Transcriptome analysis of *Polygonum minus* reveals candidate genes involved in important secondary metabolic pathways of phenylpropanoids and flavonoids

Kok-Keong Loke¹, Reyhaneh Rahnamaie-Tajadod¹, Chean-Chean Yeoh², Hoe-Han Goh¹, Zeti-Azura Mohamed-Hussein¹,², Zamri Zainal¹,², Ismanizan Ismail¹,² and Normah Mohd Noor¹

¹ Institute of Systems Biology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
² School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia

**ABSTRACT**

**Background:** *Polygonum minus* is an herbal plant in the Polygonaceae family which is rich in ethnomedicinal plants. The chemical composition and characteristic pungent fragrance of *Polygonum minus* have been extensively studied due to its culinary and medicinal properties. There are only a few transcriptome sequences available for species from this important family of medicinal plants. The limited genetic information from the public expressed sequences tag (EST) library hinders further study on molecular mechanisms underlying secondary metabolite production.

**Methods:** In this study, we performed a hybrid assembly of 454 and Illumina sequencing reads from *Polygonum minus* root and leaf tissues, respectively, to generate a combined transcriptome library as a reference.

**Results:** A total of 34.37 million filtered and normalized reads were assembled into 188,735 transcripts with a total length of 136.67 Mbp. We performed a similarity search against all the publicly available genome sequences and found similarity matches for 163,200 (86.5%) of *Polygonum minus* transcripts, largely from *Arabidopsis thaliana* (58.9%). Transcript abundance in the leaf and root tissues were estimated and validated through RT-qPCR of seven selected transcripts involved in the biosynthesis of phenylpropanoids and flavonoids. All the transcripts were annotated against KEGG pathways to profile transcripts related to the biosynthesis of secondary metabolites.

**Discussion:** This comprehensive transcriptome profile will serve as a useful sequence resource for molecular genetics and evolutionary research on secondary metabolite biosynthesis in Polygonaceae family. Transcriptome assembly of *Polygonum minus* can be accessed at [http://prims.researchfrontier.org/index.php/dataset/transcriptome](http://prims.researchfrontier.org/index.php/dataset/transcriptome).

**Subjects**  Bioinformatics, Biotechnology, Genomics

**Keywords**  De novo assembly, Hybrid assembly, Illumina sequencing, RNA-seq, *Persicaria minor*, 454 sequencing

How to cite this article  Loke et al. (2017), Transcriptome analysis of *Polygonum minus* reveals candidate genes involved in important secondary metabolic pathways of phenylpropanoids and flavonoids. PeerJ 5:e2938; DOI 10.7717/peerj.2938
INTRODUCTION

Secondary metabolites are organic compounds that are non-vital but indirectly influence plant survival, development and growth. Three major groups of plant secondary metabolites identified by chemical groups are flavonoids and phenolic compounds, terpenoids, and nitrogen/sulfur-containing compounds. Plant secondary metabolites are important natural sources for the development of medicines and natural products. The myriads of plant secondary metabolites reflect the diverse species of plants and their ecological roles, such as adaptation to different environments and defense against biotic stresses (Moore et al., 2014). Polygonum is a genus in the Polygonaceae family with up to 300 species, many of which are important as traditional medicinal plants (Narasimhulu, Reddy & Mohamed, 2014).

Polygonum minus Huds. (syn. Persicaria minor) is a culinary flavoring ingredient common in South East Asia and is also used as a remedy for different maladies ranging from indigestion to poor eyesight (Christopher et al., 2015; George et al., 2014). The leaves of Polygonum minus contain high levels of essential oils (72.54%), mainly comprised of aliphatic aldehydes, namely dodecanal (48.18%) and decanal (24.36%) (Yaacob, 1990). Less abundant aldehydes include 1-decanol, 1-dodecanol, undecanal, tetradecanal, 1-undecanol, nonanal, and 1-nonanol (Baharum et al., 2010). Furthermore, the metabolite profiling of Polygonum minus leaf revealed many terpenoids and flavonoids with antioxidant activities (Baharum et al., 2010; Goh et al., 2016). The abundance of secondary metabolites in Polygonum minus has led to the establishment of hairy root system for the production of plant secondary metabolites (Ashraf et al., 2014). β-caryophyllene was found to be the main sesquiterpenes secreted into the hairy root culture media. These studies showed the potential of developing Polygonum minus as a resource to produce natural products.

