

Supporting Information

RNA-seq-mediated transcriptome analysis of actively growing and winter dormant shoots identifies non-deciduous habit of evergreen tree tea during winters

Asosii Paul^{1*†}, Ashwani Jha^{2*}, Shruti Bhardwaj¹, Sewa Singh¹, Ravi Shankar², Sanjay Kumar¹

¹Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, P.O. Box No. 6, Palampur-176061, Himachal Pradesh, India, ²Studio of Computational Biology & Bioinformatics, CSIR-Institute of Himalayan Bioresource Technology, P.O. Box No. 6, Palampur-176061, Himachal Pradesh, India.

[†]Present Address: National Institute of Plant Genome Research, Aruna Asaf Ali Marg, P.O. Box No. 10531, New Delhi-10067, India.

*Contributed equally

Correspondence and requests for materials should be addressed to S.K. (sanjaykumar@ihbt.res.in; sanjayplp@rediffmail.com)

Supplementary method:

Read generation and analysis

Read generation and analysis was performed essentially as described previously^{1,2,3} with minor modifications. Paired-End (PE) reads (36 X 2 bp) were generated using CASAVA package (version 1.3) in fastq format obtained from Illumina Genome Analyzer Iix. To minimize sequencing error last three bases of 3' end from each read were removed which is generally known to exhibit sequencing error. FilteR¹ tool (http://scbb.ihbt.res.in/SCBB_dept/filter.php). FilteR detects adapter sequence contamination as well as poor read quality and uses the quality scoring scheme provided by Illumina. It also provides “Recommender” option to decide the suitable cut-off to perform read screening by calculating average read quality positionally. This helps selective trimming of the reads rather than discarding the entire read. K-mer frequency measurement was performed to filter out reads with lower k-mer frequency for default value, which could be a result of sequencing error. *De novo* assembling of high quality reads was performed using SOAP*denovo*⁶ program. The high quality reads were split into smaller fragments, the ‘k-mers’, for assembly to produce contigs, using the de Bruijn graph. K-mer size of 21 achieved the best balance between the number of contigs produced, coverage and average sequence length attained. PE option of assembling with distance of 200 bp was applied for more effective assembling of PE reads. The same parameters was also used to build the scaffold sequences by merging two contigs into single scaffold sequence that shared the read pairs. Protocol used in *de novo* assembling and transcript analysis of assembled sequences is shown in Supplementary Fig. S7 as described previously¹.

Sequence redundancy was removed by searching similar sequences with minimum similarity cut-off of 95% using CD-HIT-EST⁷. CD-HIT was used for further clustering with 90% similarity cut-off. The algorithms for various clustering programs differed in their approach of clustering and combined use of such clustering tools with different algorithms are reported to yield better results⁶. For the same reason clustering process was supplemented with TGICL-CAP3⁸ clustering based on terminal region matching for at least 40 bp and 90% identity. The resulting singletons and consensus contigs were merged to get the final list of assembled transcripts. A set of script was developed to detect contigs/scaffolds that had no sequence similarity but belonged to same gene's different regions. These were clustered together to represent as a single transcript. The best BLASTX hits for all contigs were looked for common NR database ID for a particular gene/peptide and all associated contigs showing highest similarity to the same sequence but its different regions, were assigned to the same ID group. Genes with same ID groups were considered unigenes.

Transcript abundance measurement

Mortazvi et al.⁴ developed read per exon kilobase per million (RPKM)-based method to analyse expression level of unigenes. In brief, filtered reads were mapped back to unigene to calculate total reads mapping the unigene. Mapping was done with 2 mismatches using SeqMap tool. RPKM value was calculated using rSeq tool (Supplementary Table S1) employing the following formula:

$$\text{RPKM} = (\text{number of reads mapping on unigene}) \times 1000 \times 1 \text{ million reads} / [(\text{length of unigene}) \times (\text{number of total reads in experiment})]^4.$$

