Transcriptome Analysis in Seabuckthorn (*Hippophae rhamnoides* L.), a Medicinally Important Plant

Prakash Chand Sharma*, Ankit Jain and Saurabh Chaudhary

**Abstract**— More than 50% crop yield is lost due to harsh, changing climatic conditions such as extreme temperatures, drought, etc. Worldwide researchers are making serious efforts towards developing high yielding crops with capability to grow under these abiotic stress conditions. Therefore, it becomes necessary to study plants tolerant to adverse environmental conditions. In this study an attempt has been made to identify abiotic stress responsive genes especially for cold tolerance through transcriptome analysis in seabuckthorn, a plant with immense medicinal value and ability to grow at extreme temperatures. The work includes a pilot study comprising of EST sequencing by Sanger’s capillary method, followed by whole transcriptome analysis by next generation sequencing. In the preliminary study 3412 ESTs have been generated from cDNA library of seabuckthorn leaf tissue, which could be clustered to 1665 unigenes. Subsequently, 88,297 putative unigenes were assembled exploiting next generation sequencing of leaf and root tissue libraries. Few of these cold tolerant genes have already been identified and validated by RT-PCR in the earlier study. DeepSAGE has also been performed for differential gene expression analysis to identify the genes for cold tolerance. Approximately 20 thousand genes were found to be differentially expressed during cold stress. Microsatellite distribution in transcriptome was also studied in *silico* and polymorphic markers are being evaluated to access genetic diversity among seabuckthorn population. This vast amount of information generated is under analysis to answer key questions pertaining to abiotic stress management in seabuckthorn and other plant species.

**Keywords**— Seabuckthorn, Transcriptome, Abiotic stress, Gene expression, Microsatellites

I. INTRODUCTION

IGH altitudinal areas are long known to harbor vegetation of great medicinal and ecological values. One important member growing among such vegetation widely in Indian Himalayas is ‘Seabuckthorn’. Sea-buckthorn is a diploid (*2n = 24*) [1], deciduous spiny shrub belonging to family elaeagnaceae. Its medicinal significance has already been recognized long ago in ancient Chinese and Tibetan medicine. Its berries are potential source of many bioactive compounds ranging from flavanoids, tocopherol to fatty acids, etc. [2]. Moreover, the plant is a rich source of vitamin C [3] and other antioxidants [4]. The extract from different plant parts works as immune modulator [5], anti carcinogenic [3] and anti-bacterial [6]. Apart from medicinal value, plant also works as ecosystem restorer as seabuckthorn shrubs are effective in preventing soil erosion due to their extensive root system [7] and symbiotic association with nitrogen-fixing actinomycetes *Frankia*, which improves the soil fertility [8].

Seabuckthorn is found to be growing in diverse areas of harsh environmental conditions ranging from seacoasts, riverbeds to cold desert. Thus, the plant species is resistant to drought and tolerates soil salinity and extreme temperatures, prevailing in foresaid areas [9]. Therefore seabuckthorn genome is expected to harbor genes, responsive to various abiotic stresses. These extreme environmental stresses further forces plant to develop alternate survival mechanism and to adapt by manifesting various physiological and morphological changes. To cope up such situations plant also uses the plasticity of gene expression and that can change dramatically over very short periods of time to withstand local environmental conditions. Studying the biochemical and molecular mechanisms by which plants tolerate abiotic stresses is therefore necessary for improving crop performance under stress. Analysis of seabuckthorn transcriptome is expected to provide necessary information about the genes involved in abiotic stress management.

Generation of expressed sequence tags (ESTs) and serial analysis of gene expression (SAGE) are most common approaches used widely for complete transcriptome profiling and differential gene expression analysis. ESTs are single pass, partial sequence reads generated from either 5’ or 3’ end of a cDNA clone [10]. As compared to whole genome sequencing which is costlier and time consuming process, EST sequencing presents a low cost alternative and a valuable starting point to characterize structural and functional genomic data in non-model or neglected but otherwise useful organism. EST database serves as a valuable resource that can be exploited for gene discovery, genome annotation, comparative genome analysis and generation of molecular markers such as SSRs and SNPs [11, 12]. In particular, ESTs from a full length enriched cDNA library provide a complete scenario of functional proteins present in respective targets [13].

