Plant Biotechnology Journal

Plant Biotechnology Journal (2013) **11**, pp. 115–127 doi: 10.1111/pbi.12018

Deep sequencing analysis of the transcriptomes of peanut aerial and subterranean young pods identifies candidate genes related to early embryo abortion

Xiaoping Chen[1], Wei Zhu[1,2], Sarwar Azam[3], Heying Li[4], Fanghe Zhu[1], Haifen Li[1], Yanbin Hong[1], Haiyan Liu[1], Erhua Zhang[1], Hong Wu[4], Shanlin Yu[5], Guiyuan Zhou[1], Shaoxiong Li[1], Ni Zhong[1], Shijie Wen[1], Xingyu Li[1], Steve J. Knapp[6], Peggy Ozias-Akins[6], Rajeev K. Varshney[1,3] and Xuanqiang Liang[1,*]

[1]*Crops Research Institute, Guangdong Academy of Agricultural Sciences (GAAS), Guangzhou, China*

[2]*College of Life Science, South China Normal University, Guangzhou, China*

[3]*International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru, India*

[4]*State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, South China Agricultural University, Guangzhou, China*

[5]*Shandong Peanut Research Institute, Shandong Academy of Agricultural Sciences, Qingdao, China*

[6]*University of Georgia, Athens, GA, USA*

Received 10 August 2012;
revised 8 September 2012;
accepted 27 September 2012.
**Correspondence* (fax +862085514269;
email Liang-804@163.com)

Keywords: peanut, aerial and subterranean pod, transcriptome, RNA sequencing.

Summary

The failure of peg penetration into the soil leads to seed abortion in peanut. Knowledge of genes involved in these processes is comparatively deficient. Here, we used RNA-seq to gain insights into transcriptomes of aerial and subterranean pods. More than 2 million transcript reads with an average length of 396 bp were generated from one aerial (AP) and two subterranean (SP1 and SP2) pod libraries using pyrosequencing technology. After assembly, sets of 49 632, 49 952 and 50 494 from a total of 74 974 transcript assembly contigs (TACs) were identified in AP, SP1 and SP2, respectively. A clear linear relationship in the gene expression level was observed between these data sets. In brief, 2194 differentially expressed TACs with a 99.0% true-positive rate were identified, among which 859 and 1068 TACs were up-regulated in aerial and subterranean pods, respectively. Functional analysis showed that putative function based on similarity with proteins catalogued in UniProt and gene ontology term classification could be determined for 59 342 (79.2%) and 42 955 (57.3%) TACs, respectively. A total of 2968 TACs were mapped to 174 KEGG pathways, of which 168 were shared by aerial and subterranean transcriptomes. TACs involved in photosynthesis were significantly up-regulated and enriched in the aerial pod. In addition, two senescence-associated genes were identified as significantly up-regulated in the aerial pod, which potentially contribute to embryo abortion in aerial pods, and in turn, to cessation of swelling. The data set generated in this study provides evidence for some functional genes as robust candidates underlying aerial and subterranean pod development and contributes to an elucidation of the evolutionary implications resulting from fruit development under light and dark conditions.

Introduction

The most prominent feature of fruit production by which species in the genus *Arachis* are distinguished from most other plants is that of aerial flowering, self-pollination, fertilization and peg (gynophore) formation, followed by gravitropic peg elongation and penetration into the soil and then subterranean fructification. These developmental events have the biologically important value for studying organogenesis and evolution. Moreover, peg penetration into the soil, and then swelling of gynophore tips are essential processes for peanut pod development, which is the crucial determinant of peanut yield. The peg harbours developing embryos, which are quiescent until the peg penetrates into the soil, and then the peg must swell to form a pod to allow room for the embryo to grow (Feng *et al.*, 1995). Thus, the failure of peg penetration into the soil, swelling in peg tips, and resumption of embryo development following quiescence can lead to abortion or acarpia, thereby contributing to further seed yield loss. Despite the biological and agronomic importance of pod swelling and subsequent embryo resumption in pod growth and development, little knowledge is available pertaining to genes regulating the initiation of swelling in peg tips and embryo development.

Previous studies showed that pod swelling and embryo development were controlled by growth regulators such as auxin, kinetin, ABA and IAA (Jacobs, 1951; Ziv and Zamski, 1975; Ziv and Kahana, 1988; Shlamovitz *et al.*, 1995) and peg elongation required the development of proembryos which synthesize growth regulators that stimulate cell division (Zamski and Ziv, 1975). Due to aerial flowering and geocarpy, both air and soil temperature and soil water content can affect development of peanut pegs into pods (Lee *et al.*, 1972; Balasubramanian and Yayock, 1981; Golombek and Johansen, 1997; Sexton *et al.*, 1997; Varaprasad *et al.*, 1999, 2000). It is not surprising that light and dark are also involved in the pod development. Their effects on the cessation and reactivation of embryo and pod development have been well characterized (Zamski and Ziv, 1975; Stalker and Wynne, 1983; Thompson *et al.*, 1985; Shlamovitz *et al.*, 1995; Nigam *et al.*, 1997). Light was found to promote peg elongation and to inhibit pod formation (Shlamovitz *et al.*, 1995). Several studies indicated that the intercalary meristem caused the

peg to elongate in the light, and the peg did not cease to elongate until a pod started to develop (Smith, 1950; Zamski and Ziv, 1975; Pattee *et al.*, 1988).

