# *De novo* sequencing and assembly of *Azadirachta indica* fruit transcriptome

**Neeraja M. Krishnan[1,+], Swetansu Pattnaik[1,2,+], S. A. Deepak[1], Arun K. Hariharan[1], Prakhar Gaur[1], Rakshit Chaudhary[1], Prachi Jain[1], Srividya Vaidyanathan[1], P. G. Bharath Krishna[1] and Binay Panda[1,2,]***

[1]Ganit Labs, Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology, Biotech Park, Electronic City Phase I, Bangalore 560 100, India
[2]Strand Life Sciences, Kirloskar Business Park, Bellary Road, Hebbal, Bangalore 560 024, India

***Azadirachta indica* (neem) is a unique, versatile and important tree species. Many parts of the plant are traditionally used as pesticide, insecticide, fungicide and for other medicinal purposes. *Azadirachta* fruits and seeds, a good source of oil, are widely used for agriculturally important pest management. Neem oil and its derivatives also support multiple cottage industries in India. Past efforts have been mostly concentrated towards identifying, characterizing and synthesizing one of its principal components, i.e. azadirachtin from seed kernels. Despite diverse use of the neem plant, a modern drug-development programme which systematically exploits the therapeutic ability of *Azadirachta* fruits remains to be fully established. Next generation sequencing technology that helps decode genomes and transcriptomes has transformational impact on medicine, agriculture, bio-fuel and biodiversity studies. Here, we report sequencing, assembly and analysis of *Azadirachta* fruit transcriptome using next-generation sequencing technology. We believe that our study shall offer valuable insights towards realizing the larger vision of understanding the key medicinally active compounds and their pathways.**

**Keywords:** *Azadirachta indica*, *de novo* sequencing, fruit and seed, transcriptome.

*AZADIRACHTA INDICA* A. Juss, the Indian neem (margosa) is a wonder tree of sorts, largely so in the Indian context. Endemic to the Indian subcontinent[1,2], it has also been propagated to other parts of the world. It belongs to the mahogany family, Meliaceae[3], and is known to have immense medicinal and agricultural value. The scientific name of neem originates from the word azadirachtin (Azad + Darakht + i, implying the 'free tree of India' in Persian[2]), the most well-studied compound present in the neem seeds[3] that plays a vital role in conferring insecticidal properties to neem, along with other steroids. Azadirachtin, extracted primarily from the seed kernels, is a complex, modified tetranortriterpenoid and a powerful insecticide[4,5]. Mature fruits of neem are yellowish drupes with a bitter-sweet mesocarp and white inner endocarp that encloses a single, elongated seed with a brown coat. Neem oil extracted from fruits and crushed seeds is often used in cosmetics and some forms of traditional medicine[3,6]. The residual crushed seed material when mixed with soil, serves as a nematicide[7]. The rest of the neem tree, namely leaves, twigs, bark and roots also have various medicinal properties[3,6]. Neem flowers and fruits are used in certain food preparations in some parts of India. Due to its diverse and multitude of health benefits, neem is populalry known as nature's pharmacy.

The neem tree, being of such commercial interest and diverse usage, naturally emerges as an important candidate for sequencing. Despite the wealth of information on the usefulness of the neem tree, there are few molecular studies done on the plant, including lack of any sequence information on its genome and transcriptome. It is surprising that there is virtually no prior information on either the coding sequence, ESTs or any form of gene expression data available on neem fruit, seed or from any other organ. This is primarily due to the bias towards characterization and synthesis of the bioactive compounds of neem for commercial benefits and a lack of systematic approach to study physiological aspects of neem. Also in India, the facilities and manpower to conduct studies on high-throughput genomics were not available till recently. Fruit, being one of the most important organs, and the primary source of azadirachtin, was the obvious first choice for our transcriptome study.

In this article, we describe the *de novo* sequencing of transcriptome from neem fruit bearing the seed, its assembly and preliminary analysis using RNA-Seq, a high-throughput next-generation sequencing technology, which allows one to study the full transcriptome profile of an organism, without probe or primer design, by directly sequencing the cDNA from the target sample. RNA-seq, is used to reveal transcriptome landscape and dynamics of living organisms with high sensitivity and accuracy[8–13].

Commercial next-generation sequencing instruments use either DNA polymerase- or DNA ligase-based chemistries to produce sequence reads. As we used sequencing-

[+]Equal contribution.
*For correspondence. (e-mail: binay@ganitlabs.in)

by-synthesis method utilizing DNA polymerase and modified fluorescent nucleotides, we will explain the technology a bit further. Essentially, the concept behind the technology is the same as Sanger sequencing, except that the process is made massively parallel, making it possible to generate huge amount of data in a single day. This parallelization helps read hundreds of millions of growing chains of DNA, rather than tens and hundreds as done with capillary Sanger sequencing. Using modified DNA polymerase with improved processivity, better 5′–3′ exonuclease activity and low error-incorporation rate (with minimum error and maximum fidelity) along with the use of modified nucleotides (for example, reversible terminator nucleotides as in the case of Solexa sequencing), the process has been made robust and capable of sequencing with read lengths up to 150 nucleotides (with Solexa technology). This robust chemistry along with the use of better optics that can image the growing nucleotide chain flawlessly for days and with a speed that is comparable with the processivity of the enzyme has made the entire process of high-throughput DNA sequencing seamless and free from user intervention. Different technologies available for high-throughput DNA sequencing and methods involved are extensively reviewed elsewhere[14,15]. High-throughput DNA sequencers produce large amount of data (in the order of terabytes of raw image data) that need to be stored, analysed, managed, interpreted and retrieved for future use. Managing these data deluge along with the complexity of data storage, archival, sharing and security combined with biological interpretation and follow-up validation are some of the key issues in the next-generation sequencing field.

