,840

The clustering or assembly process on 1529 floral ESTs generated a total of 229 contigs (consist of 805 ESTs) and 724 singletons. The combined set of these contigs and singletons gave a total of 953 assembled sequences representing putative transcripts found in *C. manan* inflorescence cDNA libraries. This clustering result is summarized in Table 2. The assembly analysis resulted in a 24% redundancy. In the other EST projects which generated large number of ESTs, such as in *A. thaliana* (Höfte et al. 1993) and *Citrus sinensis* (Bausher et al. 2003), the redundancy rates were higher. However, this redundancy value is only an approximation as there may have been some errors caused by sequencing, chimeric cDNAs, truncation, and the presence of non-overlapping bases of the same gene that may affect the EST assembly and consequently the EST redundancy. There was one contig with 80 ESTs (alpha/beta hydrolase; E-value $6e^{-8}$ and score 137) which was the largest contig in this dataset. The smallest contig contained 2 ESTs (139 contigs), and 10 contigs had 10 sequences or more.

The BLASTX analysis of the ESTs resulted in the classification of the 1529 ESTs into four groups as summarized in Figure 1. Group 1 was defined as significant match group that contained protein sequences with known function in the NCBI database with an E-value of $\leq e^{-10}$ and score values of ≥ 100 consists of 761 ESTs (143 contigs, 315 singletons). This group represented 49.8% of the 1529 ESTs. Group 2 was the group that showed significant matches to unknown or hypothetical protein sequences and has the same E-value and score as those of group 1. Group 2 consisted of 119 ESTs (15 contigs, 78 singletons), which is

equivalent to 7.8% of the inflorescence ESTs. Therefore we can conclude that 880 ESTs (57.6%) of the 1529 ESTs from this dataset had significant matches to protein sequences found in the NCBI database. This value is fairly high as more than half of the total ESTs showed significant matches. The number of significant hits for these matches is within the 50 – 70% region as observed over total number of ESTs in other EST projects, such as *Lotus japonicus* flower buds EST set (Endo et al. 2000) and *Medicago truncatula* EST set (Covitz et al 1998) where their significant matches were 58.5% and 67.1% respectively.

Table 2. Summary of clustering result of *C. manan* floral ESTs.

Clustering result	Quantity
Contigs	229
Singletons	724
Assembled sequences	953
Redundancy	24%

With the setting of E-value $>e^{-10}$ and score <100, 34.1% of the 1529 ESTs (62 contigs, 242 singletons) representing 428 sequences did not matched significantly to protein sequences from other organisms in the NCBI database and this group of ESTs was defined as group 3. About 76.4% of these ESTs showed homology to other plant ESTs while 15.3% had homology to other genes classified in the database [3.3% homology to simple eucaryotes, 0.9% to invertebrates and 11.1% to vertebrates such as human (*Homo sapiens*), cow (*Bos taurus*), fish (*Gonostoma gracile*), cat (*Felix catus*) and rat (*Mus musculus*)]. The remaining 8.3% (127 ESTs) with 15 contigs and 89 singletons were not homologous to any of the above groups and were classified as unknown. These ESTs were classified into the last group which was defined as group 4. This group of ESTs may contain some genes of interest that may be involved in the regulation of flowering. Further studies will be conducted on this set of ESTs to predict putative functions for the EST through motif and domain analysis.

Group 1 ESTs were then assigned putative gene functions based on the initial BLASTX matches using modified MIPS Functional Catalogue Database (Table 3). It showed that group 1 ESTs were largely related to the "metabolism" function with a total of 231 ESTs. 213 ESTs were linked to proteins involved in "protein synthesis" and 171 ESTs were related to "organisation and cellular function". This was followed by ESTs involved in "defence and stress regulation" (84 ESTs), "miscellaneous" (26 ESTs), "development" (11 ESTs) and "flowering" (25 ESTs). The result is shown in Table 3 and summarized in percentage in Figure 2. Both male and female inflorescence ESTs that were generated had similar profiles in the expressed genes.