While many secondary metabolites have been identified in Polygonum minus, its biosynthetic pathways remain unclear due to the limited genomic information that is available for the plant. Previously, a total of 3,352 expressed sequence tags (ESTs) were generated from standard cDNA libraries of Polygonum minus leaf, root and stem tissues (Roslan et al., 2012). This study indicated the abundance of flavonoid biosynthesis-related genes in the root tissue. The emergence of next generation sequencing has made transcriptomic analysis of plant possible with increasing speed and affordability. RNA-sequencing (RNA-seq) allows novel gene discovery and identification of transcripts of interest in various biological processes. This is especially suitable for many non-model organisms with limited genomic information (Varshney et al., 2009; Ward, Ponnala & Weber, 2012). In the past few years, this platform has been repeatedly utilized to discover and identify genes involved in the biosynthesis of secondary metabolites. For examples, alkaloid biosynthesis in Uncaria rhynchophylla (Guo et al., 2014), ginsenoside biosynthesis in Panax ginseng (Jayakodi et al., 2015), glucosinolate biosynthesis in Raphanus sativus (Wang et al., 2013), biosynthesis of capsaicinoids in Capsicum frutescens (Liu et al., 2013), biosynthesis of flavonoids in safflower (Li et al., 2012), and caffeine biosynthesis in Camellia sinensis (Shi et al., 2011). To date, there are only two plants from the Polygonum
genus with RNA-seq data deposited to the public SRA database, namely *Polygonum cuspidatum* (Hao et al., 2012) and *Polygonum tinctorium* (Minami, Sarangi & Thul, 2015). The limited genetic information from this important family of medicinal plants hinders further study on molecular mechanisms underlying the production of bioactive compounds.

To profile transcripts related to the biosynthesis of secondary metabolites in *Polygonum minus*, RNA-seq was performed on the leaf and root tissues. Sequence data generated from 454 and Illumina platforms were assembled, both independently and together, for comparison. This new dataset was compared to all *Polygonum minus* EST transcripts previously deposited to the NCBI database (dated Sep 2014) for validation of the assembly quality. The combined de novo assembly from two different sequencing platforms allowed us to overcome limitations of each technology. We also performed KEGG pathway annotation to identify transcripts related to the biosynthesis of secondary metabolites. This study reveals candidate genes involved in the biosynthesis of secondary metabolites, especially on the biosynthesis of phenylpropanoids and flavonoids in *Polygonum minus* and serves as an invaluable genetic resource for its development as a commercial herbal crop.

**MATERIALS AND METHODS**

**Sample preparation and transcriptome sequencing**

Root and leaf tissues of cultivated *Polygonum minus* grown in compost soil without fertilizer were sampled independently from the experimental plot (3°16’14.63”N, 101°41’11.32”E) at Universiti Kebangsaan Malaysia. For the leaf tissue, five expanded young leaves from the apical parts of the plants were collected and pooled as one biological replicate. Samples acquired from 45 day old plants were rinsed with distilled water and flash frozen in liquid nitrogen before stored at −80 °C. Total RNA was isolated using the Lopez-Gomez method with modifications (López-Gómez & Gómez-Lim, 1992) by adding 50% PVP-40 due to high polysaccharide and phenolic compounds in *Polygonum minus*. RNA quality and quantity were assessed using gel electrophoresis, ND-1000 Nanodrop spectrophotometer (Thermo Scientific) and Agilent 2100 Bioanalyzer with a minimum RNA integrity number of 7.

For the root sample, 250 ng of poly(A) RNA was prepared from 800 ng of total RNA using PolyATtract mRNA isolation kit (Promega, Madison, WI, USA) and used as a starting material for the Roche GS FLX sequencing at Malaysia Genome Institute. The cDNA preparation was done according to the cDNA Rapid Library Preparation Method Manual of Roche. The emulsion polymerase chain reaction (PCR) condition was performed using long fragment Lib-A emPCR amplification condition for amplicons that are 550 bp or greater. The conditions are as follows: 94 °C for 4 min, 50 cycles of 94 °C for 30 s and 60 °C for 10 min.

For the leaf sample, total RNA from two biological replicates were used for the Illumina HiSeq™ 2000 sequencing with an average read length of 90 bp through the standard library (200 bp) preparation and paired-end sequencing workflow established at BGI-Shenzhen, China.
Transcriptome de novo assembly

Raw reads from both 454 pyrosequencing platform and Illumina HiSeq™ 2000 were filtered to remove adapter sequences with sequence pre-processing tools, Cutadapt (Martin, 2011) and Trimmomatic (Bolger, Lohse & Usadel, 2014), respectively. High quality Illumina raw reads with Phred score ≥ 25 were kept for assembly. For root transcriptome, iAssembler pipeline (Zheng et al., 2011), which includes MIRA (Chevreux et al., 2004) and CAP3 (Huang & Madan, 1999) assemblers, was executed with the filtered dataset. The analysis pipeline includes three consecutive runs of MIRA with default parameters followed by CAP3 assembly to obtain the final assembled transcripts. Trinity pipeline (Grabherr et al., 2011) was used to assemble the leaf transcriptome from two leaf libraries, whereas the combined transcriptome was assembled from two leaf libraries and one simulated root library. Assembled reads from the root transcriptome were clipped to 90 bp pseudo reads with 5 bp overlap using an in-house PHP script (http://gitlab.inbiosis.ws/open-source/rnaseq-utils) to simulate Illumina sequencing output for accommodating Trinity assembler short read requirement. For this assembly, leaf raw read datasets were normalized with digital normalization following Khmer 1.0 mRNASeq protocol (Brown et al., 2014).