In this article, we report, *de novo* assembly and analysis of transcriptome from mature neem fruit with sequencing-by-synthesis method using Solexa/Illumina GAIIx instrument.

## Materials and methods

### Sample collection, identification and RNA extraction

Seed-bearing mature fruits of *Azadirachta indica* A. Juss were collected from locally grown tree. Genus and species of the plant was confirmed at the South Regional Centre, Botanical Survey of India, Tamil Nadu Agricultural University Campus, Coimbatore, India, using the herbarium of the twigs bearing fruits. The seed-bearing fruit samples were collected and immediately flash-frozen in liquid nitrogen and stored at –80°C until further use. Total RNA was extracted using Plant Total RNA extraction kit (Bioteke, China). We assessed the quality and quantity of RNA by using NanoDrop and Qubit methods. We checked RNA integrity by running total RNA on Agilent Bioanalyzer RNA 6000 Nano chip.

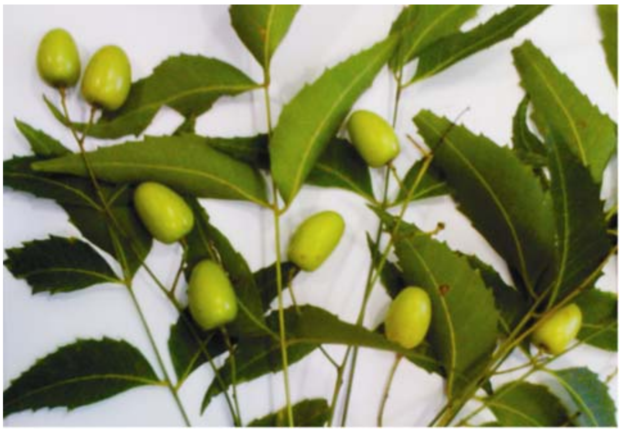

**Figure 1.** Morphology of the twig of *Azadirachta indica* plant bearing fruit (the ones used for the sequencing study).

### Sequencing library preparation

RNA library was prepared using TruSeq RNA library prep kit (Illumina) following the manufacturer’s instructions. Four micrograms of RNA was subjected to mRNA selection using poly-T oligo-attached magnetic beads followed by mRNA fragmentation to sizes between 100 and 300 nt by incubating in fragmentation mix for 8 min at 94°C. First-strand cDNA was synthesized by adding 1 μl of SuperScript II reverse transcriptase (Invitrogen) to the solution containing primed RNA and first-strand master mix followed by incubation at 25°C for 10 min, 42°C for 50 min and 70°C for 15 min. The complementary second-strand cDNA was synthesized by incubating first-strand cDNA in second-strand master mix containing RNAase H and DNA polymerase I at 16°C for 1 h. End repair was performed to remove the 3′ overhangs and fill the 5′ overhangs by incubating the DNA in end-repair mix containing T4 polynucleotide kinase, T4 DNA polymerase and large (klenow) fragment of DNA polymerase I for 30 min at 30°C, and purified using Agencourt AMPure XP beads (Beckman Coulter) according to the manufacturer’s recommendations. A-tailing of DNA was performed at 37°C for 30 min with klenow fragment followed by ligation of TruSeq adaptors using T4 DNA ligase by incubating at 30°C for 10 min. The reaction was stopped by adding stop-ligase mix and purified by using Agencourt AMPure XP beads. The adaptor ligated DNA was subjected to PCR enrichment with adaptor complementary primers for 15 cycles followed by clean-up using Agencourt AMPure XP beads. The quality and quantity of the library were estimated by NanoDrop and Picogreen method (Qubit, Invitrogen), and the size distribution was analysed on Agilent Bioanalyzer using high sensitivity DNA chips.

### Quantification of prepared library and sequencing

Accurate quantification of prepared library was performed using SYBR-Green-based qPCR reagents (Kapa

Biosystems). The qPCR results were compared with the predetermined concentration of the phiX library. Six picomoles of the *Azadirachta* seed-bearing fruit-RNA library was seeded for cluster generation in cBot (Illumina). A 72 bp paired-end sequencing was performed on the Genome Analyzer IIx platform (Illumina) following the manufacturer's recommendations.

### *Sequence quality control and pre-processing*

Raw sequence reads were analysed by in-house written scripts and checked for good quality (≥ 20 Phred score) bases in the forward and reverse reads for the entire run.

### *De novo transcriptome assembly*

The genome-independent transcriptome assembler, Trinity[16], was run with default parameters (kmer size of 25 and minimum contig length of 48) using the following command: *perl Trinity.pl -seqType fq -output Trinity -SS_lib_type FR -paired_fragment_length 150 -left Fruit_R1.fastq -right Fruit_R2.fastq -CPU 4 -run_butterfly -bfly HeapSpace 10000M*, on the paired-end sequenced reads. The paired-fragment length specified as 150, corresponds to the inner insert size of the sequenced reads, as indicated by the Agilent Bioanalyzer run (Supplementary File 1). The transcripts identified are processed through the similarity-based analyses pipeline for annotation and functional classification.

