Rothamsted Repository Download

A - Papers appearing in refereed journals


The publisher's version can be accessed at:

- [https://dx.doi.org/10.1038/s41597-019-0111-9](https://dx.doi.org/10.1038/s41597-019-0111-9)

The output can be accessed at:


© 22 July 2019, Please contact library@rothamsted.ac.uk for copyright queries.
Transcriptome sequencing, molecular markers, and transcription factor discovery of *Platanus acerifolia* in the presence of *Corythucha ciliata*

Fengqi Li1, Chunyan Wu2, Mengzhu Gao3, Mengmeng Jiao1, Cheng Qu1, Asier Gonzalez-Uriarte3 & Chen Luo1

The London Planetree (*Platanus acerifolia*) are present throughout the world. The tree is considered a greening plant and is commonly planted in streets, parks, and courtyards. The Sycamore lace bug (*Corythucha ciliata*) is a serious pest of this tree. To determine the molecular mechanism behind the interaction between the London Planetree and the Sycamore lace bug, we generated a comprehensive RNA-seq dataset (630,835,762 clean reads) for *P. acerifolia* by sequencing both infected and non-infected leaves of *C. ciliata* using the Illumina Hiseq 4000 system. We assembled the transcriptomes using the Trinity De Novo assembly followed by annotation. In total, 121,136 unigenes were obtained, and 80,559 unigenes were successfully annotated. From the 121,136 unigenes, we identified 3,010,256 SNPs, 39,097 microsatellites locus, and 1,916 transcription factors. The transcriptomic dataset we present are the first reports of transcriptome information in *Platanus* species and will be incredibly useful in future studies with *P. acerifolia* and other *Platanus* species, especially in the areas of genomics, molecular biology, physiology, and population genetics.

Background & Summary

Transcriptional sequencing technology is used in biological research for the gene expression profile investigation, the biological molecular evolution, and molecular marker acquisition1–4. The technology is particularly convenient for non-model organisms, for which there is no genome data available5–6. Abundant transcriptome data of some garden trees are reported as the demand for continuous development of urban landscaping7–9.

The London Planetree (*Platanus acerifolia*) is a hybrid cross between the American sycamore (*P. occidentalis*) and the Oriental Planetree (*P. orientalis*)10. *P. acerifolia* is a woody arbor plant with a large crown that grows rapidly, provides dense shade, and is tolerant to urban pollution11. This species is commonly grown around the world and is known as “the king of street trees”12. Despite its widespread use, there is a lack of research regarding the molecular biology of the tree, and there are no publicly available genome or transcriptome resources for the species or the genus. For this reason, research on genetic diversity and work on genetic engineering using molecular biotechnology is limited.

A particularly harmful pest to *P. acerifolia* is the sycamore lace bug (*Corythucha ciliata*), which is native to North America but was introduced to Europe in the 1960s13. The bug was first found in Hunan province in China in 2002 and has since spread to Hubei, Shanghai, Shandong, Henan, and Beijing, where heavy infestations have been reported14,15. The sycamore lace bug specifically damages *Platanus* trees, causing chlorotic or bronzed foliage and premature senescence of leaves16. Currently, transcriptome resources are not available for the genus *Platanus*,

1Institute of Plant and Environment Protection, Beijing Academy of Agriculture and Forestry Sciences, Beijing, 100097, China. 2College of Life Sciences, Capital Normal University, Beijing, 100037, China. 3Rothamsted Research, Computational and Analytical Sciences Department, Harpenden, ALS 2JQ, UK. These authors contributed equally: Fengqi Li and Chunyan Wu. Correspondence and requests for materials should be addressed to A.G.-U. (email: asier.gonzalez@rothamsted.ac.uk) or C.L. (email: luochen1010@126.com)
Ciliata, which were raised according to previous research. The experiments were performed in a growth chamber (25 ± 2°C, 50–70% RH, 16:8 L:D). The insects on the leaves were treated for 24 h, 48 h and removed with a soft brush. Control leaves (control) were grown as the others but without Platanus acerifolia infestation. After treatment, each plant leaf sample was collected for RNA extraction. Each treatment was performed in three biological replicates.

RNA isolation, cDNA library, and illumina sequencing. Total RNA was extracted using the TRIzol reagent (Invitrogen, CA, USA). The integrity and the purity of total RNA were verified using an Agilent Bioanalyzer 2100 and RNA 6000 Nano LabChip Kit (Agilent Technologies, CA, USA) with a minimum RNA integrity number of 7. Approximately 10 μg of the total RNA representing a specific adipose type was subjected to isolate Poly (A) mRNA with poly-T oligo-attached magnetic beads (Invitrogen, CA, USA). After purification, the poly(A)− or the poly(A)+ RNA fractions were fragmented into small pieces using divalent cations under elevated temperatures. The cleaved RNA fragments were reverse-transcribed to create the final cDNA library in accordance with the protocol for the mRNA-Seq sample preparation kit (Illumina, San Diego, USA). The average insert size for the paired-end libraries was 300 bp (±50 bp). The paired-end sequencing was performed on an Illumina Hiseq 4000 following the vendor’s recommended protocol.