This project was registered at NCBI’s BioProject with the accession number PRJNA208436. All raw read datasets were deposited to NCBI SRA database (http://www.ncbi.nlm.nih.gov/sra) with the accession number SRX669305 (leaf) and SRX313492 (root) (Loke et al., 2016). Assembled transcripts were deposited to NCBI TSA database (http://www.ncbi.nlm.nih.gov/genbank/tsa) with the accession number GCJZ00000000. The assembled transcripts with annotation can also be accessed at http://prims.researchfrontier.org/index.php/dataset/transcriptome.

Transcript functional annotation

An annotation pipeline from Trinotate was performed to annotate assembled transcripts (Grabherr et al., 2011). The Trinotate annotation pipeline includes several software packages such as BLASTX, BLASTP, PFAM search, SignalP, and RNAmmer that are essential in transcriptome functional annotation. All analyses were performed in parallel using assembled FASTA sequences.

Functional annotation for all transcripts was performed by running BLASTX similarity search against Trinotate Swiss-Prot protein database (September 2015) with E-value <1e−5 considered as significant hits. For the leaf and combined transcriptomes, Trinotate annotation reports were generated using the standard annotation pipeline (http://trinotate.github.io). Gene Ontology (GO) and Conserved Domain Database (CDD) were used to annotate the transcripts based on similarity. Translated peptides were generated using the Transdecoder program embedded in the Trinity assembly pipeline for protein-based analysis using Eukaryotic Orthologous Group (KOG) classification. All results were deposited into Trinotate-provided SQLite database template and a spreadsheet summary report was generated from Trinotate using BLASTX E-value cutoff of 1e−5.
KEGG pathway mapping and enrichment analysis

KEGG pathway mapping was performed by associating Enzyme Commission (EC) number from BLAST search results with UniProt database (http://www.uniprot.org/mapping, September, 2015). Metabolite pathway maps with custom color codes were generated using KEGG online Mapper API (http://www.genome.jp/kegg/tool/map_pathway2.html, updated April 1, 2015), with all associated EC number. Pathway enrichment analysis with hypergeometric test was performed by using “Annotate” and “Identify” subprograms in KOBAS version 2.0 with Benjamini–Hochberg correction (Xie et al., 2011). The leaf and root transcriptomes were compared against combined transcriptome as a background for enrichment analysis.

Transcript abundance estimation

To estimate the relative abundance of transcripts in the leaf and root transcriptomes, filtered 454 (original unclipped) and Illumina raw reads were aligned to the combined transcriptome assembly using RSEM (Li & Dewey, 2011). RSEM statistical model is based on Expectation–Maximization (EM) algorithm to compute maximum likelihood abundance estimates. Transcripts per kilobase million (TPM), which normalizes for transcript length first then sequencing depth, was used as an estimate for the relative expression level (based on proportion of mapped reads) of each transcript in the leaf and root tissues. The same approach was used to identify the list of transcripts present in leaf, root or both tissue types for KEGG pathway mapping by using RSEM estimated count value to determine the presence (TPM > 0) or absence (TPM = 0) of a transcript.

RT-qPCR of selected transcripts

DNase-treated (Ambion, Huntingdon, UK) RNA (1 μg) was reverse transcribed using iScript™ cDNA Synthesis Kit (Bio-Rad, Hercules, CA, USA) per manufacturer’s protocol. The expression of seven transcripts related to the phenylpropanoid and flavonoid biosynthetic pathways were selected from the transcriptome dataset and specific primer pairs (Table S1) were designed using PrimerBlast software. For RT-qPCR analysis, 1:20 dilution of cDNA was used as template in 20 μL volume and reactions were performed in the iQ™5 Real-Time PCR detection System (Bio-Rad, Hercules, CA, USA) using the iTaq Universal SYBR® Green SuperMix kit (Bio-Rad, Hercules, CA, USA). The amplification was executed with the following cycling program: 3 min at 95 °C, 40 cycles of 10 s at 95 °C, 30 s at 60 °C, and 30 s at 72 °C; and 0.06 s for plate reading at 65 °C followed by a melting curve analysis. Primer efficiencies were determined through standard curves of five cDNA dilution factors in triplicate. Calcium-Dependent Protein Kinase (CDPK) and Polyubiquitin (UBQ) were selected as references to normalize expression level. RT-qPCR was performed using three biological replicates, each with three technical replicates. Calculation was based on $2^{-\Delta\Delta Ct}$ method (Nolan, Hands & Bustin, 2006) for fold-change (FC). Correlation plot was generated in MS Excel.
RESULTS