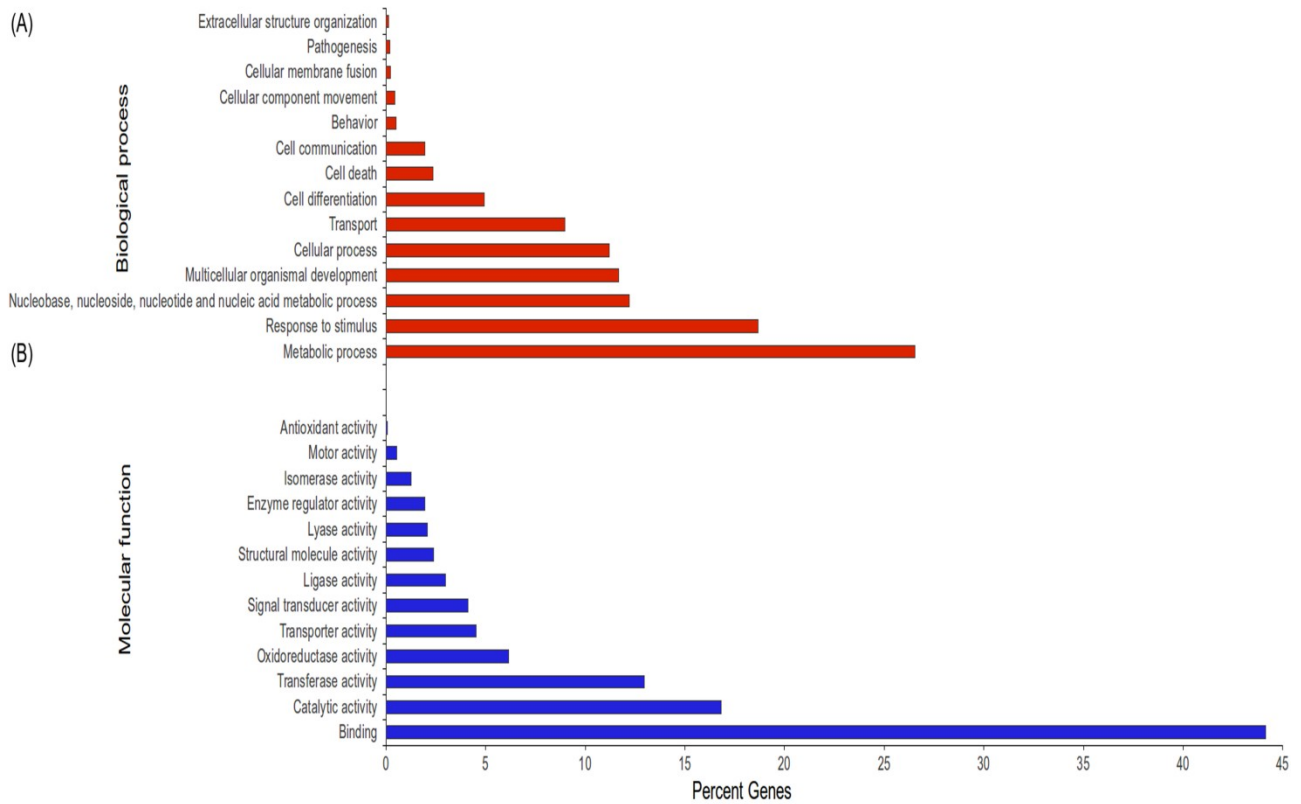
Reads mapping to unigenes

Since this is a *de novo* assembling problem where reference of guide sequence is unavailable, alignment to gene at genomic structure was not possible, nor it would be reliable to align with some known genome of some related species. The quantification of genes abundance was measured by mapping/aligning the reads over the assembled unigene sequences, as per the well established protocol described previously. Significant differentially expressed unigenes were identified using edgeR tool which compares the RPKM values of unigene in two different conditions and statistically evaluates the significance change in gene expression (Supplementary Fig. S7).