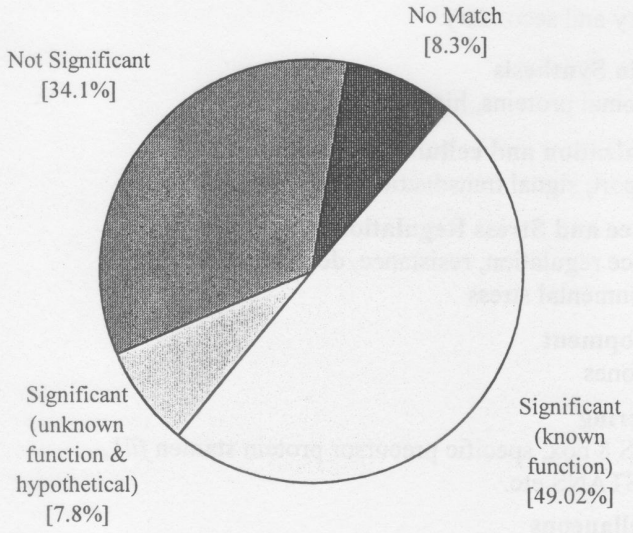

Figure 1. 1529 ESTs from *C. manan* inflorescence libraries categorized into four groups: significant with known function, significant with unknown function & hypothetical proteins, not significant and no match.

The highly abundant transcripts found in this dataset showed high levels of similarity to genes related to metabolism/energy category such as dehydrogenase (various types). This finding suggests that cells in inflorescence tissues are metabolically active, similar to the observation made in EST analysis of Chinese cabbage flower bud (Lim et al. 1996). Other abundant ESTs were the ABC-type transporter, 60S ribosomal and 40S ribosomal proteins, histone and transcriptional regulator. Most of these transcripts contain more than one subunit of these proteins, have more than one copy of the gene or are transcribed at a high level.

Eight floral genes were identified from the inflorescence libraries of *C. manan*. These floral-related ESTs were MADS 8 (e^{-47}), stamen specififc *fil*1 ($2e^{-12}$), CONSTANS ($3e^{-34}$), stigma/stylar cysteine-rich adhesin precursor (e^{-28}), Men-7 (e^{-7}), flower-specific gamma-thionine precursor ($9e^{-19}$), FRIGIDA (e^{-6}), anther-specific proline-rich protein (0.022) and Early Flowering 5 (0.17). However, floral genes related ESTs for FRIGIDA, anther-specific proline-rich protein and early flowering 5 were classified in group 3 (not significant match) due to the high E-value ($>e^{-10}$). Nevertheless these ESTs may be valuable for use in sex differentiation. Thus, further study is needed to investigate the presence of this gene in the male and female plants.

Table 3. Putative functional groups of floral ESTs.

Functional group	Numbers of ESTs
Metabolism Primary and secondary	231
Protein Synthesis Ribosomal proteins, histones and regulator	213
Organization and cellular function Transport, signal transduction and ubiquitination	171
Defence and Stress Regulation Defence regulation, resistance, detoxification and environmental stress	84
Development Hormones	11
Flowering MADS 8 box, specific precursor protein stamen *fil*1, CONSTANS etc.	25
Miscellaneous Protein homologues, unclassified proteins	26

Most researchers reported that the number of flower specific genes identified amongst the ESTs generated have been low i.e 3-5 genes (Hofte et al. 1993; Lim et al. 1996; Endo et al. 2000). Since a desired outcome of this project is to generate molecular markers that may be used in sex determination to facilitate designing a fruitful seed orchard for the production of this commercial rattan, the eight flower specific genes listed above may be potential candidates for marker design.

