Plant Tissue Cult. & Biotech. **16**(2): 95-104, 2006 (December)

PTC&B

Expressed Sequence Tags (ESTs) from a Jute (*Corchorus olitorius*) cDNA Library

J. Matthew Taliaferro, Ahmad S. Islam and Kanagasabapathi Sathasivan*

Molecular, Cell and Developmental Biology, School of Biological Sciences,
The University of Texas at Austin, USA-78712

Abstract

Analysis of the genome of jute (*Corchorus olitorius*) was done by creating a new cDNA library of expressed sequence tags (ESTs) in pBluescript as the previous libraries reported earlier in this journal yielded only small DNA fragments from chloroplast and mitochondrial DNA. This report discusses results from a cDNA library constructed using poly A+ mRNA purified from 7-day old etiolated jute seedlings. Out of 700 recombinant plasmids obtained, 250 were analyzed using WU-BLAST (www.arabidopsis.org) for similar EST sequences in *Arabidopsis thaliana* and other higher plants. So far the analysis of the library has yielded several significant sequences, including the complete open reading frame of the 60S acidic ribosomal protein P3 and a partial cDNA of Class I chitinase. These results and future EST sequences from this library will be made available in Genbank and the sequence information will be used to clone full length DNA through PCR.

Introduction

Jute, *Corchorus olitorius* L. and *C. capsularis* L. grown mainly in South Asia, is the second largest fiber crop in terms of cultivation next to cotton. In 2005, 2.9 million tons of jute plants were harvested. Although it was originally used as a natural fiber crop, it has now grown in importance as a possible candidate for paper production, a component in textile industry and a potential source for cellulosic ethanol production. The progress made in these areas has been hindered by the high content of lignin in its fibers. The long term goal of this project is to better understand the jute genome and to produce transgenic jute varieties that will have higher cellulose and lower lignin contents than that of current cultivars without compromising their other important agronomic traits such as disease resistance, photoperiod insensitivity, strong and lustrous fibers, etc. In addition, we aim to establish a database on jute genome analysis, repository of ESTs, complete cDNA and genomic DNA fragments that will be freely available to jute researchers all over the world.

*Corresponding author: sata@mail.utexas.edu

Last year the progress of this project was published (Islam *et al.* 2005) reporting isolation of several gene fragments from the cDNA and genomic DNA library constructed for us by a commercial company in the USA. Of these fragments, base sequences of two from *C. capsularis* var. CVL-1 and one from *C. olitorius* var. O4 were deposited in the GenBank. EST's from *C. capsularis* were RNA polymerase C1 (rpoC1) gene and putative phosphate transport ATP-Binding protein (Ingram et al. 2006), and that from *C. olitorius* was chloroplast 18S ribosomal RNA gene (Stone et al. 2005). The other EST's reported in that paper was a complete sequence for tRNA-Leu and partial sequence of DNA fragments encoding several proteins, such as RNA polymerase ß-subunit-1, mitochondrial DNA directed RNA polymerase and carboxytransferase ß-subunit. The ultimate object of this project is to understand more about the genome of jute, with particular emphasis on elucidating the pathway for biosynthesis of lignin in which three genes, namely, caffeoyl-CoA-O-methyltransferase, cinnamyl alcohol dehydrogenase and 4-coumaryl CoA-ligase are involved. Already complete cDNA sequences of the first two genes from *C. capsularis* have been deposited in GenBank by Basu et al. (2004). There is no such report of isolation of these genes in *C. olitorius,* which produces superior fibers (International Jute Study Group (IJSG), Tejgaon, Dhaka).

Significant progress has been made in the last few years to understand the genome sequences of *Corchorus* species. As of December 2006, there are over 427 nucleotide sequences from jute plant available in the Genbank, including 180 cDNA sequences. Many of the cDNA sequences are incomplete establishing a good start of expressed sequence tag (EST) collection. ESTs are important to understand the genes that are expressed in jute plant. Most of the published ESTs are from *C. capsularis* sequenced by S. K. Sen's lab (sequences submitted by Basu et al. (2004) and Sadhukan et al. (2006). We are reporting the initial sets of ESTs from *C. olitoius*. To understand the commonly expressed genes and to identify genes of agronomic interest, we decided to construct initially a cDNA library of *C. olitorius* var. O4 using the pBluescript II XR cDNA Library Construction Kit (http://www.stratagene.com/). The present paper describes the procedure followed to construct cDNA library and some of the EST sequences obtained from the library. One of the interesting EST reported in this paper is about class I chitinase. It is our long-term goal to identify as many genes as possible and to clone genes involved in lignin biosynthesis and in developing disease resistance.