### *Similarity-based analysis pipeline*

The transcripts were first processed through 'blastn' program[17] with the 'megablast' option, against the entire non-redundant nucleotide database. The transcripts that did not yield any blast hits in this step were further processed using the 'blastx' program, against the entire non-redundant (nr) protein database. This step translates the query sequence into all six reading frames, before searching the protein database. Due to the lack of availability of sequencing data for the Meliaceae family, the blast-positive hits in the second step, with an expect value above 0.001 were filtered out. The transcripts which yielded no hits or no significant hits using blastx (≥ 0.001) were serially processed using the blastn program[17] with the 'megablast' option, against the refseq RNA[18], expressed sequence tags (EST)[19] and transcriptome shotgun assembly (TSA) databases respectively, each time using only those transcripts which failed to yield any hits with the previous blast database.

### *Functional classification*

As described earlier, an expect value cut-off of 0.001 was used for the blastx analysis for predicting the biological functionality of all fruit transcripts at the first level. The blastx hits were subsequently mapped using Blast2GO[20] to their corresponding gene ontology (GO) accession and GO terms. Blast2GO is the ideal choice for whole-transcriptome analysis, as it is one of the most integrated tools for end-to-end analysis of sequencing data and is available as a platform-independent JAVA tool. However, the RAM requirements for annotating blastx XML output file are overwhelming; thus making the JAVA tool not feasible to use. We used the command line version, wherein we had to allocate at least 32 GB RAM for the process to run smoothly and generate the GO annotations (.annot files). The annotation files were imported into the Blast2GO JAVA interface and processed further to yield functional annotation based on biological process and molecular function.

### *Phylogenetic analyses*

Complete sequences of 24 plant genomes were obtained, from Phytozome v7.0 (http://www.phytozome.net/) for a majority of the species; from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/plants) for *Malus x domestica*, *Theobroma cacao*, *Brassica rapa*, *Lotus japonicus*, *Solanum tuberosum* and *Solanum lycopersicum*, and from https://strawberry.plantandfood.co.nz/ for *Fragaria vesca.* Transcriptome sequences were obtained from TIGR (http://plantta.jcvi.org/cgi-bin/plantta_release.pl). The evolutionarily conserved plastid-encoded ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit (*rbcL*) gene sequence was downloaded for all 24 plants from the NCBI nucleotide database. The sequence for the same was extracted from the Trinity-reconstructed transcripts for the *Azadirachta* fruit, based on the blast annotations. The *rbcL* gene sequences from the 25 plant species were aligned using ClustalW multiple alignment[21] (http://www.genome.jp/tools/clustalw), with options set to 'Slow/Accurate' and 'DNA' and the corresponding ClustalW dendrogram with branch lengths was obtained (Figure 2).

### *Comparison of annotated Azadirachta fruit transcripts with other species*

The transcripts of 24 plant species used in constructing the *rbcL*-based phylogeny downloaded from the TIGR (http://plantta.jcvi.org/cgi-bin/plantta_release.pl) were processed using 'blastn' program[17] with the 'megablast' option. The annotations thus obtained for *Azadirachta* fruit transcriptome (*A*) were compared with the first-level megablast-annotations (*B*). The pairwise common intersection percentage of annotated genes was calculated using the following formula:

$$\left(\frac{A \cap B}{A + B - A \cap B}\right) \times 100.$$

### Nucleotide content analysis

The percentage compositions of A, T, G and C nucleotides were calculated for each transcript. Their frequencies are further analysed across the entire distribution of transcripts.

### Analysis of gene expression indices

The average k-mer abundance value acts as a proxy for transcript expression, and is returned by Trinity as approximate indices for expression levels (FPKM) for each transcript[16]. A frequency histogram of these indices was obtained to analyse the spectrum of transcripts falling into low, medium and high expression-level categories.

## Results and discussion

### Overview of Trinity transcriptome and downstream analyses

This article focuses on *de novo* sequencing, assembly and analysis of transcriptome from mature seed-bearing neem fruit. In *de novo* transcriptome sequencing, there is no reference genome for alignment of the sequenced reads. The reads need to be assembled into transcripts in a genome-independent manner[13]. We assembled the RNA-Seq reads into transcripts using Trinity (http://trinityrnaseq.sourceforge.net/), a reference genome-independent assembler[16]. The transcript identification by Trinity is divided into three steps: Inchworm, Chrysalis and Butterfly[16]. Together, they assemble the RNA-seq reads into contigs, cluster the contigs, construct de Bruijn graphs[22,23] for each cluster (representing the transcriptional complexity for a gene), partition the reads among each graph, and finally trace the paths in each graph to report full-length transcripts, for alternatively spliced isoforms as well as for paralogous genes. Trinity has been found to efficiently reconstruct the transcriptome, inclusive of the splicing events and transcripts resulting from recent duplication events, better than other available *de novo* transcriptome assemblers[16]. We further annotated the Trinity-detected transcripts using megablast and blastx, against the non-redundant nucleotide and protein databases, the refseq RNA and EST databases. We found approximately 5% of the neem fruit transcripts to be unannotable using similarity-based analyses, suggesting them to be fruit-specific in nature. The Trinity-detected transcripts annotated using blastx, were also functionally classified in terms of molecular function and biological processes using Blast2GO mapping[20].