De Novo assembly, unigene annotation, and functional classification. Fastp was used to remove the readings that contained adapter contamination, low quality bases, and undetermined bases. The sequence quality was verified via FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), including the Q20, the Q30, and the GC-content of the clean data. The downstream analyses were based on high-quality clean data. De Novo assembly of the transcriptome was performed with Trinity 2.4.0. Next, TransRate and BUSCO were used to assess De Novo transcriptome assembly quality. The assembled unigenes were aligned against the Nr protein (http://www.ncbi.nlm.nih.gov/), Pfam, COG, and the SwissProt (http://www.expasy.ch/sprot/) databases using BLASTx with an E-value threshold of <0.00001. The gene ontology (GO) annotations were obtained using Blast2GO (http://www.blast2go.com/b2ghome). Metabolic pathway analysis was performed using the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/).

SNPs, SSRs, and transcription factor identification. SAMtools package was used to detect potential SNPs. SNPs were filtered based on the following criteria: (1) the number of reads to cover a candidate SNP above 8; (2) remove low quality where base calls with low Phred quality below 25; (3) frequency of mutated bases among all reads covering the position above 30%. For all unigenes, SSRs were identified using MISA. This data descriptor provides an opportunity to identify the functional genes and molecular marker for Platanus acerifolia. This comprehensive Platanus acerifolia transcriptomic information can be utilized to promote the insect defense mechanisms in Platanus acerifolia.

Table 1. Characteristics of the Platanus acerifolia transcriptome sequencing project.
by Primer3 (http://primer3.sourceforge.net/releases.php)26. The transcription factor families were identified using the Plant Transcription Factor Database PlantTFDB 4.0 (http://planttfdb.cbi.pku.edu.cn/prediction.php)27.

Data Records
The annotation, molecular markers, and transcription factor output files were provided in Figshare28. Raw FASTQ files for the RNA-Seq were deposited to the NCBI SRA database under SRA accession number SRP15664029. The final assembled unigenes sequences were deposited at NCBI GenBank (GXXZ00000000.2)30.

Technical Validation
High throughput sequencing generated 46,890,842–57,342,752 pairs of raw reads per sample 29, and the Q20 scores (the average quality value) were greater than 97%. The GC content of clean reads was similar, ranging from 46.14% to 47.36% (Online-only Table 1). The total length of the combined reads for the 12 samples that represented the different stages of damage was 202,095,905 bp and 121,136 unigenes 28; the average length was 1015.15 bp with an N50 of 1579 bp and an E90N50 of 1762 bp (Table 2).

All 121,136 unigenes found in P. acerifolia leaves were functionally annotated using six public databases (Table 3). Of unigenes, 62.91% (76,203) were annotated to the NR database, 43.55% (52,758) were annotated to proteins in the Swiss-Prot database, 40.06% (48,527) were annotated to proteins in the Pfam database, 7.31% (8,849) were annotated to the COG database, 47.88% (57,997) were annotated to the GO database, and 28.23% (34,193) were annotated to the nucleotide sequences in the KEGG database. In total, 66.5% of unigenes (80,559) were annotated to a database.

The similarity analysis of the NR database demonstrated that there were 39,436 unigenes with significant homology (E-values < 1e−30) to other sequences in the Nr database and 36,767 unigenes with E-values between 1e−5 and 1e−30. The NR annotation species distribution analysis showed that 22,670 unigenes had higher homology with nelumbo_nucifera, which accounted for 29.94% of the total (Fig. 1)28. In addition, Swiss-Prot and Pfam annotation results were deposited in Swiss-prot_annotation.xls and Pfam_annotation.xls, respectively28.

After COG based annotation, a total of 8,849 unigenes were assigned to 24 functional categories (Fig. 2). For COG annotation, the two largest COG categories were "Translation, ribosomal structure, and biogenesis" (803, 16.85%) and "Posttranslational modification, protein turnover, chaperones" (550, 11.54%). The following abundant groups were "General function prediction only" (457, 9.59%), "Energy production and conversion" (324, 6.80%), "Carbohydrate transport and metabolism" (300, 6.30%), "Signal transduction mechanisms" (265, 5.56%), and "Amino acid transport and metabolism" (262, 5.50%). The two groups involving "Cell motility" (7, 0.147%) and "Nuclear structure" (3, 0.063%) represented the smallest COG classifications. Lastly, 43 unigenes (0.902%) were classified into "Defense mechanisms".