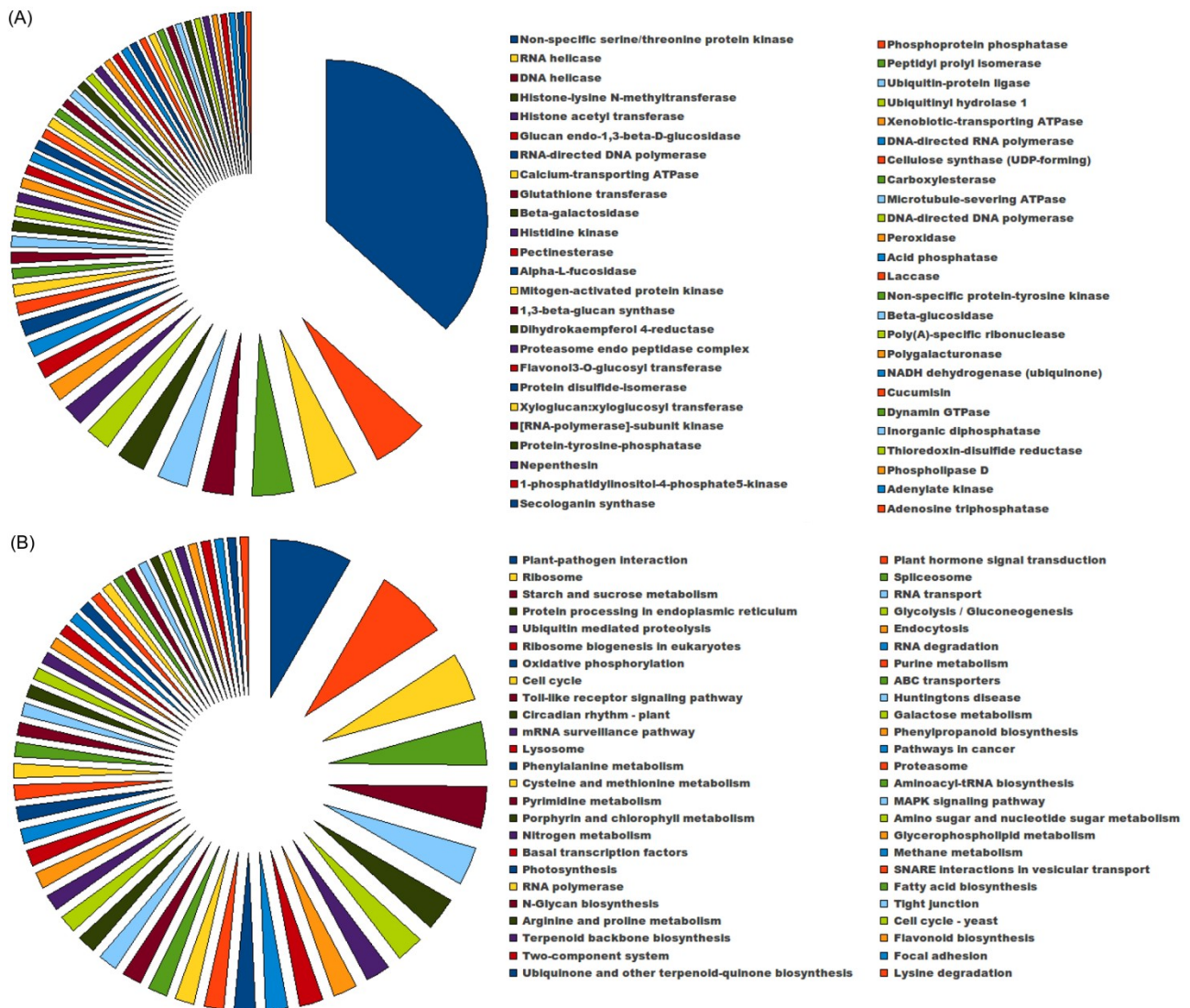
References:

1. Gahlan, P. *et al.* *De novo* sequencing and characterization of *Picrorhiza kurroa* transcriptome at two temperatures showed major transcriptome adjustments. *BMC Genomics* **13**, 126 (2012).
2. Thakur, K. *et al.* De-novo sequence assembly and transcriptome analysis of *Venturia inaequalis*, the deadly apple scab pathogen. *PLoS ONE* **8**, e53937 (2012).
3. Bhardwaj, J. Comprehensive Transcriptomic Study on horse gram (*Macrotyloma uniflorum*): De novo Assembly, Functional Characterization and Comparative Analysis in Relation to Drought Stress. *BMC Genomics* **14**, 647 (2013).
4. Mortazavi, A *et al.* Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621-628 (2008).
5. Li, W. *et al.* Sequence clustering strategies improve remote homology recognitions while reducing search times. *Protein Eng.* **15**, 643-649 (2002).
6. Li, R. *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265-272 (2010).
7. Huang, Y. *et al.* CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**, 680-682 (2010).