Compared with the extensive studies on physiological and environmental factors influencing peanut peg development and pod initiation (Underwood *et al.*, 1971), studies on the underlying genetic determinants controlling this process have received far less attention. Especially, knowledge regarding genes and their expression patterns related to aerial and subterranean pod development is yet to be gained. During the past several years, expressed sequence tag (EST) sequencing studies generated a large number of cDNA sequences (Guo *et al.*, 2008, 2009; Bi *et al.*, 2010; Tirumalaraju *et al.*, 2011). However, with traditional methods of sequencing randomly selected cDNA clones from various tissues, low coverage of less-abundant or rare transcripts, which usually play vital roles, was obtained. Recently, a new approach to transcriptome profiling using next-generation sequencing technologies (NGS), termed RNA sequencing (RNA-seq), has been developed (Varshney *et al.*, 2009; Wang *et al.*, 2009). It has been widely applied in plant biology, both in model species such as Arabidopsis (Weber *et al.*, 2007) and also in crop plants including rice (Lu *et al.*, 2010), maize (Li *et al.*, 2010), and recently, in peanut (Pandey *et al.*, 2012; Zhang *et al.*, 2012).

Here, we used NGS to globally investigate and compare the transcriptomes of peanut aerial and subterranean young pods (for convenience, hereafter, we will not use peg, but aerial pod and subterranean pod) to improve our understanding of peanut aerial and subterranean pod development. This study was undertaken: (i) to define and annotate the transcriptomes of peanut aerial and subterranean pods, (ii) to identify differentially expressed genes (DEGs) between aerial and subterranean pod development, (iii) to discover functional genes related to embryo abortion in aerial developing pod, and (iv) also to provide a valuable transcriptomics sequence resource of aerial and subterranean early-stage pod development for the peanut community.

Results

Comparison of embryo development of aerial and subterranean pods

It is still unknown why aerially developing pods can not swell normally compared with those that penetrate into soil. For better understanding of the early stage of pod development, histochemical analysis was conducted to compare the early embryo development of aerial and subterranean pods. The subterranean pod tissues were collected 0, 2, 4 and 6 days after penetration into soil (DAPS) and the aerial pod tissues were collected 8, 10, 12 and 14 days after flowering (DAF) corresponding to time points for subterranean pod collection. We compared developing embryos of aerial and subterranean young pods through staining of paraffin sections (Figure 1). No changes were found between the first and second time points for embryos in both aerial and subterranean young pods. Starting from 4 DAPS, the embryo of the subterranean pod had enlarged compared with that in 12-DAF aerial pods. Eventually, the embryo of the aerial pod was found to cease growth and finally aborted.

Transcriptome sequencing and *de novo* assembly

In this study, three cDNA libraries (AP, SP1 and SP2, see Materials and methods section for detailed library information and Figure S1) were constructed using total RNA from aerial and subterranean pods, respectively, and each was subjected to a half run on a 454 GS FLX Titanium platform. Approximately, 274, 290 and 238 Mb of sequence data were obtained for AP, SP1 and SP2, respectively, in the form of 704 738, 711 496 and 609 841 reads averaging 396 bp in length. Transcript reads containing adaptor sequences were cleaned, and the sequence data were filtered for low-quality reads at high stringency. This resulted in a total of 1 683 455 high-quality reads containing 664 846 199 bases with an average read size of 395 bp (Table 1). A great number of reads were distributed around 500 bp with lengths ranging from 40 to 852 bp (Figure S2). The sequencing data generated in this study are deposited in NCBI Sequence Read Archive (SRA, http://www.ncbi.nlm.nih.gov/sra) with accession number SRA053198.

We employed TGICL2.0 and Newbler (v2.6) for *de novo* assembly of peanut young pod transcriptomes and compared their performance. The parameters used for the two programs are described in Materials and methods section. Summary of these assemblies generated by the two programs has been given in Table 2. The TGICL program, as compared to Newbler, produced a much larger number of contigs with 28% of contigs 1 kb or greater in length. In contrast, the Newbler program generated more than 63% of contigs 1 kb or greater in length, much higher than TGICL. The frequency distribution of contig lengths showed that Newbler generated a larger number of long