In addition the 1529 inflorescence ESTs were compared against 371 fruit ESTs and 338 leaf ESTs. The comparison was carried out for the significant protein match from the 880 inflorescence ESTs, 259 fruit ESTs and 212 leaf ESTs (Table 4). Table 4 provides a comparison of the 1351 significant ESTs of inflorescence, fruit and leaf ESTs that have been clustered into the putative functional group. Homologous ESTs that were identified were then classified into taxonomical groups and gene function. About 57.5% (880 ESTs out off 1529 ESTs) inflorescence ESTs, 69.8% (259 ESTs out off 371 ESTs) fruit ESTs and 68.7% (232 ESTs out off 338 ESTs) leaf ESTs were found to be comparable to other plants genes. The remaining 42.5% (649) of inflorescence ESTs, 30.2% (112) fruit ESTs and 31.3% (106) leaf ESTs were of non-plant genes origin and had matches to genes from other organisms such as bacteria, yeasts, animals or viruses.

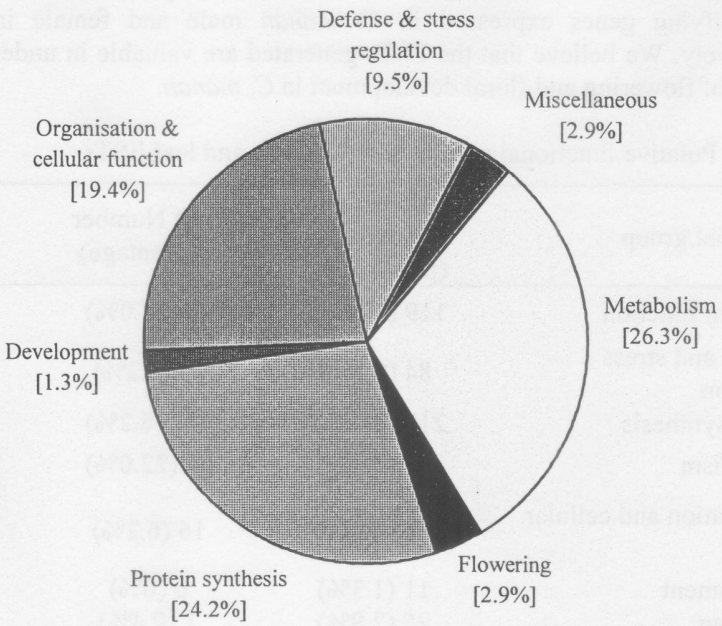

Figure 2. Group 1 EST distribution using functional categories based on a modified MIPS classification system.

Compared with the fruit and leaf ESTs, the highest percentage of *C. manan* floral genes (26.3%) were clustered into the metabolism category same as leaf ESTs (29.7%), while fruit ESTs showing 22%. A total of 24.2% of *C. manan* floral transcripts were involved in protein synthesis, with 16.2% and 14.6% of fruit and leaf ESTs respectively categorized into the same putative functional grouping. This was followed by the organization and cellular function, defence and stress regulation, development, flowering, and miscellaneous, respectively, as illustrated in Table 4. The amounts of unknown function proteins in EST sets of fruit (27.0%) and leaf (24.5%) were larger than floral ESTs (13.5%). As we compared the results among the inflorescence, fruit and leaf ESTs, we may conclude that although almost all genes from all functional protein groups may be present in all tissues, their level of transcription may vary. Expression of these genes may be spatially and temporally variant in the three different tissues (Nadarajah 1999).

However it is not possible for us to make conclusive remarks on the gene pattern and expression profiles in all these tissues based on ESTs that were generated from these libraries. These ESTs only represents small percentage of the libraries. Therefore to obtain a better perspective of gene content and expression levels in these libraries, a larger number of cDNA clones will have to be sequenced and analysed. It is hoped that additional sequencing works will yield candidate

floral specific genes that may be developed into markers for use in the sex determination of rattan. Overall, this EST database has provided useful information in identifying genes expressed in *C. manan* male and female inflorescences respectively. We believe that the ESTs generated are valuable in understanding the process of flowering and floral development in *C. manan*.

Table 4. Putative functional groups of floral, fruit and leaf ESTs.