Revolution in sequencing technology i.e shift from Sanger’s sequencing method to high throughput next generation
sequencing technology has further offered the opportunity to obtain vast sequence information in a time and cost effective manner. Next generation sequencing technologies hold a great potential in many areas of advanced research, including large scale re-sequencing of sequenced genome, transcriptome and micro RNA sequencing, and DNA methylation studies. The efficiency and speed of gene discovery has improved dramatically due to the emergence of next generation sequencing technologies [14, 15]. The various platforms developed for next generation sequencing have potential to generate millions of tags and elucidate the complete transcriptome profiling of any organism. However, due to generation of millions of short read tags during NGS data generation as compared to Sanger’s capillary sequencing method, the de novo assembly and annotation of short reads into well representing form is the major challenging task.

SAGE is another promising technique used for the quantification of the transcripts of an organism in a given environment. The major advancement in the technique is DeepSAGE a combination of LongSAGE and the high throughput next generation sequencing. The DeepSAGE technique provides million’s of tags, including those representing rare transcripts with higher frequency than the tags obtained in original SAGE and its variant as well as microarray based expression analysis [16]. Furthermore, it is cost effective and less laborious as compare to other gene expression analysis techniques.

Although the ESTs generation by Sanger’s capillary sequencing, tag based high throughput next generation sequencing and DeepSAGE technology are well established techniques for transcriptome studies of any organism including non-model plants, each technique has its own significance and shortfalls. By keeping all this in mind, we have initiated experiment to characterize the transcriptome of seabuckthorn by generating ESTs from leaf tissues by traditional Sanger sequencing method as a pilot study followed by complete transcriptome profiling by next generation sequencing from leaves and roots tissues of seabuckthorn. Finally, we have employed DeepSAGE to identify genes differentially expressed during abiotic stress in seabuckthorn.

II. MATERIALS AND METHODS

A. Sample preparation and total RNA isolation.

Leaf tissues of *H. rhamnoides* for the pilot study were harvested from the premises of DIHAR, Leh, India and the samples were then transported in liquid nitrogen and stored -80°C until further use. However, for NGS and gene expression studies, *in vitro* germinated *H. rhamnoides* seedlings were taken. The total RNA was isolated from tissues using the modified CTAB RNA isolation protocol [17].

B. cDNA library construction and generation of expressed sequence tags

cDNA library was constructed using the SMART cDNA library construction kit (Clontec, Japan) from high quality RNA. Percentage insert of the library was estimated by digesting few transformed clones with restriction enzymes *Sma*I and *Xba*I (Fermentas). Transformed colonies were stored as glycerol stocks in -80°C for further use. Transformed colonies were inoculated in LB broth (HiMedia) and incubated overnight at 37°C with continuous shaking. Plasmid isolation was performed in 96 well format through Perfectprep Plasmid Isolation Kit (Eppendorf). Quality and quantity of plasmid DNA was evaluated through agarose gel electrophoresis against lambda *HindIII* DNA marker. Inserts were sequenced by MegaBACE sequencer (based on Sanger’s capillary method), using Dye terminator chemistry (GE Biosciences, USA). *Phred* software [18, 19] was used for base-calling of chromatogram obtained as a result of sequencing to generate sequence reads, which were then subjected to *crossmatch* software to mask vector sequences and trims off low quality sequences. After quality trimming the sequences were submitted to dbEST in NCBI.

C. Clustering of expressed sequence tags and functional annotation of unigenes

CAP3 software [20] was used for clustering of EST’s into Unigenes. These unigenes (contigs and singletons) were subjected to BLASTX similarity search. Function was assigned to sequences depending upon the best hit against non redundant protein database (of NCBI). For Gene Ontology annotation, BLAST2GO [21] software was used to assigns GO term to the BLASTX positive hits. Gene investigator [22] was used for selection of abiotic stress responsive genes.

D. In silico microsatellite screening

Sequence datasets were screened for microsatellites using a perl script MISA [23] with default parameters. Mononucleotides repeats were not included in the search parameters considering presence of polyA/ poly T repeats in the sequences due to polyA/ T tails in eukaryotic transcripts.

E. Next Generation Sequencing

High quality RNA isolated from leaf and root tissues of *in vitro* germinated seabuckthorn seedlings was analyzed for RIN (RNA Integrity Number) value using Bioanalyzer. RNA with RIN value 8 or more was selected for Next Generation Sequencing. Two separate libraries from root and leaf tissues were constructed and sequenced commercially on Illumina Solexa GA II platforms by Ocimum Biosolution Pvt Ltd., Hyderabad, India.