Materials and Methods

Total RNA isolation: Extreme care was taken to isolate poly A+ RNA because it makes up only a small fraction: i.e., 1 - 4% of the total RNA content, 90% of the remainder being 28S and 18S ribosomal RNA, 2 - 4% transfer RNAs, and the final 1 - 2% other small RNAs, including microRNAs. All RNA isolated over a period of two months was obtained from 7 - 10-day old etiolated seedlings of the "Tossa" variety of jute (*C. olitorius* var. O4). Our initial efforts at total RNA isolation using TRI Reagent (http://www.ambion.com) from Ambion produced mixed results, possibly because of the high polysaccharide content of these seedlings. This difficulty prompted us to use an Invitrogen-made extraction reagent specifically made for plant RNA extraction called "Concert Plant RNA Isolation Reagent" (https://www.invitrogen.com). Its quality was then checked using a denaturing formaldehyde-agarose gel using the Northernmax solutions containing formaldehyde in both the gel buffer and loading buffer (http://www.ambion.com) to get rid of any secondary RNA structure.

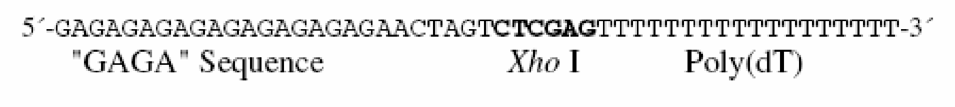
mRNA purification : mRNA was then isolated from the total RNA using affinity chromatography (http://www.ambion.com/catalog/CatNum.php?AM1922).

We preferred to use a system, where the oligo-dT sequences were attached to metal beads. These extraction materials were part of the Poly(A)Purist kit from Ambion. Using the beads, the mRNA was first bound to the oligo-dT sequences contained on the magnetic beads, thereafter these beads were pulled to the side of the container using a magnet. The unbound RNA still in solution was then pipetted out of the tube. mRNA was then eluted off of the beads and precipitated for subsequent use.

Due to the rapid degradation of RNA, this procedure was done in a sterile environment to get rid of any contaminating RNases. Approximately 700 ng of mRNA was isolated for cDNA library construction.

Overview of cDNA library construction: The library was constructed using the pBluescript II XR cDNA Library Construction Kit form Stratagene. It uses pBluescript as a vector and also came with a sample of high quality test mRNA meant for use as a positive control throughout construction.

First strand synthesis: The synthesis of the first strand complementary to the extracted mRNA again takes advantage of the poly-A tail on the 3′ end of every transcript. The reverse transcriptase (Stratascript RT) needs a primer to start, so the following primer was added:



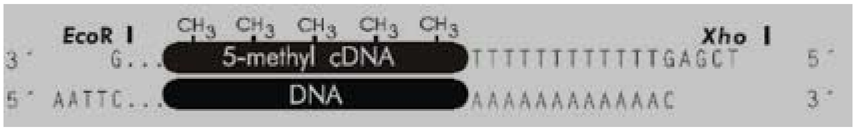
5´-GAGAGAGAGAGAGAGAGAGAACTAGT**CTCGAG**TTTTTTTTTTTTTTTTTT-3´

"GAGA" Sequence *Xho* I Poly(dT)

The "GAGA" sequence was designed to protect the *Xho* I restriction site and this *Xho I* site allows the finished cDNA to be inserted into the pBluescript vector in a sense orientation (5' -*Eco*R I–>*Xho* I -3'). The poly(dT) region anneals to the 3´ poly(A) tail of the mRNA template to synthesize the first-strand cDNA.

The Xho I restriction site became important later in the orientation and insertion of the insert into the plasmid. The dNTP mixture added in this reverse transcription reaction contained methylated dCTP. This ensured at the resulting cDNA was hemimethylated, protecting it from degradation by *E. coli* after transformation and by restriction enzymes in subsequent library construction steps. Due to safety concerns, no radioactive materials were used in the creation of this library.