Sequencing and assembly

We performed de novo assembly of *Polygonum minus* leaf and root transcriptomes using reads from two different sequencing platforms. Root transcriptome from 454 sequencing showed a typical length distribution ranging from 100 to 600 bp which peaked around 460 bp (Fig. S1). It yielded the lowest coverage with the highest number of redundant reads compared to leaf transcriptome from Illumina sequencing (Table 1). To obtain a more comprehensive reference transcriptome library, we performed combined assembly using normalized Illumina and trimmed 454 reads (Fig. 1). After digital normalization of combined raw reads, 34.37 million high quality reads were assembled into 188,735 transcripts with a median length (N50) of 1,009 bp, and a total length of 136.67 Mbp (Table 1). The number of predicted coding sequence (CDS) also increased for the combined assembly. Combined assembly has a smaller size range compared to leaf assembly alone. This indicates that combined assembly pipeline helps in reducing erroneously long transcripts (>5 kbp) and closing the gaps in the root transcriptome with reduced number of transcript in the combined assembly. This is possibly due to low coverage of root transcriptome resulting in abundant short transcripts (Fig. 2). The combined assembly of the leaf and root transcriptomes also increased the number of transcript, especially for transcripts with length below 1,500 bp. Furthermore, a cross comparison of 3,538 *Polygonum minus* ESTs from a previous study against the combined assembly shows 100% reciprocal hits (Table S2) which supports a comprehensive coverage of the combined assembly.

Similarity search

A summarized assembly and annotation workflow is illustrated in Fig. 1. A total of 163,200 (86.5%) transcripts from the *Polygonum minus* combined (leaf and root)
transcriptome assembly were annotated with putative functions based on similarity to sequences in the Swiss-Prot protein database using BLASTX. Among all the BLAST hits, 37% were distributed in the range of E-value lower than $1 \times 10^{-50}$, 23% with E-value between $1 \times 10^{-25}$ and $1 \times 10^{-50}$, and majority of the hits (40%) were distributed in the range of $1 \times 10^{-5}–1 \times 10^{-25}$ (Fig. 3A). BLAST hit similarity distribution analysis showed that

**Figure 1** Schematic diagram showing analysis workflow of *Polygonum minus* transcriptome combined assembly and downstream analysis.
majority of the BLAST hits (64%) were in between 40 and 80%; 16% of the transcripts had at least 80% sequence similarity compared to 20% with less than 40% sequence similarity (Fig. 3B). For BLAST hit species distribution (Fig. 3C), we identified that majority of the homologous matches were from *Arabidopsis thaliana* (59%), which is the top most represented plant species in Swiss-Prot database followed by *Oryza sativa* (4%), *Vitis vinifera* (0.9%), and *Glycine max* (0.5%). The rest were classified as “Others” (35.6%) which include mostly hypothetical proteins.

**Functional annotation and classification of the assembled transcripts**

A total of 1,109,319 GO terms were assigned to the annotated *Polygonum minus* transcripts classified based on BLASTX similarity search. In many cases, multiple terms were assigned to the same transcript, and all of the GO terms were classified into 53 functional groups under three categories of cellular components, biological processes, and molecular functions (Fig. 4). For cellular components, most assignments were cell and cell part (76,052; 76.1%). Within the molecular function category, the GO terms were predominantly assigned to binding (63,382; 63.4%) and catalytic activities (54,855; 54.9%). In biological processes, transcript sequences assigned to cellular (64,870; 64.9%) and metabolic (55,632; 55.6%) processes were the most abundant, followed by biological regulation (22,516; 22.5%), pigmentation (20,141; 20.1%) and response to stimulus (19,847; 19.8%). Similar annotation was observed in KOG classification with the most abundant transcript found in signal transduction

![Figure 2](image-url)
Figure 3 BLASTX analysis of *Polygonum minus* combined assembly. (A) E-value distribution, (B) Similarity distribution, (C) Homologous gene-species distribution.
mechanism (Fig. S2). These findings revealed that the main functions of the annotated transcripts were responsible for fundamental biological regulation and metabolism common in plants.

**KEGG pathway mapping and gene enrichment analysis**

KEGG pathways were mapped with *Polygonum minus* transcripts based on the GO annotation from BLASTX results. A full list of mapped transcripts related to plant secondary metabolite biosynthesis is available as an online material (Data S1). Table 2 shows the KEGG pathways related to selected secondary metabolite biosynthesis which were mapped with *Polygonum minus* combined assembly in comparison with that of publicly available *Polygonum minus* EST sequences. A total of 28,643 transcripts were mapped to 474 out of 949 enzymes in KEGG Orthology (KO) pathways related to the biosynthesis of secondary metabolites. Pathways with majority number of entries mapped include “KO000941 Flavonoid biosynthesis” (17; 89%), “KO00944 Flavone and flavonol biosynthesis” (10; 83%), and “KO00901 Indole alkaloid biosynthesis” (8; 80%). The “KO00940 Phenylpropanoid biosynthesis” pathway has the greatest number of transcript count of 3,753, followed by “KO00906 Carotenoid biosynthesis” (1,259), and “KO00900 Terpenoid backbone biosynthesis” (1,122). The transcript abundance could reflect the importance of these pathways in *Polygonum minus*. Figure 5 depicts the KEGG phenylpropanoid pathway mapped with the combined assembly compared
to the previous EST library. This shows that current transcriptome has greatly expanded the number of discovered transcripts related to plant secondary metabolite biosynthesis in *Polygonum minus*.