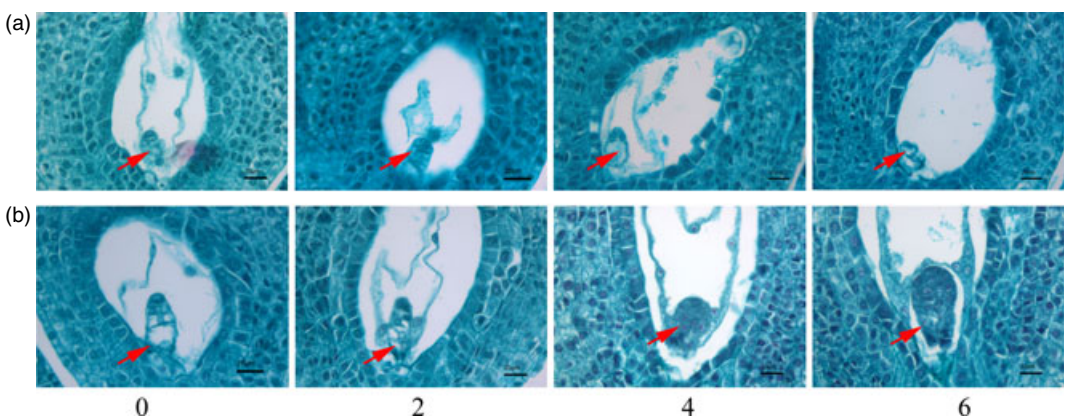

Figure 1 Comparison of embryo development at different stages in aerial (a) and subterranean (b) young pods through staining of paraffin sections. The numbers, 0, 2, 4 and 6, indicate days after penetrating soil. Embryos are indicated with red arrows. Bars = 20 μm.

Table 1 Summary of sequencing data

Library	No. of raw reads	Total length (bp)	No. of high-quality reads after removing low-quality reads	Total length (bp) of high-quality reads
AP	704 738	274 182 868	540 918	207 079 674
SP1	711 496	290 240 364	614 517	250 480 616
SP2	609 841	238 082 819	528 020	207 285 909
Total	2 026 075	802 506 051	1 683 455	664 846 199

Table 2 Summary of de novo assembly results of 454 sequence data using TGICL2 and Newbler (v2.6)

Program	TGICL			Newbler		
Parameter and/or input data type	Filtered reads (p90)	Filtered reads (p95)	Raw reads (p95)	Filtered reads	Raw reads	SFF format
Input reads	1 683 455	1 683 455	2 026 075	1 683 455	2 026 075	2 026 075
Assembled reads	1 456 254	1 443 329	1 741 599	1 423 137	1 693 752	1 681 975
Contigs	46 752	48 524	49 883	28 302	28 930	29 359
Total Size (bp)	40 150 565	41 499 671	42 040 740	40 519 221	42 485 486	41 235 047
Contigs ($\geq$100 bp)	46 487	48 231	49 429	28 279	28 886	29 330
Contigs ($\geq$1 kb)	13 074	13 476	13 476	17 830	18 147	18 770
Maximum length (bp)	7660	7660	7175	7624	7760	7760
N50 (bp)	974	971	961	1714	1717	1730
Average length (bp)	859	850	843	1431	1425	1447
Contigs with significant hit (%)	39.081 (84.1)	40 150 (83.2)	40 653 (82.2)	25 808 (91.2)	26 229 (90.8)	26 682 (91.0)
Contigs with 80% or greater coverage*	9407	9602	9468	15 122	15 356	15 758
Soybean protein hits†	20 912	21 202	21 240	15 155	15 300	15 301
Soybean proteins with 80% or greater coverage‡	9876	10 073	9940	10 083	10 197	10 320

*The number of contigs with 80% or greater coverage of soybean proteins.

†The number of unique soybean proteins to which contigs show significant similarity ($\leq$1E−5).

‡The number of unique soybean proteins to which contigs show significant similarity and $\geq$80% coverage.

contigs that were more evenly distributed across different length ranges (Figure S3A). Newbler used SFF format data as input and generated an assembly with an N50 of 1730 bp and an average contig length of 1447 bp. Although the number of contigs, N50 and average length showed differences between the two programs, the maximum length of contigs and the total size of assemblies generated by both programs were comparable (Table 3 and Figure S3B). Overall, TGICL generated a larger number of contigs representing the diversity of transcripts, while Newbler output the longer average contig length and lower number of contigs showing a remarkably higher contiguity.

To validate the assemblies generated by TGICL and Newbler, we used the soybean proteome predicted from the genome sequence (Schmutz *et al.*, 2010) as a reference sequence. The quality of assembly should be substantiated by the coverage of the soybean proteome. We identified the number of soybean proteins to which the assembled contigs exhibited significant similarity (Table 2). Comparison of the summed proportion of soybean proteins covered by contigs from various assemblies showed that a larger number of soybean proteins were represented in TGICL assembly (Figure S4). Fewer contigs resulting from Newbler showed significant similarity to soybean proteins. Although assemblies generated by Newbler had a higher N50, longer average lengths and a larger number of contigs 1kb or greater in length, the number of unique soybean proteins to which contigs generated by the two assemblers showed 80% or greater coverage was comparable (Table 2). When the coverage percentage decreased to $\geq$50%, the TGICL assembly showed significant similarity to over 15 000 unique soybean proteins much higher than that of Newbler (<13 000 hits) (Figure S4).