Second strand synthesis: The RNA was removed from the reaction mixture containing a RNA/DNA hybrid to allow synthesis of a second strand complementary to the newly produced strand. RNase H was added to nick the RNA at several places leaving many RNA fragments. These fragments served as primers for DNA polymerase I, which synthesized the second strand using the process known as nick translation. Uneven termini were then filled in with *Pfu* DNA polymerase to produce blunt ends, to which EcoRI adapters were ligated. This step allowed each insert to have an EcoRI adapter on both ends. It is to be noted here that the Xho I restriction site from the original primer used in first strand synthesis remained on one end. Inserts were then digested with Xho I to release the EcoRI adapter from one end of the cDNA; it produced a directional cDNA insert as shown below:



The cDNA inserts were then precipitated using ethanol and prepared for ligation into the vector, pBluescript.

Ligation: The inserts were ligated into pBluescript by digesting the plasmid with EcoRI and Xho I, then adding T4 ligase and rATP. Size-fractionation was omitted because the amount of mRNA retrieved from our material was low, and we feared that this step would result in the loss of many inserts.

Transformation: The recombinant plasmids were then transformed into XL-10 gold competent cells using chemical transformation. The competent cells have holes in the membrane through which even large molecules such as plasmids can diffuse into. After the plasmids were allowed to diffuse in, the holes in the membrane were repaired by heat shocking the bacteria at 42ºC for 30 seconds. This initiates various repair mechanisms in the *E. coli* cell, including membrane repair. The cells were then allowed to grow for 1 h at 37 ºC in NZY+ broth. They were then plated on agar plates containing ampicillin, X-gal, and IPTG. The recombinant plasmids were selected based on blue-white color selection. Part of the library was then plated while the rest was amplified and saved for later analysis. The part that was plated produced approximately 700 bacterial colonies that contained recombinant plasmids. These were stored at –70ºC in glycerol cultures for later analysis. White colonies were plated again and cultured in liquid media for plasmid DNA isolation and restriction digestion with PvuII enzyme.

DNA isolation and sequencing: The plasmid DNA was isolated using the plasmid DNA miniprep kit from Invitrogen and then sent for automated sequencing in the Institute for Cell and Molecular Biology at UT Austin.

Results and Discussion

We obtained approximately 80-120 µg of total RNA per g of plant tissue. This process was repeated until approximately 1 mg of RNA was isolated. Fig. 1 shows RNA bands in the 2 - 3 kb range including that of small tRNAs (~80-100 bp) towards the bottom of the gel. The sharpness of the more prevalent 28S and 18S rRNA bands revealed that both the mRNA, and rRNA were of good quality. The same figure shows mRNA as a faint smear dispersed in the background between these bands accounting for varying lengths of mRNA transcripts. The sharp and clear profile of rRNAs in the gel indicated that those and the rest of the RNA including the mRNA were not degraded.

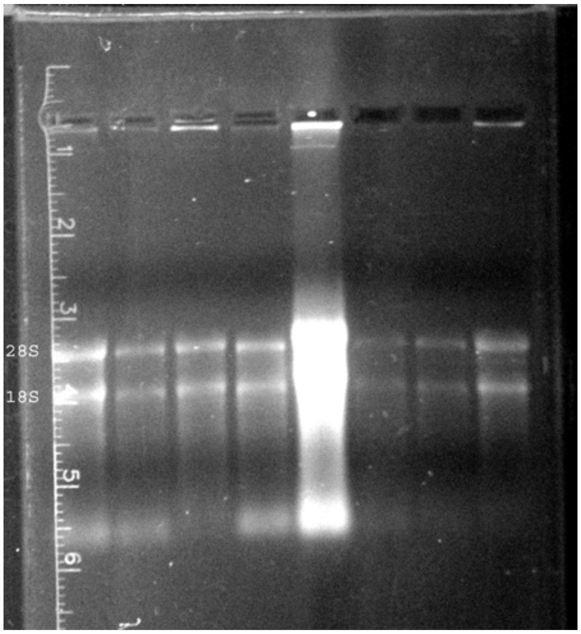


Fig 1. A denaturing RNA gel profile containing total RNA isolated from jute, *C. olitorius* var. O4. 28S and 18S bands are visible as a pair of bands approximately one-third of the way through the gel. Bands containing tRNA are visible at the bottom, while the mRNA comprises a faint smear seen behind all of the bands. The brightness seen in lane 5 is due to overloading.

Earlier results show that approximately 1 in 8 contain significant size (150 bp or larger) inserts. After isolation of the plasmid DNA, it was digested with Pvu II. Pvu II cuts outside of the multicloning site, producing a plasmid fragment of 2.6 kb together with two others, one that contains the insert and another approximately 400 bp from the plasmid (Fig. 2).