To further examine the relationship between tissue types and biological processes, KEGG pathway enrichment analysis was performed with KOBAS. A list of significant

![Table 2 KEGG pathways related to selected secondary metabolite biosynthesis mapped with RNA-seq assembly compared to that of publicly available *Polygonum minus* EST sequences.](Image)

<table>
<thead>
<tr>
<th>KEGG Pathway</th>
<th>Total entry</th>
<th>Mapped entry</th>
<th>EST</th>
<th>RNA-seq</th>
<th>Transcript count</th>
</tr>
</thead>
<tbody>
<tr>
<td>KO01110 Biosynthesis of secondary metabolites</td>
<td>949</td>
<td>127</td>
<td>2%</td>
<td>12</td>
<td>13%  474</td>
</tr>
<tr>
<td>KO00909 Sesquiterpenoid and triterpenoid biosynthesis</td>
<td>66</td>
<td>5</td>
<td>11%</td>
<td>9</td>
<td>28%  1,259</td>
</tr>
<tr>
<td>KO00906 Carotenoid biosynthesis</td>
<td>46</td>
<td>4</td>
<td>10%</td>
<td>15</td>
<td>36%  107</td>
</tr>
<tr>
<td>KO00904 Diterpenoid biosynthesis</td>
<td>53</td>
<td>11</td>
<td>21%</td>
<td>28</td>
<td>53%  1,122</td>
</tr>
<tr>
<td>KO00940 Phenylpropanoid biosynthesis</td>
<td>32</td>
<td>9</td>
<td>28%</td>
<td>23</td>
<td>72%  3,753</td>
</tr>
<tr>
<td>KO00902 Monoterpenoid biosynthesis</td>
<td>22</td>
<td>1</td>
<td>5%</td>
<td>12</td>
<td>55%  54</td>
</tr>
<tr>
<td>KO00941 Flavonoid biosynthesis</td>
<td>19</td>
<td>7</td>
<td>37%</td>
<td>17</td>
<td>89%  893</td>
</tr>
<tr>
<td>KO00942 Anthocyanin biosynthesis</td>
<td>14</td>
<td>1</td>
<td>7%</td>
<td>4</td>
<td>29%  985</td>
</tr>
<tr>
<td>KO00943 Isoflavonoid biosynthesis</td>
<td>13</td>
<td>0</td>
<td>0%</td>
<td>5</td>
<td>38%  34</td>
</tr>
<tr>
<td>KO00944 Flavone and flavonol biosynthesis</td>
<td>12</td>
<td>1</td>
<td>8%</td>
<td>10</td>
<td>83%  680</td>
</tr>
</tbody>
</table>

![Figure 5 KEGG pathway of flavonoid biosynthesis mapped with combined assembly. Orange/darker shading: found in combined transcriptome and EST library. Yellow/lighter shading: found only in combined transcriptome.](Image)
pathways in *Polygonum minus* leaf and root is showed in Table 3. In leaf, pathways related to photosynthesis, primary metabolite metabolism and biosynthesis of secondary metabolites (carotenoids and diterpenoids) were significantly enriched. Conversely, pathways for phenylpropanoid biosynthesis, regulation of actin cytoskeleton, and plant signaling were significantly enriched in the root.

**RT-qPCR validation of RNA-seq relative expression estimation**

We estimated the expression level of each transcript in the leaf and root tissues based on TPM values (Data S2). To validate the relative expression levels from the transcript abundance estimation, seven transcripts related to phenylpropanoid and flavonoid biosynthetic pathways were chosen for RT-qPCR, namely phenylalanine ammonia-lyase (*PAL*), 4-coumarate–CoA ligase (*4CL*), trans-cinnamate 4-monooxygenase (*C4H*), chalcone synthase (*CHS*), chalcone isomerase (*CHI*), flavanone 3-dioxygenase (*F3H*), and dihydroflavonol-4-reductase (*DFR*) (Fig. 6A). All of the transcripts related to the seven enzymes were identified from the transcriptome sequences based on pairwise sequence alignment BLAT hit result at E-value cutoff of 1e–5 (Table 4, Data S3). Clustered transcripts (contigs) showed variable expression in the leaf and root tissues (Fig. 6B). Most of the contigs had higher TPM values in the leaf tissue; while some contigs were higher in the root tissues, especially for C4H and F3H. Consistently, all of the selected transcripts showed higher expression in the leaf than the root (Fig. 6C). Despite that contig c49638 not showing the expected expression, RT-qPCR results still showed a good correlation ($R^2 > 67\%$) to that of estimated based on RNA-seq TPM values (Fig. 6D).