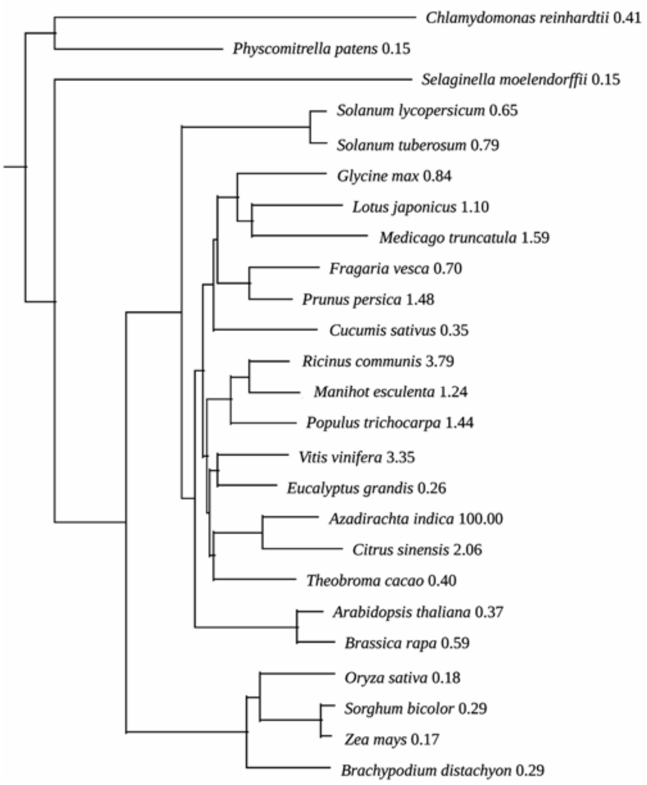

**Figure 2.** *rbcl*-based plant phylogeny. The phylogeny was constructed using ClustalW alignment of *rbcL* gene sequences. Numbers adjacent to each species indicate the percentage overlap between annotated genes according to megablast (blastn) to the non-redundant nucleotide database of the transcriptome for that species with that of the *Azadirachta indica* transcriptome.

### Identifying the transcriptome of neem fruit

Results from the RNA-seq transcriptome study are summarized in Table 1. The RNA Seq[12] analyses of the mature neem fruit yielded 2,865,266 paired reads. About 93.94% reads contained 75% or more of the bases with Phred quality score ≥ 20, and 93.6% of the bases in all reads had a Phred quality score greater than 20. Inchworm extracted 57,453,100 kmers of length 25 nucleotides from the mature fruit RNA-Seq reads. The kmers were assembled into 154,372 contigs of lengths varying from a minimum of 48 nucleotides to 6385. The cumulative transcriptome size of the neem fruit was estimated to be 18,789,522 nucleotides. Out of the total 26,908 reconstructed transcripts, ~57% could be annotated using megablast (Table 1, Supplementary File 2). These transcripts were found to be significantly similar to various mRNAs from other plant species (Supplementary File 2). The spectrum of blast similarities varied from 1 to 49 plant species (Supplementary File 2). Nearly 86% of the megablast-negative transcripts were further annotated using blastx (Table 1, Supplementary File 3). The remaining transcripts were annotated using megablast against the refseq RNA and EST databases respectively, to a small extent: 0.12% and 18.36%. No hits resulted from further megablast against the TSA database. About

5% of the total transcripts failed to be annotated using the similarity-based annotation pipeline, and could potentially contain transcripts unique to the neem fruit.

The Trinity approach is known to provide a consolidated solution for transcriptome reconstruction in any sample, especially in the absence of a reference genome[16]. It performs better than the other *de novo* transcriptome assemblers such as Oases, since it does not construct scaffolds by inserting Ns between two transcripts, and instead, retains intact transcript sequences (data not shown).

## *Comparison of annotated transcripts across species*

Annotated transcripts for the neem fruit when compared with those for the other species (Figure 2), yielded the best overlap with *Ricinus communis* (3.79), *Vitis vinifera* (3.35) and *Citrus sinensis* (2.06). The lowest overlap indices were observed for the mosses: *Physcomitrella patens* and *Selaginella moellendorfii* (0.15). The extent of overlap with different species approximately fits phylogenetic expectations (Figure 2). The overlap accounts for variation in the number of annotated genes for the different species and hence is comparable between species. It is however, subject to bias resulting from variation in the actual number of transcripts identified for different species, a factor that cannot be normalized, given the difference in technologies used for sequencing the different transcriptomes. The overall low overlap indices for all species versus neem can be explained because the transcriptome of the neem fruit is being compared to the pooled transcriptomes of the other plants.

**Table 1.** Similarity-based analyses attributes of the mature neem fruit transcriptome reconstructed by Trinity

| | |
|---|---|
| No. of paired reads | 2,865,266 |
| Total transcripts | 26,908 |
| No. of inchworm kmers | 57,453,100 |
| No. of inchworm contigs | 154,372 |
| Minimum contig length | 48 |
| Maximum contig length | 6385 |
| Size of the transcriptome | 18,789,522 |
| Megablast positive | 15,386 |
| Megablast negative | 11,522 |
| Megablast negative, blastx positive hits | 9859 |
| Megablast negative, blastx negative hits | 1663 |
| Megablast negative, blastx negative, refseq rna positive hits | 2 |
| Megablast negative, blastx negative, refseq rna negative hits | 1661 |
| Megablast negative, blastx negative, refseq rna negative, est positive hits | 305 |
| Megablast negative, blastx negative, refseq rna negative, est negative hits | 1356 |

The similarity-based analyses pipeline was run on the Trinity-derived *Azadirachta* fruit transcripts. The intermediate step statistics during the transcriptome reconstruction by Trinity is summarized here. The number of similar and dissimilar hits obtained during the various intermediate steps in the pipeline is also listed.

The plastid-encoded *rbcL* phylogeny is commonly used in plant evolution, as it agrees well with the known phylogeny for Viridiplantae members[24]. *C. sinensis* (orange), the closest phylogenetic species to neem does feature in the top three, even though it is not the best overlapping species in annotated transcripts. The closest overlapping species to neem is *R. communis* (castor). Indeed, evolution of the fruit transcriptome does not have to match the actual course of evoution of plants, and can be regulated similarly in two not so closely related plants. Interestingly, *R. communis* and *C. sinensis* share more annotated transcripts than *R. communis* and neem fruit, and the three together share 46 common annotated transcripts (Supplementary File 4).