The cDNA library constructed in pBluescript cloning vector (www.stratagene.com) into the EcoRI/XhoI restriction sites and stably maintained in the *E. coli* host cell XL-10-Gold contained a sufficient number of primary clones comprising about 65 % recombinants. The average insert size was 150 to 500 bp. The short strands of cDNA could possibly be due to some inhibitory compound that might have co-purified with the jute RNA during the process of isolation. Special attention to prevent such contamination may help in future cDNA synthesis and library attempts.

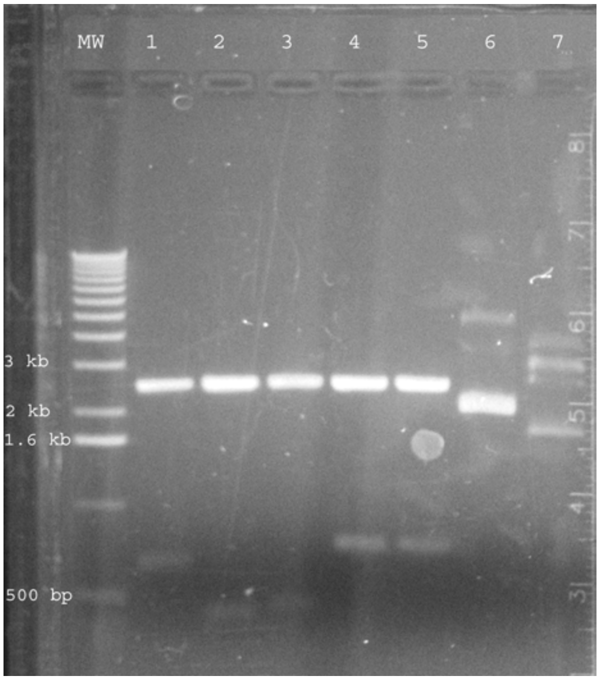


Fig. 2. Gel profile containing isolated plasmid DNA after digestion with Pvu II. Lanes 1-7 correspond to clones 194-200, respectively. The straight line of bands at 2.6 kb represents the uniform piece of plasmid that is cut out after digestion. Smaller fragments below are the cDNA inserts plus approximately 400 bp of the vector. Insert size is thus band size minus 400 bp.

The cDNA library was constructed in pBluescript cloning vector (www.stratagene.com) using the EcoRI/XhoI restriction sites. The recombinants were stably maintained in the *E. coli* host cell XL-10-Gold. The library contained a sufficient number of primary clones comprising about 65 % recombinants. The average insert size was 150 to 500 bp. The short strands of cDNA could possibly be due to some inhibitory compound that might have co-purified with the jute RNA during the process of isolation. Special attention to prevent such contamination may help in future cDNA synthesis and library attempts. The same problems were also encountered in previous attempts at the creation of a jute cDNA library (Islam et al. 2005).

Randomly selected white colonies were cultured on LB ampicillin (50 μg/ml) overnight; the plasmid DNA was isolated, checked for recombinants and the DNA was sent for sequencing using T3 primer. The cDNA sequences were identified based on the poly(A) tail region after removal of the vector sequences. Such well-defined cDNA sequences were used to search for similar sequences through WU-BLAST with introns and untranslated regions excluded (AGI Genes (-introns, +UTRs) (Washington University – BLAST 2.0) listed in TAIR (www.arabidopsis.org) web site. The results of sequence analysis of selected clones are summarized in Table 1.

Table 1. This table indicates a sample of the EST's that have been isolated thus far that show significant homology to known sequences in other plants. The sample corresponds to the clone number from the above-mentioned library. The length indicates the size of the stretch over which the homology was present, while the score and P values are statistical indicators of the strength of the match.

Sample	Length (bp)	Homology	Score	P value
004	438	60S acidic ribosomal protein P3	819	2.90E-32
165	263	Chitinase-like protein (Class I)	288	1.80E-07
198	238	Expressed protein	169	0.98
225	151	Expressed protein on chromosome 1	159	0.086
124	77	Subunit of Photosystem II: PSBO2,	157	0.16
185	102	Plastid gene expression	148	0.41
118	88	Cytochrome P450 family protein	146	0.45
207	115	Myb family transcription factor	143	0.52
194	160	Calmodulin-family binding protein	141	0.67
226	142	Auxin responsive protein/IAA-induced protein	135	0.81
148	44	Chloroplast outer membrane protein	129	0.98
229	81	Cytochrome P450 71B7	124	0.998