<table>
<thead>
<tr>
<th>KEGG Pathway</th>
<th>Total entry</th>
<th>Tissue-specific entry</th>
<th>$P$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>(A) Leaf transcriptome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO00040—pentose and glucuronate interconversions</td>
<td>305</td>
<td>186</td>
<td>0.00432</td>
</tr>
<tr>
<td>KO00196—photosynthesis—antenna proteins</td>
<td>101</td>
<td>68</td>
<td>0.01666</td>
</tr>
<tr>
<td>KO00280—valine, leucine, and isoleucine degradation</td>
<td>328</td>
<td>188</td>
<td>0.02246</td>
</tr>
<tr>
<td>KO00380—tryptophan metabolism</td>
<td>166</td>
<td>102</td>
<td>0.02422</td>
</tr>
<tr>
<td>KO00906—carotenoid biosynthesis</td>
<td>198</td>
<td>118</td>
<td>0.02965</td>
</tr>
<tr>
<td>KO00071—fatty acid degradation</td>
<td>377</td>
<td>211</td>
<td>0.03102</td>
</tr>
<tr>
<td>KO00500—starch and sucrose metabolism</td>
<td>927</td>
<td>487</td>
<td>0.03692</td>
</tr>
<tr>
<td>KO00904—diterpenoid biosynthesis</td>
<td>70</td>
<td>47</td>
<td>0.04188</td>
</tr>
<tr>
<td><strong>(B) Root transcriptome</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KO00940—phenylpropanoid biosynthesis</td>
<td>645</td>
<td>26</td>
<td>0.00005</td>
</tr>
<tr>
<td>KO00360—phenylalanine metabolism</td>
<td>472</td>
<td>20</td>
<td>0.00020</td>
</tr>
<tr>
<td>KO04810—regulation of actin cytoskeleton</td>
<td>439</td>
<td>14</td>
<td>0.01727</td>
</tr>
<tr>
<td>KO04075—plant hormone signal transduction</td>
<td>1,086</td>
<td>26</td>
<td>0.04266</td>
</tr>
<tr>
<td>KO04024—cAMP signaling pathway</td>
<td>411</td>
<td>12</td>
<td>0.04525</td>
</tr>
<tr>
<td>KO00562—inositol phosphate metabolism</td>
<td>326</td>
<td>10</td>
<td>0.04919</td>
</tr>
<tr>
<td>KO00592—alpha-linolenic acid metabolism</td>
<td>282</td>
<td>9</td>
<td>0.04937</td>
</tr>
</tbody>
</table>
DISCUSSION

Hybrid assembly generated a comprehensive reference transcriptome

We performed combined assembly using simulated reads from root transcriptome assembly with Trinity pipeline when the overlap-layout-consensus approach using MIRA failed to combine 454 and Illumina reads. Comparative analyses of de novo transcriptome assemblers based on both simulated and real RNA-seq data (Mundry et al., 2012).

Figure 6 RT-qPCR validation of RNA-seq relative expression estimation. (A) The integrated phenylpropanoid (black) and flavonoid (orange) biosynthesis pathway showing selected genes. Each gene is followed in parentheses by the number of contigs homologous to gene families encoding this enzyme which are expressed in the leaf and root tissues. (B) Expression heatmap of contigs expressed in the leaf and root tissues sorted in descending order of root expression values (Log_{10} TPM). Contigs chosen for RT-qPCR are underlined whereas contigs showing higher expression in the root are marked with an asterisk. (C) RT-qPCR analysis showing the fold of expression level in the root relative to expression level in the leaf. Error bars show the confidence intervals calculated from 1 SE of ΔΔCt. (D) Correlation plot between the RT-qPCR fold-change (FC) compared to FC calculated from TPM values of RSEM estimates. PAL, phenylalanine ammonia-lyase; 4CL, 4-coumarate—CoA ligase; C4H, trans-cinnamate-4-monoxygenase; CHS, chalcone synthase; CHI, chalcone isomerase; F3H, flavanone 3-dioxygenase; DFR, dihydroflavonol-4-reductase.

Loke et al. (2017), PeerJ, DOI 10.7717/peerj.2938
suggested that Trinity performed better than MIRA, which is computationally more intensive. Furthermore, MIRA and CAP3 were previously reported to be conservative in merging reads, resulting in a high number of redundant short contigs but fewer chimeric contigs (Clarke et al., 2013). This appeared to be the case for the root transcriptome in which the 190,269 transcripts from iAssembler pipeline were significantly reduced to 26,301 proportionally longer transcripts (Fig. S3) when constructed with Trinity pipeline using the simulated Illumina reads (Table S4). The comparison between the independent and combined assemblies indicates an improvement in the de novo transcriptome quality through the combined assembly approach, based on sequence statistics (Table 1), transcript length distribution (Fig. 1) and validation using EST sequences (Table S2).