Defining aerial and subterranean young pod transcriptomes

To estimate how many genes were expressed and to profile the aerial and subterranean young pod transcriptomes comprehensively, we constructed a reference transcript data set generated by a combination of *de novo* assembly in this study and the UGA Tifrunner reference data. After removing redundant sets of sequences and short sequences below 100 bp, we obtained a total of 151 533 transcript assembly contigs (TACs) representing 114 867 135 (114.9 Mb) of the sequence with an average length of 758 bp and a N50 of 866 bp. To further characterize the peanut transcriptome, GC content of transcripts for peanut, soybean, *M. truncatula*, *L. japonicus*, *Arabidopsis* and rice was also computed (Figure 2). Peanut has approximately 65% of transcripts with GC content in the range of 35%–40% (Figure 2a). The range of GC content of peanut transcripts was narrow when compared with other species used in this study (Figure 2b). The average GC content of peanut transcripts is approximately 40.5% similar to that of some other legume species such as soybean (41.5%), *M. truncatula* (39.5%) and chickpea (40.3%) (Garg *et al.*, 2011a,b), but slightly lower than that of *Arabidopsis* (42.5%) and *L. japonicus* (42.4%).

Table 3 Top 30 DEGs with putative functions

					UniProtKB Blast			
GeneID	R	AP	SP1	SP2	Acc. No.	Species	E Value	Description
AHTC10019122	3792	33286	17389	13970	Q9AVH2	*P. sativum*	4E−88	Putative senescence-associated protein m
AHTC10119531	1355	281	1166	2688	Q43374	*A. hypogaea*	2E−138	Mannose/glucose-binding lectin
AHTC10022137	1053	5314	3135	1300	G7K0D4	*M. truncatula*	9E−40	RRNA intron-encoded homing endonuclease
AHTC10010404	880.9	7282	3640	3041	G7LEH3	*M. truncatula*	6E−54	ATP synthase subunit beta
AHTC10035505	836.4	7357	5193	2658	G7K0C9	*M. truncatula*	1E−49	Cytochrome P450 likeTBP
AHTC10014520	652.8	4526	2264	1615	D7L5C8	*A. lyrata*	2E−22	Expressed protein
AHTC10022115	588.9	2169	3400	4456	Q43374	*A. hypogaea*	2E−148	Mannose/glucose-binding lectin
AHTC10047690	578.6	14	109	700	G7JJ73	*M. truncatula*	2E−10	Peroxidase
AHTC10047992	535	6	84	609	No hits found			
AHTC10034863	533.7	3058	1419	965	G7K0D6	*M. truncatula*	4E−54	RRNA intron-encoded homing endonuclease
AHTC10004231	470.6	232	233	1023	G1EUI6	*P. elegans*	2E−86	Lectin
AHTC10014910	458.1	637	92	8	G7JTH4	*M. truncatula*	6E−164	Caffeic acid 3-O-methyltransferase
AHTC10038092	453.6	4484	2237	2227	G7LEH3	*M. truncatula*	1E−59	ATP synthase subunit beta
AHTC10022196	421.3	3180	2213	1047	G7K0E0	*M. truncatula*	7E−68	Tar1p
AHTC10000195	400.7	63	479	759	Q43375	*A. hypogaea*	3E−124	Galactose-binding lectin
AHTC10046418	399.9	10	94	501	No hits found			
AHTC10020355	389.3	1596	581	414	Q9AVH2	*P. sativum*	3E−41	Putative senescence-associated protein
AHTC10003317	387.9	4	123	488	G7ZVN7	*M. truncatula*	1E−10	Beta-glucosidase
AHTC10051154	369.1	7	101	466	B9II82	*P. trichocarpa*	3E−10	Predicted protein
AHTC10002092	343.8	712	184	56	G7ZYC4	*M. truncatula*	2E−136	Cell wall-associated hydrolase
AHTC10015904	326.3	0	25	325	Q6WNU4	*G. max*	0E+00	Subtilisin-like protease
AHTC10015759	321.1	2351	1185	875	G7K0E9	*M. truncatula*	1E−33	Cytochrome P450 likeTB
AHTC10022119	301.2	736	205	90	G7ZYC4	*M. truncatula*	3E−134	Cell wall-associated hydrolase
AHTC10026674	288.5	328	16	2	E5FHZ8	*A. hypogaea*	3E−161	Late embryogenesis-abundant protein group 9 protein
AHTC10046625	268	2	49	308	No hits found			
AHTC10046509	267.7	7	47	322	Q9ZRE7	*A. thaliana*	1E−18	ATFP3
AHTC10033787	259.7	31	115	421	E9RHS6	*L. japonicus*	2E−178	9/13-hydroperoxide lyase
AHTC10004563	252.5	1901	2057	755	O24320	*P. vulgaris*	0E+00	Lipoxygenase
AHTC10008955	232.6	246	8	0	Q43437	*G. max*	3E−141	Photosystem II type I chlorophyll a-/b-binding protein
AHTC10051658	226.2	7	69	299	No hits found			
AHTC10004755	218.8	82	588	223	A2Q4Q3	*M. truncatula*	3E−129	Polyphenol oxidase