The information in the above table was compiled from all sequences collected from August to December, 2006. Any entered sequences that gave homologous sequences with a maximum score of 50 bits or less were discarded. The scores and error values for the homologous sequences were recorded. The highest score for each input sequence was noted. The probable location of the genes of interest, their placement in the chromosome, and the specific protein encoded by each individual gene were also taken note of. The sequences reported above contain only the 3′ end of the open reading frame and the 3′ UTR downstream from the open reading frame. This was evidenced by the presence of poly (A) region in the cDNA sequences, obtained using the T3 primer located upstream of the sequence, that gave homology in WU-BLAST. This also explains some of the low scores obtained from the WU-BLAST analysis. The reason for the relatively low scores is two-fold. First, all of these EST's were relatively short (75-500 bp). This limited the number of nucleotides available for a homologous match. Second, these EST's contain the 3′ UTR. Since the UTR is not translated, it is not under the evolutionary stress to remain conserved from generation to generation, leading to variability from organism to organism and reducing the score and the number of homologous matches found.

The two most significant matches are those of the 60S acidic ribosomal protein P3 and the Class I chitinase-like protein. A multiple alignment analyzing the 60S acidic ribosomal protein P3 sequence and similar P3-encoding cDNA sequences from *Arabidopsis thaliana* and *Lactuca saligna* was done using ClustalW (http://www.ebi.ac.uk/clustalw/). The *Arabidopsis thaliana* and *Lactuca saligna* sequences were found using NCBI BLAST (http://www.ncbi.nlm.nih.gov/). This showed significant homologies between these three sequences over the length of the entire transcript, including strong homology around the ATG that was found to be the start of the open reading frame for translation of the protein. Further analysis of this sequence using BLASTX, a program that translates the nucleotide sequence into an amino acid sequence and then compares that sequence against all of the proteins in its database (http://www.ncbi.nlm.nih.gov/BLAST/) showed that the translated sequence also had significant homology to the 60S acidic ribosomal protein P3 in *Arabidopsis* with a score of 104 and an E value of 2 X 10^{-21}. The homology started with a methionine residue, the same amino acid shown in the nucleotide sequence that signifies the beginning of the open reading frame, and continued encoding for 67 amino acids as shown below (Fig. 3).

```
Jute         52   MGVYTFVCKSSAGEWTAKQFEGELEGSAPSTYELQRKLVQAAAAADSSGGVCSSFSLITPTSAVFQV   252
                  MGV+TFVCK+  G W+AKQ EGELE SA STYELQRKLVQ + +ADSSGGV SSFSL++PTSAVFQV
Arabidopsis  1    MGVFTFVCKNGGGAWSAKQHEGELESSASSTYELQRKLVQVSLSADSSGGVQSSFSLVSPTSAVFQV    67
```

Fig. 3. Homology between the translated EST isolated from *C. olitorius* and the 60S acidic ribosomal protein P3 from *A. thaliana* obtained using BLASTX.

Similar analysis of the Class I chitinase-like sequence showed less but still significant homology to the Class-I chitinase sequences in *Arabidopsis*. Unlike the 60S ribosomal protein sequence, it did not contain the entire open reading frame; rather it contained only approximately the last 30 amino acids and the 3′ UTR. Upon investigation of the sequence using BLASTX, the translated nucleotide sequence of this EST showed significant homology to the Class I chitinase in

Arabidopsis with a score of 49.3 and an E value of 6 X 10^{-5}. The homology, as shown in Fig. 4 was seen towards the carboxy-terminus of the protein. This is consistent with earlier findings that the EST contained the poly-A tail, 3′ UTR, and 3′ end of the open reading frame.

```
Jute           11   LDLLGIGREQAGPHDELTCAEQVAFNPT      94
                    LDL+GIGRE AGP+DEL+CAEQ  FNP+
Arabidopsis   286   LDLMGIGREDAGPNDELSCAEQKPFNPS     313
```

Fig. 4. Homology between the translated EST isolated from *C. olitorius* and the Class I chitinase from *A. thaliana* as revealed in BLASTX.