A total of 57,997 unigenes were annotated in the GO database, 53.14% (29,079) for the biological process, 58.80% (49,763) for the molecular function, and 56.13% (32,553) for the cellular component. The categories

<table>
<thead>
<tr>
<th>Type</th>
<th>Resource</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total transcripts number</td>
<td>199,080</td>
<td></td>
</tr>
<tr>
<td>Total unigenes number</td>
<td>121,136</td>
<td></td>
</tr>
<tr>
<td>Total sequence base</td>
<td>202,095,905</td>
<td></td>
</tr>
<tr>
<td>Largest</td>
<td>18,931</td>
<td></td>
</tr>
<tr>
<td>Smallest</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>Average length</td>
<td>1015.15</td>
<td></td>
</tr>
<tr>
<td>N50</td>
<td>1579</td>
<td></td>
</tr>
<tr>
<td>E90N50</td>
<td>1762</td>
<td></td>
</tr>
<tr>
<td>GC percent</td>
<td>41.50</td>
<td></td>
</tr>
<tr>
<td>Mean mapped reads</td>
<td>1442.7636918</td>
<td></td>
</tr>
<tr>
<td>TransRate score</td>
<td>0.25922</td>
<td></td>
</tr>
<tr>
<td>BUSCO score</td>
<td>72.1% (3.3%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Assembly information of the Platanus acerifolia transcriptome dataset.

<table>
<thead>
<tr>
<th>Unigene number (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NR</td>
</tr>
<tr>
<td>Swiss-Prot</td>
</tr>
<tr>
<td>Pfam</td>
</tr>
<tr>
<td>COG</td>
</tr>
<tr>
<td>GO</td>
</tr>
<tr>
<td>KEGG</td>
</tr>
<tr>
<td>Total_annotation</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

Table 3. Annotation information of the Platanus acerifolia transcriptome dataset.
“cellular process,” “metabolic process,” and “single-organism process” were most abundant among the biological process GO category. Within the cellular component category, the “cell” and “cell part” terms were most abundant. For the molecular function, the unigenes were chiefly related to “binding” and “catalytic activity” (Fig. 3)28.

We mapped the unigenes to the reference authoritative pathway in KEGG for further functional classification and annotation. In total, 34,193 unigenes were distributed among 130 KEGG pathways, and 11,229 (32.84%) were related to metabolic pathways. The largest number of unigenes involved were in the “Carbohydrate metabolism” (2741) category, followed by the “Amino acid metabolism” (1771) category, whereas “Glycan biosynthesis and metabolism” (309) was the smallest group (Fig. 4 and kegg_annotation.xls)28.

We screened the P. acerifolia unigene dataset to determine potential SNPs and SSRs for future populations and genetics analysis. Among unigenes sequences, we detected 28,144 unigenes containing SSRs and 6,053 unigenes containing more than one SSR. According to the repeat motif, the SSR loci can be divided into six categories: mono-nucleotide repeats (21,895), di-nucleotide repeats (11,388), tri-nucleotide repeats (5,353), tetra-nucleotide repeats (373), penta-nucleotide repeats (55), and hexa-nucleotide repeats (33) (Fig. 5, ssr_repeats.xls, ssr_analysis_details.xls)28.

A total of 3,010,256 SNPs were obtained from the twelve leaves samples. Among these SNPs, 1,503,269 and 1,506,987 SNPs were obtained from the CK and insect treated samples, respectively. And, 1,005,449 SNPs were homo-type, 2,004,807 were hete-type (snp_homo_hete_statistics.xls, snp_detail.xls)28. Among them, 1,349,858 were putative transitions, and 791,734 were putative transversions. The transition-type SNPs include four classes (A/G, C/T, G/A, and T/C) and the transversion-type SNPs include eight classes (A/C, A/T, C/A, C/G, G/C, G/T, T/A, and T/G). (snp_transition_tranversion_statistics.xls, snp_detail.xls)28.

In order to promote functional gene research in P. acerifolia, we identified a series of transcription factors, which included 35 gene families. Among them, MYB_superfamily had as many as 311 unigenes, and C2C2 and

Fig. 1 Species distribution of the NR annotation.

Fig. 2 The COG functional categories.
AP2/ERF had 168 and 166 unigenes, respectively. NAC had 132 unigenes, bHLH had 122 unigenes, and both WRKY had 107 unigenes (Fig. 6, Transcription_Factor_annotation.xls)

The comprehensive datasets we present are the first reports of transcriptome information in Platanus species and will facilitate the identification of insect defense-related genes in the future. The annotated unigenes are a significant improvement on the sequence information available for P. acerifolia and other closely related species. The identified SNPs and SSR locus resources will be of help in population genetic structure, gene flow studies, and parentage analysis for P. acerifolia. The reported transcription factors in this dataset will be useful resources to further explore the physiological and biochemical mechanisms of growth development and stress response in P. acerifolia and other Platanus species.

**Code Availability**

The following software were used for data analysis:

1. Fastp was used for preprocessing for FastQ files. https://github.com/OpenGene/fastp.
3. Trinity 2.4.0 was used to de novo transcriptome assembly. https://github.com/trinityrnaseq/trinityrnaseq.
5. Blast2GO was used for GO annotation. http://www.blast2go.com/b2ghome.
6. KEGG database was used for metabolic pathway annotation. http://www.genome.jp/kegg/.
7. SAMtools was used for detecting SNPs. https://samtools.github.io/bcftools/howtos/variant-calling.html.
8. MISA was used to identify SSRs. http://pgrc.ipk-gatersleben.de/misa/misa.html.

References