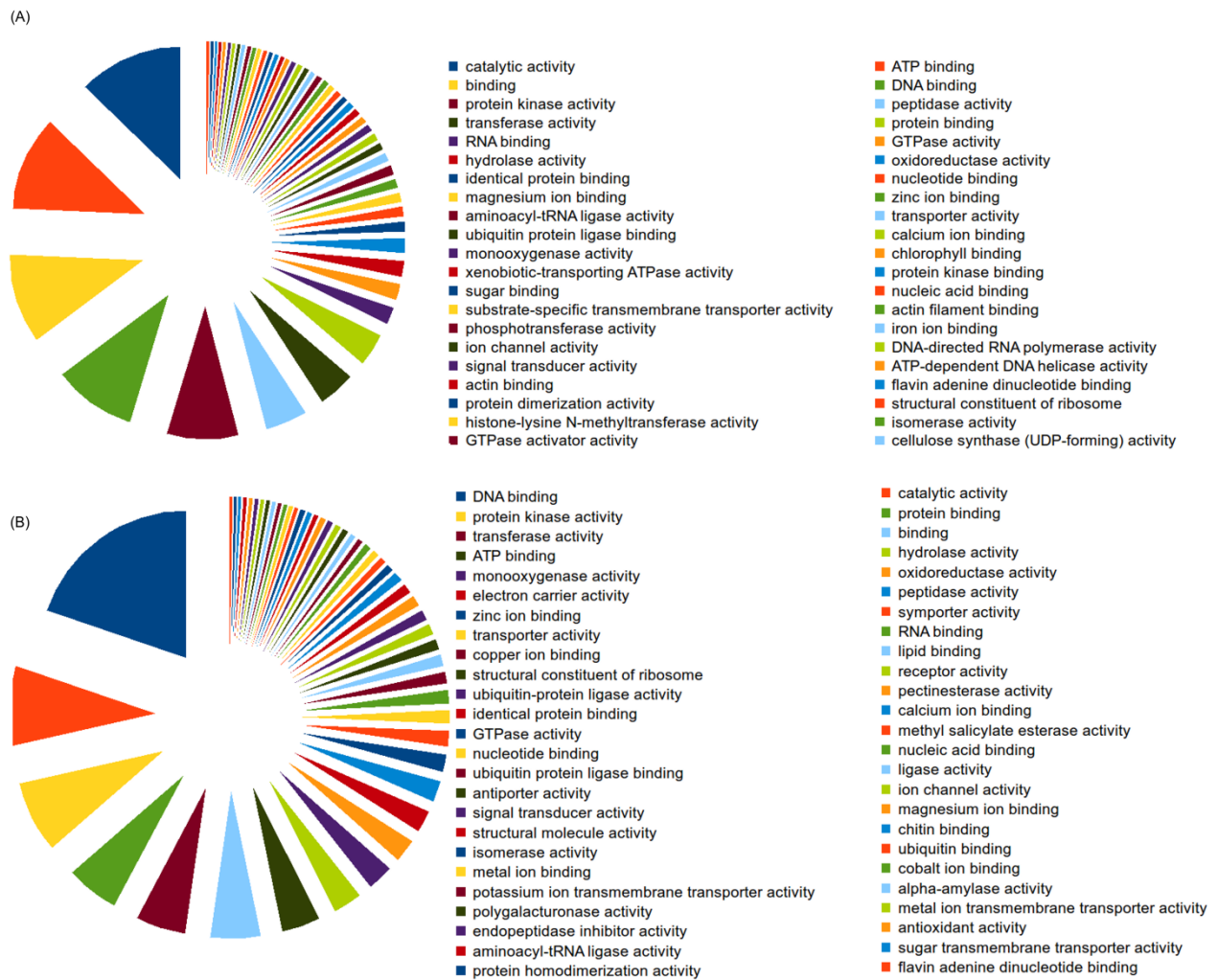
8. Pertea, G. *et al.* TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**, 651-652 (2002).