To provide deep insights into peanut aerial and subterranean young pod transcriptomes, we used SSAHA2 v2.5.3 tool to align every individual library's data against the reference transcript sequence. A total of 99% of the sequencing reads could be mapped to the reference transcriptome (Figure 3a). Approximately, 73.79%, 76.30% and 75.17% of reads for AP, SP1 and SP2, respectively, mapped uniquely to the reference transcriptome assembly and 23-26% of reads were filtered as multiple-mapped and ignored in subsequent analyses. The unmapped reads accounted for <1% in each individual data set, indicating the reference transcriptome assembly could cover almost all transcripts of aerial and subterranean young pod transcriptomes. The number of reads mapped to a TAC ranged from 1 to 33 286 with a median of 10.5 for AP, 10.9 for SP1 and 9.1 for SP2 (Figure 3b and Table S1). A total of 74 974 TACs were detected by at least one sequence read in three libraries and 49 632, 49 952 and 50 494 TACs were identified in AP, SP1 and SP2, respectively (Figure 3c). Although the three libraries, AP SP1 and SP2, were pooled from different developmental stages and environments (light and dark), the number of TACs among the three libraries was almost equivalent and they also shared roughly 28 000 TACs, accounting for approximately 55% of expressed genes in each library. Approximately, 66% (50 686) of TACs were supported by expression data in at least two libraries. A comparison between aerial and subterranean transcriptomes indicated that 84% of genes that were expressed in the aerial pod library (AP) were also detected in either of the two subterranean young pod libraries (SP1 and SP2) (Figure 3c). A gradient of shared gene expression was also detected among three pairs of libraries: AP and SP1 (7070) > SP1 and SP2 (6632) > AP and SP2 (5938). We measured gene expression levels in reads per kilobase of gene per million reads (RPKM) (Mortazavi *et al.*, 2008). Based on this analysis, the gene expression levels in aerial and subterranean transcriptomes were classified into five categories (rare, low, moderate, high and extremely high) (Figure 3d). The largest portion of transcripts exhibited low expression (RPKM >3–10) followed by moderate expression (RPKM >10-50). A small fraction (1.2%–1.4%) of transcripts was expressed at extremely high levels (RPKM >100). Furthermore, the average coverage of transcripts for AP, SP1 and SP2 was 15.2, 14.3 and 15.6 RPKM, respectively. Approximately, 16.74% of uniquely mapped reads were derived from 67 highly expressed genes, whereas over 50 000 genes were represented by <10% of transcripts reads (Table S1). There were clear linear relationships in the gene expression levels between AP and SP1 (R^2 = 0.9263) as well as between SP1 and SP2 (R^2 = 0.9269),

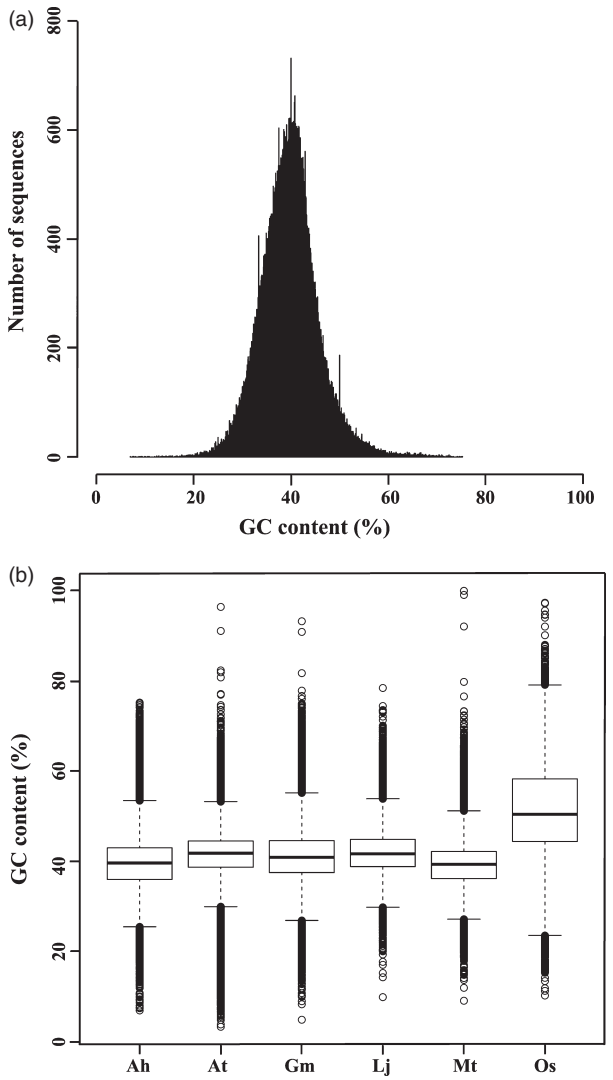

Figure 2 GC content analysis of peanut transcripts. (a) Frequency of GC content of peanut transcripts. (b) Distribution of GC content of transcripts for peanut (Ah), Arabidopsis (At), Soybean (Gm), *L. japonicus* (Lj), *M. truncatula* (Mt) and rice (Os). The boxes display the likely range of the GC content variation (the interquartile range or IQR). The upper and lower bars represent upper and lower inner fences, respectively. The circles depict outliers in the GC content distribution.