## *Functional annotation*

About 93% (25,102 out of a total of 26,908) of transcripts were blastx-positive and were processed further by Blast2GO gene annotation tool[20]. GO terms and accessions were assigned to about 85% (21,351 out of a total of 25,102) of blastx-postive fruit transcripts using Blast2GO. The functional classification based on molecular function and biological process terms is depicted in Figure 3 *a* and *b* respectively. A frequency distribution of the number of sequences mapped to biological process and molecular function is depicted by Figure 3 *c* and *d* respectively. Mapping of the transcripts to fully detailed biological process and molecular function categories is presented in Supplementary Files 5 and 6. These attributes correlate well with the physiological state of the fruit. For instance, there are more number of transcripts catering to metabolic processes and stress response, and a lower frequency of transcripts relating to cell growth and differentiation, or metabolic processes involved in utilization of micro nutrients. The blastx hits that were not assigned GO terms are putatively neem-specific transcripts and are currently being validated by orthogonal experimental approaches. This study will help us understand both generic and specific aspects of transcript function in the neem fruit and its seed.

The RNA-Seq technology allows selective sequencing of transcripts and transcriptional profiling of the tissue of interest[13]. Following data generation, the subsequent stages of data analysis and interpretation involve heuristics-based approaches to assemble the transcriptome and forge its functionalities. In the absence of physiological data pertinent to the neem fruit, we attempted to verify the predicted functionalities based on existing transcriptome data for the fruit tissue of a nearest sequenced neighbour. Based on the phylogenetic proximity of neem to *C. sinensis*, and availability of transcriptome data for *C. sinensis*, we narrowed down our comparison to this closely related species. The second-level mapping of GO terms to the neem fruit transcripts matches well with that

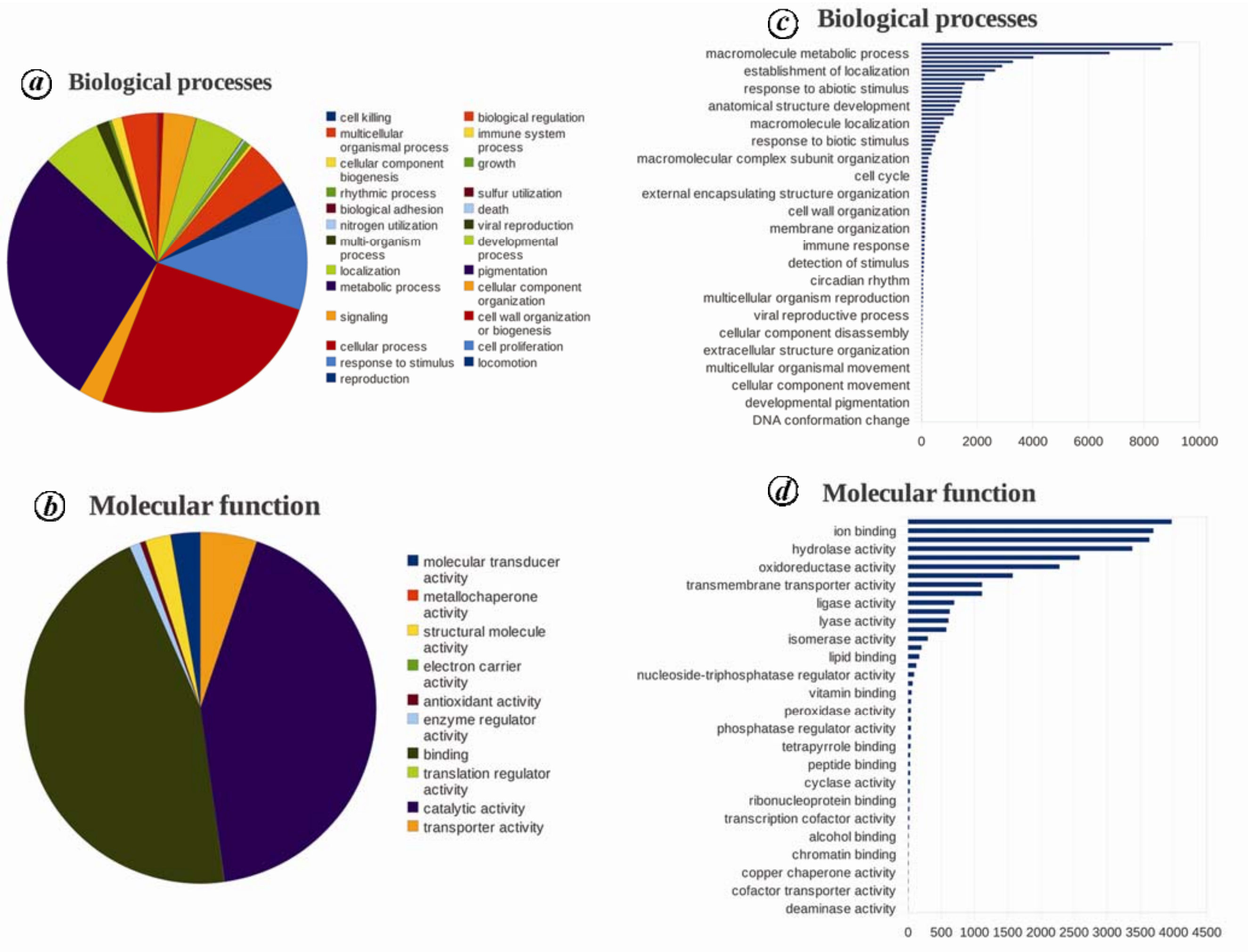

**Figure 3.** Gene Ontology (GO)-term-based functional categorization of genes/transcripts. Blastx (*e*-value cut-off of 0.001) results for fruit transcripts were formatted as xml, and subject to Blast2GO mappings. The GO terms and scores were categorized based on biological processes (***a*** and ***c***) and molecular function (***b*** and ***d***). The pie charts depict the categorization at the second level, and the frequency histograms depict the same at a deeper third level.