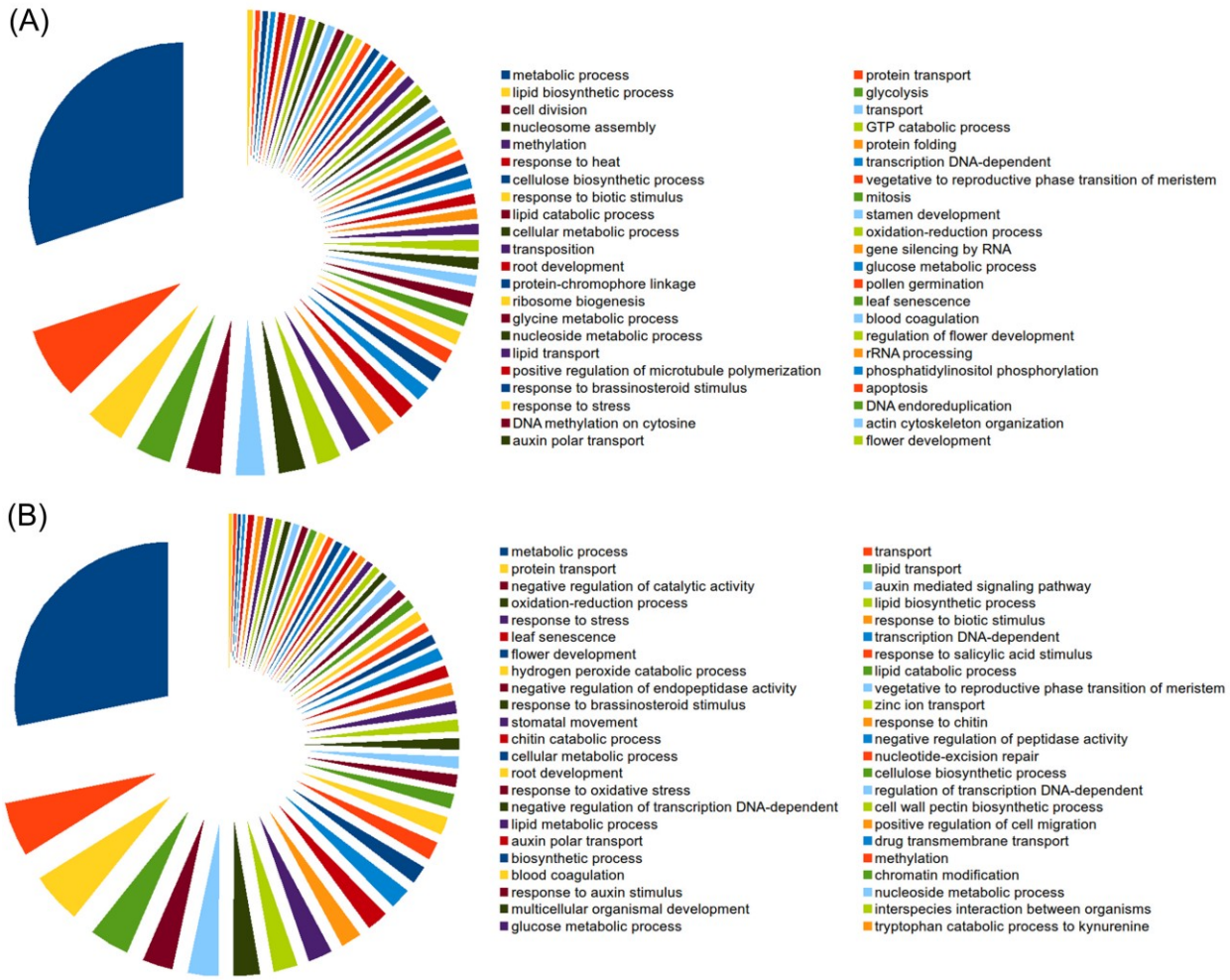
Supplementary Figure S1. Percent of the total unigenes belonging to (A) biological process, and (B) molecular function category of Gene Ontology (GO) classification. Supplementary Table S2 has details on GO classification for all the unigenes.



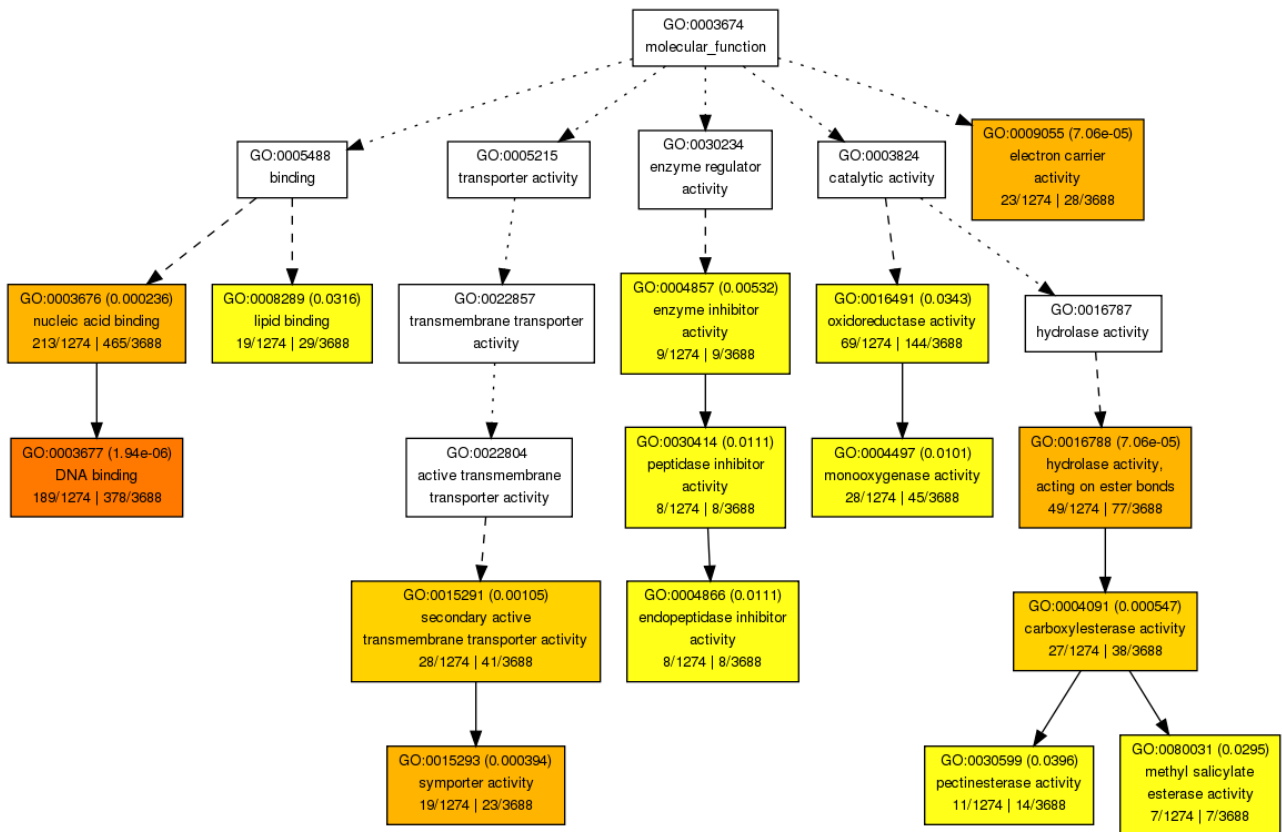
Supplementary Figure S2. Representation of (A) top 50 enzyme classes based on Enzyme Commission (EC) classification, and (B) top 50 pathways based on Kyoto Encyclopedia of Genes and Genomes (KEGG). Supplementary Table S3 has details for EC classification and KEGG pathways for all the unigenes.