higher than between AP and SP2 (R^2 = 0.8676) (Figure 3e). This is consistent with their developmental stages, providing additional evidence that samples with one replicate could also produce reliable data. The AP library was constructed using pooled aerial developmental pods and the SP1 pooled early subterranean pods. Thus, AP shared a larger number of expressed genes with SP1 than SP2. Genes that were not expressed in subterranean pods (SP1 and SP2) but were detected in aerial pods (8892) probably represent genes that were expressed exclusively in the aerial pod development.

Functional annotation of aerial and subterranean young pod transcriptomes

To identify the putative functions of TACs related to aerial and subterranean young pod development, sequence similarity search was carried out against protein sequences available at UniProtKB database using BLASTX algorithm with an *E* value threshold of 1E−5. A total of 59 342 TACs showed significant similarity to proteins in the Uniprot protein database and matched 28 633 unique Uniprot protein accessions. Many peanut TACs also showed similarity to uncharacterized proteins defined as unknown, hypothetical and predicted proteins as well. In addition, approximately 20% of transcripts had no hits, indicating these transcripts presumably were peanut-specific genes.

To identify the functional category of the annotated TACs, gene ontology (GO) was employed to classify the transcripts annotated by known proteins. An in-house script was used to obtain GO terms for annotated transcripts. A broad range of GO categories was covered by the assigned functions of TACs (Figure 4). In total, 42 955 TACs with significant similarity to Uniprot proteins were assigned to GO terms. Of them, 38 280 TACs were assigned to the Molecular Function, followed by Biological Process (31 152) and Cellular Component (17 881). Among the various biological process categories, Catalytic Activity and Binding within Molecular Function category were dominantly represented, accounting for more than 20% of annotated TACs, respectively. Among the Biological Process category, Cellular Process and Metabolic Process accounted for 13.39 and 6.33%, respectively. Under the category of Cellular Component, 11.42 and 4.53% TACs were located into Cell and Membrane, whereas, only a few TACs were assigned to Extracellular Region, Macromolecular Complex and Cell Junction. In addition, a fraction of TACs were identified to be involved in other important biological processes such as Biological Regulation, Transporter Activity, Transcription Factor Activity, Structural Molecular Activity and Response to Stimulus and Signalling. Cell Part and Organelle represented the majority of the Cellular Component Category.

For further understanding of the biological functions and interactions of genes, pathway-based analysis was conducted based on the Kyoto Encyclopaedia of Genes and Genomes (KEGG) Pathway database, which records the networks of molecular interactions in the cells, and variants of them specific to particular organisms. TACs annotated by UnitProtKB proteins were mapped to KO database using KOBAS (for KEGG Orthology-Based Annotation System, v2.0) (Xie *et al.*, 2011). Results showed that a total of 2968 TACs were assigned to 174 KEGG pathways for both aerial and subterranean young pod transcriptomes (Table S2). Of the 174 KEGG pathways, 96% (168 pathways) were shared by aerial and subterranean transcriptomes with exception of six pathways involving only eight TACs, suggesting that most genes expressed in aerial young pods were also present in subterranean young pods even if they were subject to completely disparate conditions. The functional analysis was consistent with the analysis of the linear relationship in the gene expression levels between aerial and subterranean young pods.

Differentially expressed genes between aerial and subterranean young pods

In this study, we utilized RNA-seq to investigate the overall expression of aerial and subterranean pod transcriptomes. Differentially expressed genes among the three developmental stages were identified using an R package named 'DEGseq' (Wang *et al.*, 2010). Based on this analysis, the DEGs in AP vs SP1, AP vs SP2 and SP1 vs SP2 were identified as 1413, 2090 and 1846, respectively (Table S3). A total of 3436 TACs were identified as DEGs in at least two libraries. Among them, 171 genes were expressed differentially in all of three libraries. For