observed with citrus fruit transcriptome[25]. These surprising similarities in the biological process and molecular function among the neem and *C. sinensis* fruits, corroborate the systematic classification that tags them under the same order Sapindales. The correlation of different biological processes and molecular functions to physiology of the fruit allows us to compartmentalize the differentiating factors specific to the neem fruit and also aids in identification of traits common to fruits, in general.

## *GC content and approximate expression level index analyses of fruit transcripts*

The profiles of A and T nucleotides follow the same pattern (Figure 4 *a*): almost all the transcripts contain 20–40% As and the same composition range of Ts. There are practically no or very few transcripts with compositions of As and Ts in any other interval range. The composition profiles of G and C, on the other hand, follow a different pattern: about two-thirds of the transcripts fall into the 20–40% G and C each, per transcript category, whereas the remaining one-third of the transcripts have a G and C% each, ranging from 0 to 20% per transcript.

The approximate expression level indices of neem fruit transcripts, as reported by Trinity range between 7.82 and 78,934.03 (Figure 4 *b*). According to the frequency histogram of these indices, transcripts with very low expression levels (expression indices <1000) are highly frequent (~25,000). A rapid decrease is observed in the transcript frequency (from 25,000 to 100) for the immediate next interval of gene expression (1000–2000). The frequency of transcripts decreases to ~1–10 for intermediate to higher levels of expression (> 3000).

Combining the GC content analyses with the expression level analyses of fruit transcripts, we find that the %A and %T in each transcript appear to very tightly regulated to be mostly in the 20–40% range (Figure 4 *c*).