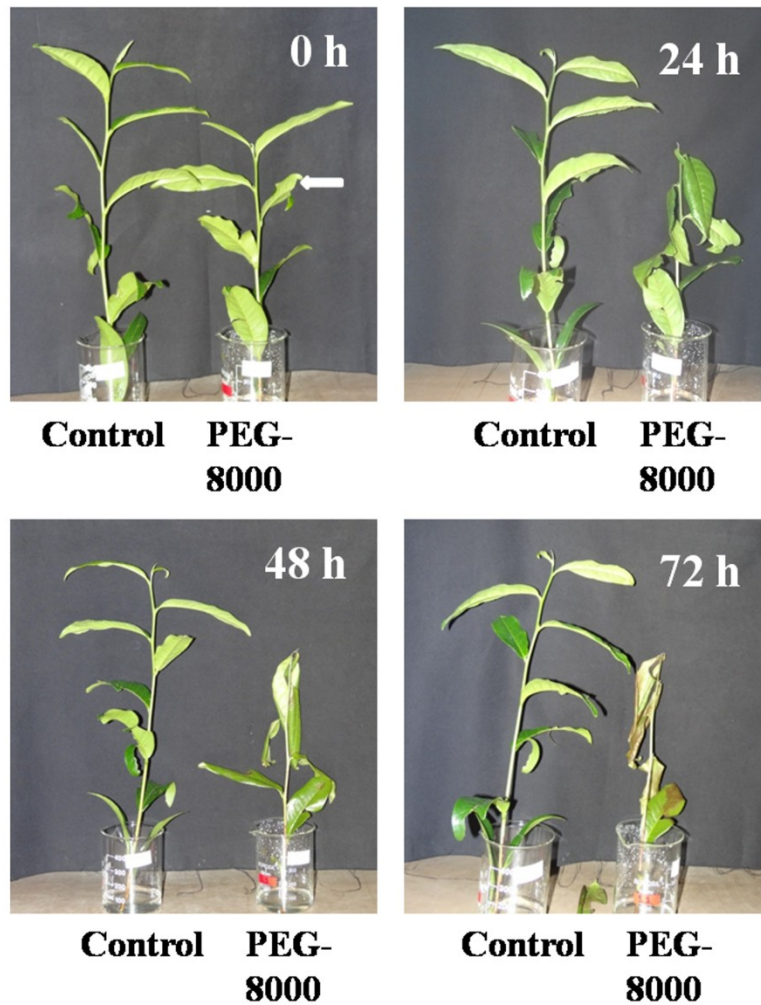
Supplementary Figure S3. Characterization of over-expressed unigenes during the period of active growth (A), and winter dormancy (B) based on molecular function.