The combined assembly provides a more comprehensive reference transcriptome of Polygonum minus compared to EST library (Roslan et al., 2012) and cDNA-amplified fragment length polymorphism (cDNA-AFLP) transcript profiling (Ee et al., 2013). All previously found ESTs were covered by the current transcriptome. For example, seven enzymes involved in the flavonoid biosynthesis pathway were identified from previous EST library compared to 17 in current transcriptome study (Table 2). Another two enzymes which were not found in the flavonoid biosynthetic pathway are coumaroylquinone (coumaroylshikimate) 3′-monooxygenase [EC:1.14.13.36] and flavanone 7-O-beta-glucosyltransferase [EC:2.4.1.185] (Fig. 5). These two enzymes might be absent or lowly expressed, or no homologous counterpart can be found in the similarity search.

### Pathway enrichment in the leaf and root tissues

Downstream analyses based on combined assembly such as functional annotation and pathway analysis provide insights into the biological processes relevant to the leaf and root tissues. “Photosynthesis” and “Starch and sucrose metabolism” were identified as significantly enriched pathways in leaf (Table 3). The products from these pathways such as xylose can be a substrate for the “Pentose and glucuronate interconversions” pathway. Glucuronate is one of the major sugars required in plant cell wall development (Seitz et al., 2000). These pathways are also involved in the phenolic

---

**Table 4** Summary of genes involved in phenylpropanoid and flavonoid biosynthesis in *Polygonum minus*.

<table>
<thead>
<tr>
<th>Gene</th>
<th>KO entry</th>
<th>EC no.</th>
<th>Enzyme name</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAL</td>
<td>K10775</td>
<td>4.3.1.5</td>
<td>Phenylalanine ammonia-lyase</td>
<td>37 23 13</td>
</tr>
<tr>
<td>4CL</td>
<td>K01904</td>
<td>6.2.1.12</td>
<td>4-Coumarate—CoA ligase</td>
<td>15 8 4</td>
</tr>
<tr>
<td>C4H</td>
<td>K00487</td>
<td>1.14.13.11</td>
<td>trans-Cinnamate-4-monooxygenase</td>
<td>20 10 6</td>
</tr>
<tr>
<td>CHS</td>
<td>K00660</td>
<td>2.3.1.74</td>
<td>Chalcone synthase</td>
<td>26 18 12</td>
</tr>
<tr>
<td>CHI</td>
<td>K01859</td>
<td>5.5.1.6</td>
<td>Chalcone isomerase</td>
<td>18 3 2</td>
</tr>
<tr>
<td>F3H</td>
<td>K00475</td>
<td>1.14.11.9</td>
<td>Flavanone 3-hydroxylase</td>
<td>17 9 6</td>
</tr>
<tr>
<td>DFR</td>
<td>K13082</td>
<td>1.1.1.219</td>
<td>Dihydroflavonol 4-reductase</td>
<td>27 13 10</td>
</tr>
</tbody>
</table>

---

Loke et al. (2017), PeerJ, DOI 10.7717/peerj.2938
and antioxidant response pathways and could be related to the high antioxidant and medicinal values of *Polygonum minus* (*Mohd Ghazali et al.*, 2014; *Shetty*, 2004). Furthermore, the “Carotenoid biosynthesis” and “Diterpenoid biosynthesis” pathways were also enriched in the leaf. This is consistent with the important role of carotenoids and terpenoids in plant protection. Carotenoids can protect the leaf from photo oxidative damage of excessive light by scavenging reactive oxygen species (*Cazzonelli*, 2011). Terpenes may act as a chemical messenger to regulate the expression of genes in plant defense as well as influencing gene expression of neighboring plants (*Cheng et al.*, 2007). The volatile diterpenes can also protect plants against insects and herbivores (*Heiling et al.*, 2010; *Seo et al.*, 2012). This also supports the discovery of multiple terpenes from chemical analysis of *Polygonum minus* leaf essential oil (*Baharum et al.*, 2010).

“Phenylpropanoid biosynthesis” and “Phenylalanine metabolism” pathways were identified to be enriched in the root transcriptome. Phenylpropanoids are known to be directly involved in plant stress responses towards temperature, drought, UV, and lack of nutrients (*Korkina*, 2007). The pathway is branched into several important pathways such as flavonoid biosynthetic pathway which plays role in plant resistance (*Treutter*, 2006). This pathway was highlighted in previous EST analysis which identified 11 ESTs encoding for seven enzymes in the biosynthesis of flavonoids (Fig. 6) (*Roslan et al.*, 2012). Other enriched pathways in the root include “Plant hormone signal transduction” and “cAMP signaling” pathways which indicate the important role of signaling in plant root. Furthermore, inositol phosphate and alpha-linolenic acid related to jasmonic acid are also important in plant signaling and stress response (*Witzany*, 2006). This supports the importance of root function in detecting environmental signals and transduce them through phytohormones.