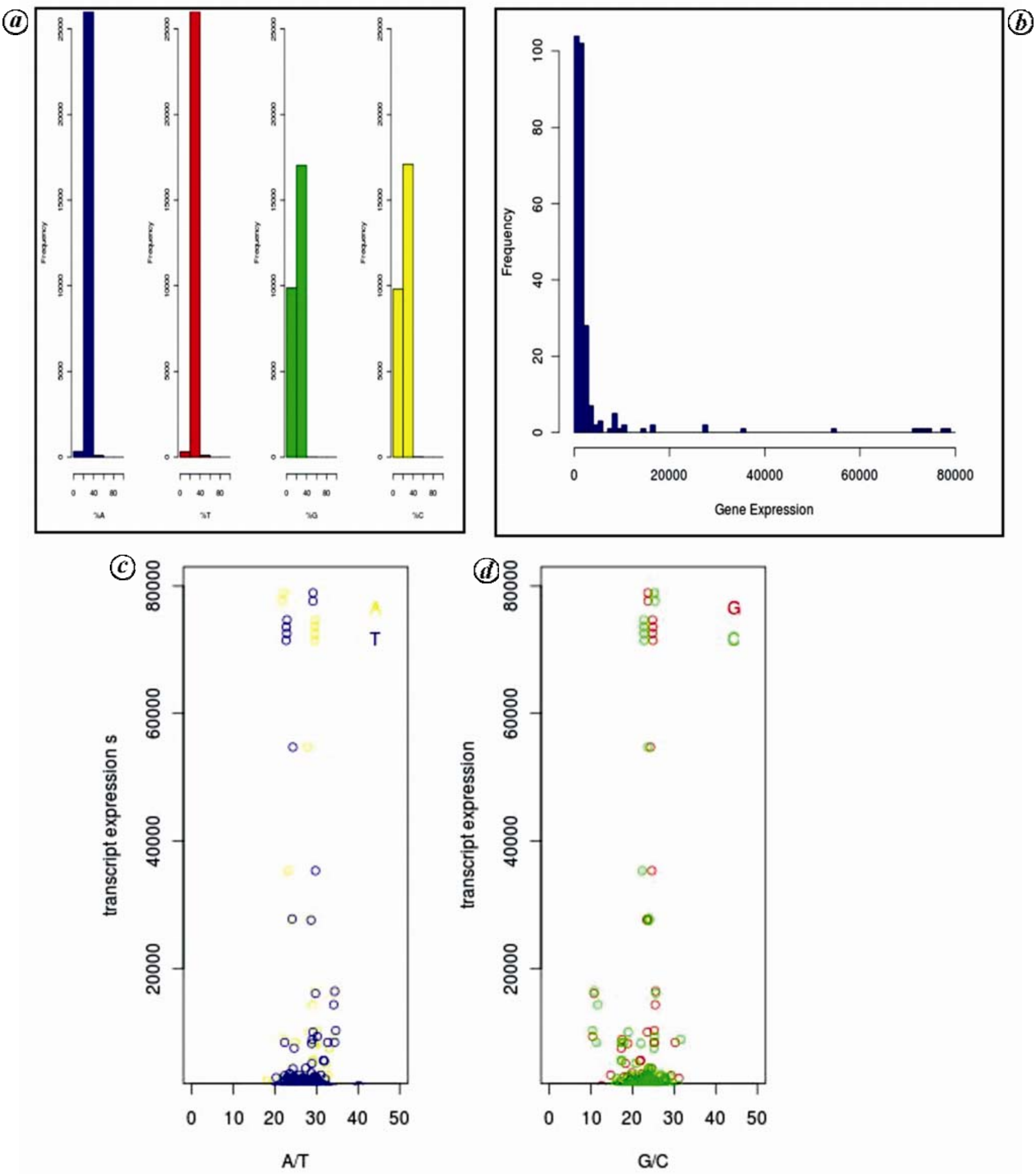

**Figure 4.** Nucleotide composition and expression level indices of neem fruit transcripts. ***a***, Percentages of A, T, G and C are calculated for each transcript and plotted as frequency histogram with bins of size 20, for the entire distribution of Trinity-identified transcripts. ***b***, The approximate gene expression indices for transcripts are obtained from Trinity. The frequency histogram is plotted at intervals of 1000, with a truncated *Y*-axis scale of 100 to avoid the very high frequency transcripts (up to 25,000) with a very low gene expression index (≤ 1000). Correlation of expression level indices with percentages of A (yellow circles) and T (blue circles) (***c***), and G (red circles) and C (green circles) (***d***) of transcripts.

Based on the profiles of %G and %C, however, there appear two populations of transcripts, one with low %GC but low to intermediate low levels of expression, and the other with higher %GC but having low to very high levels of expression (Figure 4 *d*). This provides preliminary data for correlation between GC content and gene expression levels of transcripts in neem. A detailed study is needed to establish whether there is any definitive

**Table 2.** Categorization of genes with varying ranges in expression value indices

| Expression values | | | | |
|---|---|---|---|---|
| >20,000 | 15,000–20,000 | 10,000–15,000 | 5000–10,000 | 2000–5000 |
| Allergen11S globulin precursor mRNA [71,482.42–78,934.03] | Unannotated [16,083.98–16,451.26] | Metallothionein-like protein (MT45) mRNA [10,045.76] | Unannotated [5,138.42] | Ubiquitin_precursor_ubi)_ mRNA [2,044.77] |
| | | | Cyc15 mRNA for extension [5,519.75–5,640.45] | Thaumatin-like mRNA [2,067.459] |
| 26S ribosomal RNA [54,713.66] | | Envelope glycoprotein I [10,298.31] | Metallothionein-like protein class II (MT) mRNA [7,541.78–8,425.56] | Polyubiquitin mRNA [2,279.86, 2,301.98, 2,378.83, 2,791.65] |
| Alpha tubulin mRNA [35,358.39] | | Unannotated [14,333.96] | Glycine-rich protein/ late embryogenesis abundant protein [8,437.22–8,902.03] | Translation elongation factor 1-alpha mRNA [2,291.38, 2,357.09–2,376.80, 2,393.46–2,482.25, 2,526.54–2,672.47] |
| C2 calcium/lipid-binding region, CaLB domain containing protein [27,804.33–27,596.56] | | | Unannotated [8,449.09] | |
| | | | *Ricinus communis* hypothetical protein [9,379.54] | |

The approximate expression value indices (FPKM) are obtained from Trinity. Genes annotated according to the similarity-based annotation pipeline are indicated in the various expression level categories.

relationship between the two and if there is evolutionary conservation of this phenomenon in all higher plants.

The allergen11s globulin precusor mRNA appears to be the most expressed transcript (Table 2). It is similar in sequence to that found in the *Pistacia vera*[26], associated with allergy to pistachio nut. Most tree nut allergens are known to be seed-storage proteins (particularly, the globulins), which accumulate during the process of seed development[27]. This coincides well with the developmental stage of the sequenced sample as a mature seed-bearing fruit. Expression studies on the 11S globulin gene family in *R. communis* also confirm that its seed-specific expression is developmentally regulated and is highest upon the maximum expansion of the cellular endosperm[28]. The 26S ribosomal RNA and alpha-tubulin mRNA are the other transcripts which fall into the highly expressed transcript category (> 20,000). Both are well-known internal controls in most gene expression analyses in plants[29,30]. This fact correlates well with their high expression level indices in our observations. The polyubiquitin mRNA and its precursor, and the elongation factor 1-alpha mRNA fall into the low expression level transcript category (2000–5000), while the metallothionein-like protein class II (MT) mRNAs are expressed to an intermediate level. We did not find a significant difference in the spectrum of molecular function and associated biological processes when we performed the Blast2GO mapping analyses[20] with only the highly expressed transcripts (data not shown).

Our current work on the characterization of *A. indica* fruit transcriptome is part of a larger effort to sequence the whole genome of the plant along with transcript discovery in its various organs. Full genome sequence, its relationship with other plant genomes along with the discovery of transcripts expressed in various organs will shed more light on the evolutionary significance of neem and characterize the known/novel pathways that are involved in the synthesis of biologically active compounds.

1. Stoney, C., Fact sheet on *Azadirachta indica* (neem) – a versatile tree for the tropics and subtropics. Publication of Forest, Farm, and Community Tree Network (FACT Net), Arkansas, United States, 1997; www.winrock.org/forestry/factnet.htm.
2. Niharika, A., Aquicio, J. M. and Anand, A., Antifungal properties of neem (*Azadirachta indica*) leaves extract to treat hair dandruff. *Int. Sci. Res. J.*, 2010, **2**, 244–252.
3. Biswas, K., Chattopadhyay, I., Banerjee, R. K. and Bandyopadhyay, U., Biological activities and medicinal properties of neem (*Azadirachta indica*). *Curr. Sci*., 2002, **82**, 1336–1345.
4. Govindachari, T. R., Chemical and biological investigations on *Azadirachta indica* (the neem tree). *Curr. Sci*., 1992, **63**, 117–122.
5. Veitch, G. E., Beckmann, E., Burke, B. J., Boyer, A., Maslen, S. L. and Ley, S. V., Synthesis of azadirachtin: a long but successful journey. *Angew. Chem., Intl. Ed. Engl.*, 2007, **46**, 7629–7632.
6. Chatterjee, A. and Pakrashi, S. (eds), *The Treatise on Indian Medicinal Plants*, Publications and Information Directorate, CSIR, New Delhi, India, 1994, vol. 3, p. 76.