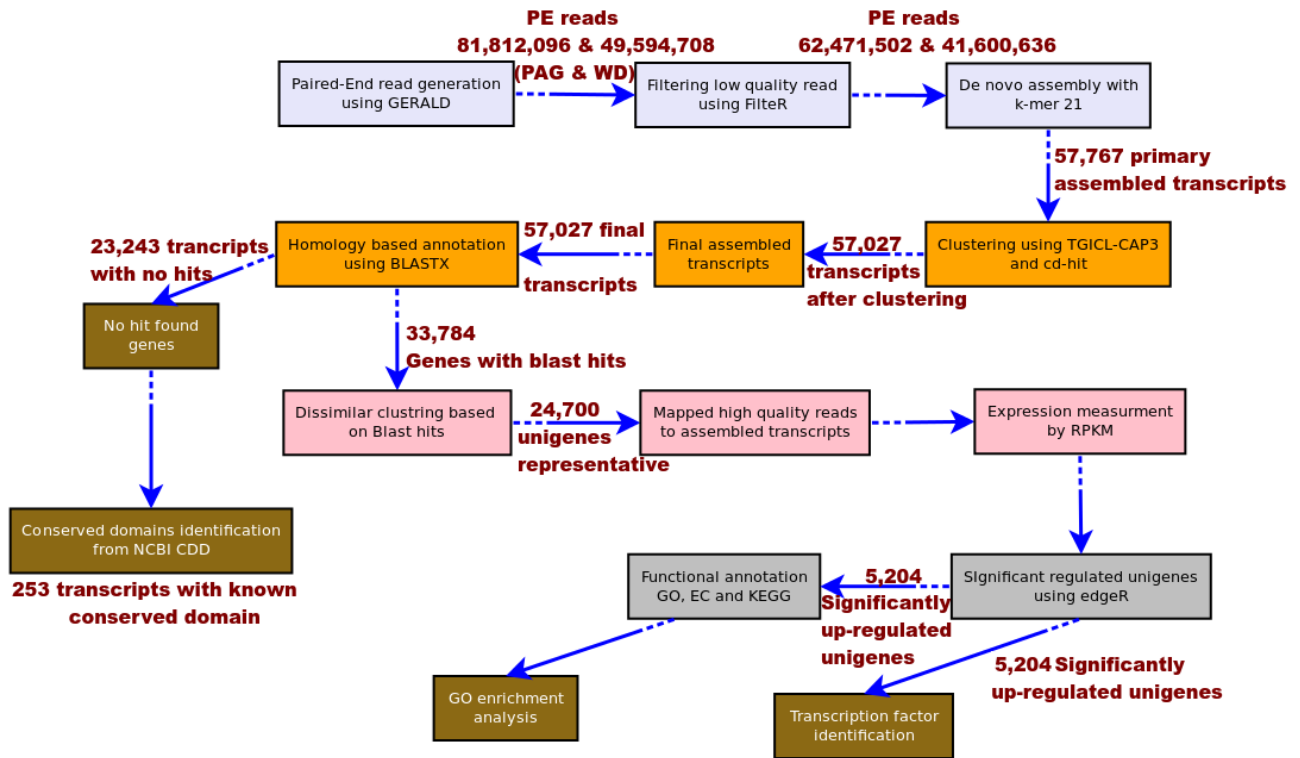
Supplementary Figure S4. Characterization of over-expressed unigenes during the period of active growth (A), and winter dormancy (B) based on biological process.



Supplementary Figure S5. Enrichment analysis of unigenes based on Gene Ontology (GO) molecular function category significantly over-expressed during the period of active growth as compared to winter dormancy. P-value is given in the bracket next to GO term in the box. Ratio given in the box is the ratio of unigenes belonging to the GO term to the total number of unigenes. First ratio is the number of unigenes in foreground and next ratio is number of unigenes in background/reference.



Supplementary Figure S6. Leaf-abscission model, wherein 10% polyethylene glycol-8000 (PEG-8000) induced abscission of tea leaves. Relatively mature leaves e.g. at position four and five (leaf position was with reference to apical bud at “0” position; the leaf adjacent to apical bud was designated to be at position 1) show the sign of abscission at 72 h and onwards. Tea leaves do not abscise in control (deionised water) in the designated time period. Fourth leaf (marked by arrow) was used for gene expression analysis and relative electrolyte leakage measurements.



Supplementary Figure S7. Detailed work-flow of developed *de novo* transcriptome assembly protocol used for analysis of *Camellia sinensis*.

Supplementary Table S4. Transcription factor families over-expressed during the period of active growth (PAG) and winter dormancy (WD).

Over-expressed during PAG	Median in PAG	Median in WD	Fold enrichment PAG/WD
CPP	3.89	0.24	16.2871463187
E2F-DP	3.08	0.31	9.8673606506
SNF2	2.17	0.39	5.5749497609
HMG	2.32	0.43	5.378945631
ARID	1.81	0.55	3.2661480962
FHA	1.75	0.57	3.0534057212
SET	1.7	0.58	2.9326575431
DDT	1.5	0.55	2.7477738115
TAZ	1.48	0.68	2.170132325
ARF	1.42	0.66	2.1627985104
PHD	1.45	0.68	2.1343911585
SWI/SNF-BAF60b	1.39	0.72	1.9411000933
GNAT	1.33	0.75	1.7655585963
Sigma70-like	1.31	0.76	1.7275243031
MADS	1.3	0.77	1.6788357034