F. Gene expression analysis of selected cold stress related genes

On the basis of differential expression pattern and from our pilot study, few cold related genes were selected for validation by real time-qPCR. First strand cDNA template was constructed using the above RNA sample as per manufacturer’s protocol. Real time RT-PCR experiments were performed on One Step Plus Real Time PCR (Applied Biosystem). Real time cycling conditions were performed (10 min 95°C, 40 cycles of 15 s 95°C and 60 s at 60°C) followed by generation of melting curve to check the specificity of amplification.
III. RESULTS AND DISCUSSIONS

A. RNA isolation and cDNA library construction

Availability of high quality RNA is the prerequisite for cDNA library construction as high quality and quantity of RNA ensures the complete representation of all transcripts. However, isolation of high quality and quantity of RNA is a difficult task in case of plants like seabuckthorn, due to the presence of high concentration of polysaccharides, polyphenolic compounds and other secondary metabolites [24] as these compounds cause hindrances in good quality RNA extraction. Traditional RNA isolation protocols [25, 26] failed to yield good quality RNA. A modified CTAB method developed in our laboratory [17] using enhanced concentration of β- mercaptoethanol and introduction of polyvinylpolypyrolidone pylonride, in CTAB RNA extraction buffer gave high quality ($A_{260}/A_{280}$ ratio= 1.7) and better RNA yield that was optimum to proceed for cDNA library construction. cDNAs generated as a result of reverse transcription of mRNA with a size range of 0.5-4.0 kb were selected for cDNA library construction. Insert percentage of library was found to be 90%.

B. EST sequencing, submission, clustering and annotation

In our pilot study, 4700 clones were selected from cDNA library and sequenced using Sanger’s method leading to the generation of 3412 high quality EST sequences, which were later submitted to dbEST, NCBI. Clustering of EST’s using CAP3 resulted into a total of 1665 unigenes comprising of 345 contigs and 1320 singletons [27]. On the basis of Gene Ontology (GO) annotation could be assigned to 1665 unigenes. Among various molecular functions unigenes involved in structural constituent of ribosome, DNA binding and zinc ion binding were highly represented.

C. Next generation sequencing

Promising results obtained from the generation of EST database of seabuckthorn provided foundation for the current larger study i.e transcriptome analysis and global gene expression profiling. Next generation sequencing of leaf and root tissue libraries yielded more than 90 million short reads comprising of approximately 8.5 billions nucleotides, which could be further assembled into 88,297 putative unigenes. The immense data generated through this project is under analysis using suitable bioinformatics tools.

D. Microsatellite distribution and marker development

In the pilot study 62 SSRs were identified in 59 unigenes. More than 50% of the SSRs were dinucleotide repeats followed by trinucleotide repeats [27]. Primers could be designed for 30 of the microsatellites identified in this study, and further assessed on genomic DNA of seabuckthorn accessions collected from DIHAR, Leh. Twenty four primers resulted into positive amplification, while nine primer pairs in H. rhamnoides genotypes and two additional primer pairs in H. salicifolia genotypes, were found polymorphic [28]. When similar study was conducted on data obtained as a result of NGS, more than seven percent putative unigenes were found to be carrying microsatellites repeat within them. Markers developed as a result these studies are currently being assessed on germplasm collected from diverse geographical conditions of Leh and Lahual.

E. Gene expression analysis of selected cold stress related genes

On the basis of GO annotation and Gene investigator software [22], in our pilot study 16 cold stress related genes were selected for the RT-PCR validation. In quantitative analysis of these selected genes, house-keeping gene ‘actin’ was used as internal control and the relative expression level of cold stress responsive genes were estimated using $2^{-\Delta\Delta CT}$ method [29]. Another parallel study of deep sequencing of tags generated from control and stressed leaf tissue RNA, revealed an average of 20,000 genes, being expresses differentially. Of which 21 highly expressed genes showing differential expression in stressed samples as compare to control, have been selected for RT-PCR and the validation is under process.

IV. CONCLUSION

Availability of advanced NGS methods has resulted into vast amount of data at low cost leading to adoption of these approaches widely for transcriptome analysis. EST’s generated as a result of our previous study provided a reference dataset and has proved to be very useful resource for further functional genomic analysis in seabuckthorn. Vast information that has been exploited using NGS is certain to reveal useful information from transcriptome of seabuckthorn, leading to the identification of important abiotic stress responsive gene as well as enriching the limited marker resources.

ACKNOWLEDGMENT

We thank DIHAR (Defense Institute of High Altitude Research), Leh, India for providing seabuckthorn samples. This work was supported by grants from DRDO (Defense Research and Development Organization), India and DBT (Department of Biotechnology), India.

REFERENCES