7. Akhtar, M., Nematicidal potential of the neem tree *Azadirachta indica* (A. Juss). *Int. Pest Manage. Rev*., 2000, **5**, 57–66.
8. Ozsolak, F. and Milos, P. M., RNA sequencing: advances, challenges and opportunities. *Nature Rev. Genet*., 2011, **12**, 87–98.
9. Wang, Z., Gerstein, M. and Snyder, M., RNA-seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet*., 2009, **10**, 57–63.
10. Marguerat, S. and Bahler, J., RNA-seq: from technology to biology. *Cell. Mol. Life Sci*., 2010, **67**, 569–579.
11. Wilhelm, B. T. and Landry, J. R., RNA-seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*, 2009, **48**, 249–257.
12. Wang, Z., Gerstein, M. and Synder, M., RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet*., 2009, **10**, 57–63.
13. Martin, J. A. and Wang, Z., Next-generation transcriptome assembly. *Nature Rev. Genet*., 2011, **7**, 671–682.
14. Metzker, M. L., Sequencing technologies – the next generation. *Nature Rev. Genet*., 2010, **11**, 31–46.
15. Shendure, J. and Ji, H., Next-generation DNA sequencing. *Nature Biotechnol*., 2008, **26**, 1135–1145.
16. Grabherr, M. G. *et al*., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnol*., 2011, **29**, 644–652.
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., Basic local alignment search tool. *J. Mol. Biol*., 1990, **215**, 403–410.
18. Pruitt, K. D., Tatusova, T., Klimke, W. and Maglott, D. R., NCBI Reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*., 2009, **37** (Database Issue), D32–D36.
19. Boguski, M. S., Lowe, T. M. and Tolstoshev, C. M., dbEST – database for ‘expressed sequence tags’. *Nature Genet*., 1993, **4**, 332–333.
20. Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M. and Robles, M., Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, 2005, **21**, 3674–3676.
21. Thompson, J. D., Higgins, D. G. and Gibson, T. J., ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., 1994, **22**, 4673–4680.
22. de Bruijn, N. G., A combinatorial problem. *K. Ned. Akad. Wet*., 1946, **49**, 758–764.
23. Good, I. J., Normal recurring decimals. *J. London Math. Soc*., 1946, **21**, 167–169.
24. Magallon, S. and Sanderson, M. J., Relationships among seed plants inferred from highly conserved genes: sorting conflicting phylogenetic signals among ancient lineages. *Am. J. Bot*., 2002, **89**, 1991–2006.
25. Liu, Q. *et al*., Transcriptome analysis of a spontaneous mutant in sweet orange [*Citrus sinensis* (L.) Osbeck] during fruit development. *J. Exp. Bot*., 2009, **60**, 801–813.
26. Beyer, K., Grishina, G., Bardina, L., Stalcup, D. and Sampson, H., Identification and cloning of 11S globulin, a new minor allergen from pistachio nut. Submitted to the Allergen nomenclature subcommittee of the International Union of Immunological Societies, and to the EMBL/GenBank/DDBJ databases, 2008.
27. Roux, K. H., Teuber, S. S. and Sathe, S. K., Tree nut allergens. *Int. Arch. Allergy Immunol*., 2003, **131**, 234–244.
28. Chileh, T., Esteban-Garcia, B., Alonso, D. L. and Garcia-Maroto, F., Characterization of the 11S globulin gene family in the castor plant *Ricinus communis* L. *J. Agric. Food Chem*., 2010, **58**, 272–281.
29. Singh, K. *et al*., 26S rRNA-based internal control gene primer pair for reverse transcription-polymerase chain reaction-based quantitative expression studies in diverse plant species. *Anal. Biochem*., 2004, **335**, 330–333.
30. Dong, L., Sui, C., Liu, Y., Yang, Y., Wei, J. and Yang, Y., Validation and application of reference genes for quantitative gene expression analyses in various tissues of *Bupleurum chinense*. *Mol. Biol. Rep*., 2010, **38**(8), 5017–5023.

ACKNOWLEDGEMENTS. The study is financed by grants from the Department of Information Technology, Ministry of Communications and Information Technology, Government of India, and the Department of Information Technology, Biotechnology and Science & Technology, Government of Karnataka. We thank Prof. N. Yathindra for encouragement.

AUTHOR CONTRIBUTION. B.P. conceived, planned, designed and directed the project, N.M.K. and S.P. designed analysis pipeline and guided the process, N.M.K., S.P., P.G., P.J., R.C. and S.V. analysed the data, D.S.A., A.K.H. and B.K. produced the sequencing data. N.M.K., S.P. and B.P. wrote the manuscript.

SUPPLEMENTARY INFORMATION. Supplementary files can be downloaded from www.ganitlabs.in/projects/AZ/CS2011/Supp

DISCLOSURE. Ganit Labs is a public–private partnership initiate between Institute of Bioinformatics and Applied Biotechnology (IBAB) and Strand Life Sciences (SLS). Data derived from this research will be uploaded on an information portal and will be available free-of-cost to all academic researchers and for not-for-profit use.

CONFLICT OF INTEREST. Authors declare no conflict of interest.

Received 7 October 2011; accepted 7 December 2011