Over-expressed during WD	Median in PAG	Median in WD	Fold enrichment WD/PAG
LFY	0.03	31.21	974.01106463
SRS	0.2	4.77	23.7423335004
LOB	0.45	2.24	5.0191383147
AP2-EREBP	0.46	2.15	4.6928735811
ARR-B	0.51	1.97	3.8666031994
C2C2-Dof	0.53	1.88	3.5238817745
ULT	0.6	1.92	3.1759302643
OFP	0.59	1.71	2.90951199
DBP	0.61	1.64	2.675443802
SBP	0.65	1.54	2.375151721
GRAS	0.67	1.5	2.2454105062
TIG	0.68	1.51	2.2120115171
WRKY	0.66	1.41	2.1375896286
CSD	0.7	1.44	2.0688033867
BES1	0.71	1.4	1.9710226136
bZIP	0.75	1.34	1.7847746005

Supplementary Table S6. Expanded form of abbreviations used in Figure 2-6.

<p>Figure 2</p>
<p>AP2-EREBP, APETALA2-ethylene-responsive element binding proteins; bHLH, Basic helix-loop-helix; bZIP, Basic region/leucine zipper motif; C2C2-Dof, DNA-binding with one finger; C2H2, Cysteine-2/Histidine-2 zinc finger protein; DBP, DNA-binding protein phosphatase; FAR1, Far-red impaired response1; G2-like, Golden 2-like; GRAS, GAI-RGA-SC; HB, Homeobox; MYB, Myeloblastosis; NAC, NAM, ATAF1,2, CUC2; Orphans, Orphans transcription factor; SBP, Squamosa promoter binding protein; WRKY, WRKY transcription factor; OFP, Ovate family proteins; SRS, SHI related sequence; BES1, Bri1-EMS-suppressor 1; LOB, Lateral organ boundaries; CAMTA, Calmodulin binding transcription activator</p>
<p>Figure 3</p>
<p>Name of each gene starts with a prefix <i>Cs</i> which stand for <i>Camellia sinensis</i>. Full name of the genes follow: <i>CsPdr</i>, <i>PDR-like ABC transporter</i>; <i>CsGalE</i>, <i>UDPglucose 4-epimerase</i>; <i>CsCullin</i>, <i>Cullin</i>; <i>CsTef</i>, <i>Transcription elongation factor</i>; <i>CsAqp</i>, <i>Aquaporin</i>; <i>CsPhot</i>, <i>Phototropin</i>; <i>CsPetrp</i>, <i>Ribosomal protein PETRP-like</i>; <i>CsMybR2R3</i>, <i>R2R3 Myb transcription factor</i>; <i>CsCop1</i>, <i>Photoregulatory zinc-finger protein COP1</i>; <i>CsGlip</i>, <i>GDSL-motif lipase/hydrolase family protein</i>; <i>CsR3H</i>, <i>Single-stranded nucleic acid binding R3H</i>; <i>CsPp2C</i>, <i>Protein phosphatase 2C</i>; <i>CsDnmt</i>, <i>DNA cytosine methyltransferase</i>; <i>CsG-protein</i>, <i>Small GTP-binding protein</i>; <i>CsCeh</i>, <i>Carboxylic ester hydrolase</i>; <i>CsLea3</i>, <i>Late embryogenesis abundant protein 3</i>; <i>CsLea4</i>, <i>Late embryogenesis abundant protein 4</i>; <i>CsZfp</i>, <i>Zinc finger protein</i>; <i>CsDhn</i>, <i>Dehydrin</i>.</p>
<p>Figure 4, 5, 6C</p>
<p>Name of each gene starts with a prefix <i>Cs</i> which stand for <i>Camellia sinensis</i>. Full name of the genes follow: <i>Cre1</i>, <i>Cytokinin receptor1</i>; <i>Arf5</i>, <i>Auxin response factor5</i>; <i>Pin1</i>, <i>Auxin hydrogen transporter</i>; <i>Pin2</i>, <i>Auxin hydrogen symporter</i>; <i>Erf2</i>, <i>Ethylene response factor 2</i>; <i>Jomt</i>, <i>Jasmonate o-methyltransferase</i>; <i>Cel2</i>, <i>Cellulase 2</i>; <i>Pg</i>, <i>Polygalacturonase</i>; <i>Ga2-ox1</i>, <i>Gibberellin 2-oxidase 1</i>; <i>Pgip1</i>, <i>Polygalacturonase inhibiting protein1</i>; <i>Pgi1</i>, <i>Polygalacturonase inhibitor1</i>; <i>Pgi2</i>, <i>Polygalacturonase inhibitor2</i>.</p>