Functional group	Inflorescence Number (percentage)	Fruit Number (percentage)	Leaf Number (percentage)
Unknown function	119 (13.5%)	70 (27.0%)	52 (24.5%)
Defence and stress regulation	84 (9.5%)	16 (6.2%)	21 (9.9%)
Protein synthesis	213 (24.2%)	42 (16.2%)	31 (14.6%)
Metabolism	231 (26.3%)	57 (22.0%)	63 (29.7%)
Organization and cellular function	171 (19.4%)	16 (6.2%)	24 (11.3%)
Development	11 (1.3%)	0 (0%)	0 (0%)
Flowering	25 (2.9%)	1 (0.4%)	0 (0%)
Miscellaneous	26 (2.9%)	57 (22.0%)	21 (9.9%)
Total Significant EST	880 (100%)	259 (100%)	232 (100%)

ACKNOWLEDGEMENTS

This research was funded by IRPA grant 09-02-02-0036-EA132 from Ministry of Science, Technology & Innovation, Malaysia.

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman D. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

Aminuddin M. 1994. Rattan in Malaysia: Conservation status biodiversity base and the strategic programme. Bamboo and rattan genetic resources and use. *Proceedings of the First INBAR Biodiversity, Genetic Resources and Conservation Working Group.* Pp. 45-48. 7-9 November 1994.

Bausher M, Shatters R, Chaparro J, Dang P, Hunter W, Niedz R. 2003. An expressed sequence tag (EST) set from *Citrus sinensis* L. Osbeck whole seedlings and the implications of further perennial source investigations. *Plant Science* **165**: 415-422.

Bon MC, Basri HAB, Joly HI. 1994. Genetic diversity of two rattan species application of isozyme markers. *Measuring and Monitoring Biodiversity in Tropical and Temperate Forest.* Empress Hotel, Chiang Mai, Thailand. 28 August – 21 September 1994.

Bouck J, Miller W, Gorrell JH, Muzny D, Gibbs RA. 1998. Analysis of the quality and utility of random shotgun sequencing at low redundancies. *Genome Research* **8**: 1074–1084.

Covitz PA, Smith LS, Long SR. 1998. Expressed sequence tags from a roots-hair-enriched *Medicago truncatula* cDNA library. *Plant Physiology* **117**: 1325-1332.

Dransfield J. 1979. *A Manual of the Rattans of the Malay Peninsula.* Kuala Lumpur: Forest Research Institute Malaysia.

Endo M, Kokubun T, Takahata Y, Higashitani A, Tabata S, Watanabe M. 2000. Analysis of expressed sequence tags of flower buds in *Lotus japonicus*. *DNA Research* **7**: 213–216.

Höfte H, Desprez T, Amselem J, Chiapello H, Caboche M, Moisan A, Jourjon MF, Charpenteau JL, Berthomieu P, Guerrier D, Giraudat J, Quigley F, Thomas F, Yu DY, Mache R, Raynal M, Cooke R, Grellet F, Delseny M, Parmentier Y, Marcillac G, Gigot C, Fleck J, Phillips G, Axelos M, Bardet C, Tremousaygue D, Lescure B. 1993. An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana*. *The Plant Journal* **4**: 1051-1061.

Lim CO, Kim HY, Kim MG, Lee SI, Chung WS, Park SH, Hwang I, Cho MJ. 1996. Expressed sequence tags of Chinese cabbage flower bud cDNA. *Plant Physiology* **111**: 577-588.

Matsubara K, Okubo K. 1993. cDNA analyses in the human genome project. *Gene* **135**: 265-274.

Nadarajah K. 1999. Stress Regulation in *Arabidopsis thaliana*. PhD Thesis University of East Anglia.

Theissen G, Becker A, Di Rosa A, Kanno A, Kim JT, Munster T, Winter K, Saedler H. 2000. A short history of MADS-box genes in plants. *Plant Molecular Biology* **42**: 115-149.

Weigel D, Nilsson O. 1995. A development switch sufficient for flower initiation in diverse plants. *Nature* **377**: 495-500.

Wickneswari R, Siti Salwana H, Norwati M, Nur Supardi MN, Aminuddin M. 1995. Genetic diversity in potential seed sources of *Calamus manan* MIQ. in Peninsular Malaysia. *Malaysian Applied Biology* **31(1)**: 49-58.