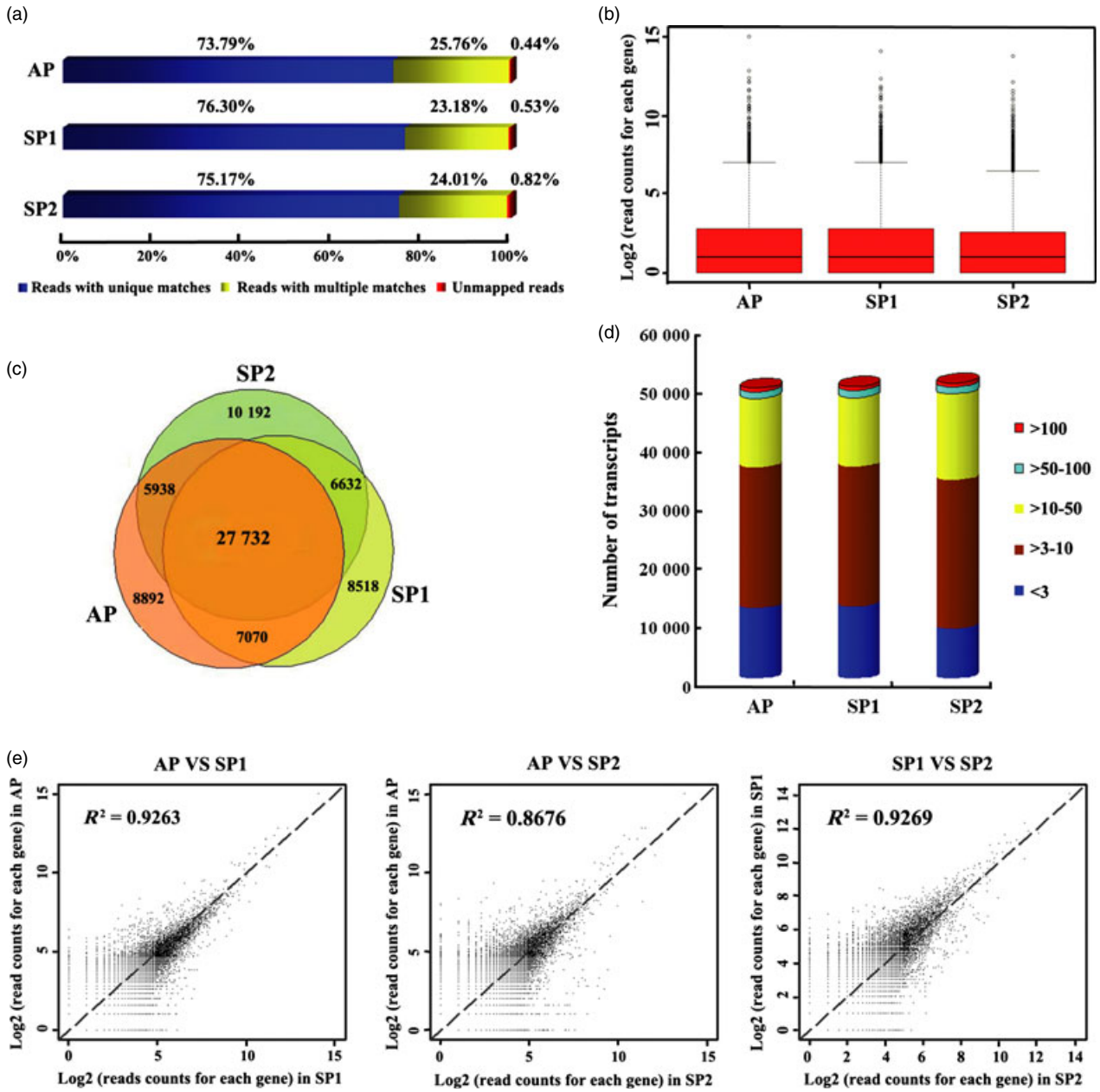

Figure 3 Summary of RNA-seq mapping data and comparison of expressed transcripts between aerial and subterranean young pod transcriptomes. (a), Overall mapping results of reads referring to reference transcript sequences. (b) Box plot for the number of reads uniquely mapped to a transcript. (c) Numbers of shared and unique genes among aerial and subterranean pods. (d), Number of transcripts with different expression levels in aerial and subterranean transcriptomes. (e), The scatter plot comparing the gene expression levels pairwise among the three libraries (between AP and SP1, between AP and SP2, as well as between SP1 and SP2, respectively).

improvement of the reliability of the DEGs and determination of the independence of read distribution within libraries, an estimation of the relative abundance defined as *R* (Stekel *et al.*, 2000) was used to identify the most highly significant differences in read abundance for each transcript among the three libraries. According to the R value, all DEGs showed more than 85% of believability ($R > 5$) except for one (Table S3). When $R > 9$, there were 2252 DEGs, which presumably represented true variation and were not false-positive results (Stekel *et al.*, 2000). The above methods normalized for the total read number and used relative abundance based on the number of mapped reads to measure gene expression levels and further identify DEGs. To correct for biases in transcript size, we also measured gene expression levels in reads per kilobase of per million mapped reads (RPKM) (Mortazavi *et al.*, 2008), which could also normalize for the total read sequences obtained in each individual library. According to the RPKM method, a total of 12 080 transcripts were differentially expressed in at least two libraries. The number was almost four times that of DEGs identified by DEGseq. The DEGs in AP vs SP1, AP vs SP2 and SP1 vs SP2 were identified as 4948, 7588 and 6769, respectively (Table S3). Approximately, 91% (3122) of DEGs ($R > 5$) identified by DEGseq could overlap with these RPKM-based DEGs (Table S3). To focus on analysing those genes which truly were differentially expressed, 2194 overlapped DEGs