All other EST's were too short to perform the same analysis with. However, work is under way to determine the sequence of the full length of the above transcripts. The technique that has been chosen for this purpose is 5′ Rapid Amplification of 5′ Ends provided by Invitrogen. (http://www.invitrogen.com/ content/sfs/manuals/ generacercdna_man.pdf). This will involve RNA ligase-mediated amplification of 5′ ends using a gene specific primer and a 5′ adaptor primer. This will be selectively used to clone genes of agronomic interest. In addition, we are planning to continue characterizing additional ESTs that could be useful in cloning the full-length genes and enable us to understand gene expression patterns in various tissues under different stress conditions. Accumulation of EST sequences of jute, now being worked upon by various labs in the world such as those under the guidance of Professor S. K. Sen at the Indian Institute of Technology, Kharagpur; Professor P. K. Gupta at Charan Singh University, Meerat, India; Professor Haseena Khan at Dhaka University and Dr. Anirban Roy at Plant Virus and Biotechnology Unit, Central Research Institute for Jute and Allied Fibers, Barrackpore, India will pave the way toward cloning new genes resulting in marker assisted breeding and genetic engineering to improve jute crop.

Acknowledgement

The authors sincerely acknowledge the technical help extended by two undergraduate students, Mohamad Wazni and Nabila Anwar, in isolating plasmid DNA from a number of cDNA library samples (stored at –70ºC in glycerol), some of which were sent for sequencing. The jute seeds were originally obtained from the Bangladesh Jute Research Institute through Professor Haseena Khan at the Department of Biochemistry and Molecular Biology, Dhaka University. Part of the funding for this project came from the Undergraduate Research Fellowship by University CoOp and The University of Texas at Austin.

References

Basu A, Ghosh M, Meyer R, Powell W, Basak SL and **Sen SK** (2004) Analysis of Genetic Diveristy in Cultivated Jute Determined by Means of SSR Markers and AFLP Profiling. Crop Sci. **44**: 678-685.

Basu A, Maiti MK and **Sen SK** (2004) Cloning and sequencing of mRNA for caffeoyl-CoA-o-methyltransferase from *Corchorus capsularis* (AY500159) complete CDS (908 bp) PLN 17-JAN-2004 from NCBI Genbank (Unpublished).

Basu A, Maiti MK and **Sen SK** (2004) Nucleotide sequence of cinnamyl alcohol cDNA isolated from *Corchorus capsularis* (AY504425) complete CD (1243 bp) PLN 13-JAN-2004 from NCBI Genbank, Unpublished.

Ingram CI, Montalvo RJ, Stone LA, Islam A and **Sathasivan K** (2006) RNA polymerase C1-like (rpoC1) gene from chloroplast partial sequence from *Corchorus capsularis* (DQ198155) PLN 12-APR-2006 from NCBI Genbank (Unpublished).

Islam AS, Taliaferro JM, Lee CT, Ingram CI, Montalvo RJ, Van der Ende G, Alam S, Siddiqui J and **Sathasivan K** (2005) Preliminary Progress in Jute (*Corchorus* species) Genome Analysis. Plant Tissue Cult. and Biotech. (15)2: 145-156.

Khan F, Islam AS and **Sathasivan K** (2004) A rapid method for high quality RNA isolation from Jute: *Corchorus capsularis* and *C. olitorius* L. Plant Tissue Cult. and Biotech. **14**: 63-68.

Montalvo R J, Ingram CM, Stone LA, Islam A and **Sathasivan K** (2005) 18S ribosomal RNA gene from chloroplast, partial sequence from *Corchorus olitorius,* partial sequence (DQ151662) PLN 24-AUG-2005 from NCBI Genbank (Unpublished).

Sadhukhan S, Basu A, Maity MK and **Sen SK** (2006) Studies on development of jute fibre through generation of ESTs from NCBI Genbank (Unpublished).

Stone LA, Ingram CM, Montalvo RJ, Islam A and **Sathasivan K** (2005) Putative Phosphate Transport ATP-Binding Protein gene from *Corchorus capsularis, partial sequence* (DQ151661) PLN 24-AUG-2005 from NCBI Genbank (Unpublished).

Company protocols used in the present investigation:

Concert Plant RNA Isolation Reagent: https://www.invitrogen.com/content/sfs/manuals/ plantrna_man.pdf

TRI Reagent: http://www.ambion.com/catalog/CatNum.php?AM9738

Northernmax gel products: http://www.ambion.com/catalog/CatNum.php?AM8676

Poly(A) Purist Purification: http://www.ambion.com/catalog/CatNum.php?AM1922

pBluescript II XR cDNA Library Construction Kit (including library construction figures): http://www.stratagene.com/manuals/200455.pdf

RACE protocol http://www.invitrogen.com/content/sfs/manuals/generacercdna_man.pdf