**Analysis of transcripts involved in the biosynthesis of phenylpropanoids and flavonoids**

Flavonoids are products of the phenylpropanoid metabolism, which bridges between primary and secondary metabolism, through $p$-coumaroyl-CoA that serves as a precursor for flavonoid biosynthesis (Fig. 6A). Based on the abundant transcripts (Data S2) and pathways enrichment analysis in *Polygonum minus* leaf and root tissues (Table 3), we focused on identifying all the transcripts related to the seven important enzymes in phenylpropanoid and flavonoid biosynthesis (Table 4, Data S3). These transcripts can be grouped as unigenes and further clustered into contigs as representative gene families (*Grabherr et al.*, 2011). In general, most of the contigs related to phenylpropanoid and flavonoid biosynthesis were more abundant in the leaf tissue based on the normalized TPM values (Fig. 6B). However, certain contigs did show otherwise which indicate differential regulation of different gene families in the leaf and root tissues as proposed in previous study (*Roslan et al.*, 2012). The EST study by *Roslan et al.* (2012) showed higher expression of three genes related to flavonoid biosynthesis in *Polygonum minus* root, namely *CHS*, flavonol synthase (*FLS*), and leucoanthocyanidin dioxygenase (*LDOX*). Indeed, our
transcriptome analysis also showed higher expression of certain transcripts in the root tissue (Data S2 and S3). The comprehensive transcriptome analysis and robust correlation with RT-qPCR results in current study showed that the expression of transcripts related to the biosynthesis of phenylpropanoids and flavonoids in Polygonum minus were more prevalent in the leaf than the root tissue. This finding is consistent with recent metabolite profiling study in different tissues of Polygonum minus which showed that the phenolic content and antioxidant activity in the leaf extract is much higher than that of root extract (Ahmad et al., 2014).

The lack of a reference genome for Polygonum minus has made it difficult to determine the exact number of genes involved in the metabolic pathways. Here, a large number of candidate transcripts could be matched with known enzymes in public databases which provides a substantial gene resource for further research in Polygonum minus. Our study also provides reference sequences for evolutionary analyses of metabolomes among the members of Polygonaceae family which is rich in ethnomedicinal plants.

CONCLUDING REMARKS
In this study, we described the first comprehensive transcriptome profile of Polygonum minus leaf and root tissues for the curation of secondary metabolite-related transcripts (Data S1). We compared the annotation results of three Polygonum minus assemblies (leaf, root, and combined) with that of previous EST library. This comparison provides a useful resource for gene discovery using Polygonum minus combined assembly. We have established a reference transcriptome profile of Polygonum minus with annotations of transcript descriptions for future investigation on specific processes or pathways, especially on the effect of environmental stresses and the biosynthesis of secondary metabolites. We also identified some of the enriched metabolite pathways in the root and leaf tissues. The identification of transcripts related to secondary metabolite biosynthesis will aid in further exploitation of the genetic resource from this herbal plant for future biotechnological development.

ACKNOWLEDGEMENTS
Sequencing was done at Malaysia Genome Institute and BGI-Shenzhen, PR China. Database and server hosting were supported by Centre for Bioinformatics Research, INBIOSIS. We also thank the two reviewers for improving this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding
The research was supported by the Fundamental Research Grant Scheme (FRGS/1/2013/SG05/UKM/01/2) from the Malaysian Ministry of High Education (MOHE) awarded to II and Research University Grant under the Arus Perdana (UKM-AP-BPB-14-2009) from Universiti Kebangsaan Malaysia awarded to NMN. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Grant Disclosures
The following grant information was disclosed by the authors:
Malaysian Ministry of High Education (MOHE): FRGS/1/2013/SG05/UKM/01/2.

Competing Interests
The authors declare that they have no competing interests.

Author Contributions
- Kok-Keong Loke analyzed the data, wrote the paper, prepared figures and/or tables.
- Reyhaneh Rahnamaie-Tajadod conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables.
- Chean-Chean Yeoh conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables.
- Hoe-Han Goh conceived and designed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables.
- Zeti-Azura Mohamed-Hussein reviewed drafts of the paper.
- Zamri Zainal reviewed drafts of the paper.
- Ismanizan Ismail conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper, acquired funding.
- Normah Mohd Noor conceived and designed the experiments, contributed reagents/materials/analysis tools, reviewed drafts of the paper, acquired funding.

DNA Deposition
The following information was supplied regarding the deposition of DNA sequences:
- BioProject: PRJNA208436.
- BioSample: SAMN02211834.
- Sequence Read Archive: SRR917962.

Data Deposition
The following information was supplied regarding data availability:
- Assembled transcripts can be found at http://prims.researchfrontier.org/index.php/dataset/transcriptome.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.2938#supplemental-information.

REFERENCES
DOI 10.3390/molecules191119220.