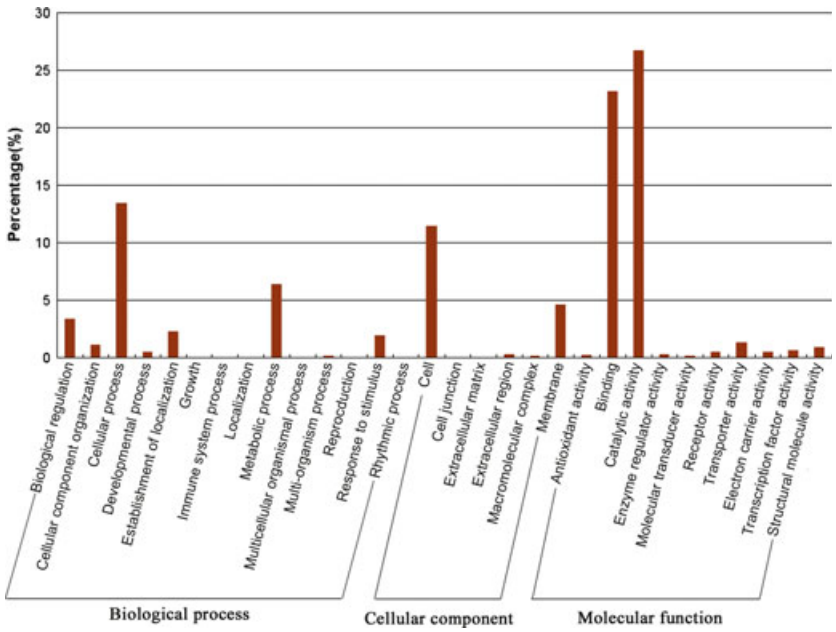

Figure 4 Gene ontology classification of peanut aerial and subterranean young pod transcriptomes.

with $R > 9$, representing a 99.0% true-positive rate (Stekel *et al.*, 2000), were used for subsequent analyses. The top thirty DEGs based on *R* value (>200) are given in Table 3. The putative functions of all DEGs are shown in Table S4. Furthermore, we identified 859 and 1068 DEGs showing up-regulation in the aerial pod transcriptome (assigned upAP-DEGs) and the subterranean pod transcriptome (SP1 and SP2, assigned upSP-DEGs), respectively. Of the upAP-DEGs, we observed that those involved in photosynthesis process and thylakoid component were notably enriched (Figure 5). Of the upSP-DEGs, we identified that those involved in cellular protein metabolic process were augmented. Detailed GO categories for upAP-DEGs and upSP-DEGs have been shown in Table S5.

To validate the results of expression profiling obtained by RNA-seq, quantitative real-time RT-PCR was performed on 20 transcripts randomly selected for differential or constitutive expression levels. A high correlation ($R^2 = 0.7310$) was found between RNA-seq and qRT-PCR (Figure 6a). The genes encoded proteins involved in photosynthesis, lectin, protein kinase, Lipoxygenase and HSP70. Of the selected genes, 65% exhibited almost similar expression profiles as determined from their respective RNA-seq data.

Photosynthesis is significantly up-regulated and enriched in aerial young pod

All AP-preferred transcripts were analysed using KOBAS to identify the metabolic pathways in which they function. KOBAS mapped 494 AP-specific transcripts to 137 pathways. Among them, five pathways were significantly up-regulated ($P < 0.05$) in aerial young pod using a *P*-value based on hypergeometric distribution (Table 4). Photosynthesis–antenna proteins pathway ranked number one with a *P*-value of 9.93E−7, followed by photosynthesis with a *P*-value of 1.51E−5. For the subterranean young pod-preferred transcripts, a total of 1174 transcripts were mapped to 159 pathways. Four of these pathways, TGF-beta signalling pathway, mismatch repair, DNA replication and fatty acid biosynthesis, were significantly identified (Table 5). Among them, the TGF-beta signalling pathway ranked number one with a *P*-value of 6.78E−3, which is involved in many cellular processes in the developing embryo including cell growth, cell differentiation, cellular homoeostasis and other cellular functions.

Pathway-based analysis showed that transcripts involved in the photosynthesis pathway were specifically expressed in aerial young pod. For further understanding of transcripts exclusively expressed in aerial and subterranean young pods, we analysed the statistically significant enrichment of specific GO terms represented in both AP and SP-preferred transcripts using Cytoscape with a plugin named BiNGO (Maere *et al.*, 2005). Like the pathway-based analysis, the photosynthesis terms in the biological process category were significantly enriched (Figure 7a). The cellular component category term thylakoid, photosynthetic membrane and photosystem terms related to photosynthesis were significantly enriched. Expression levels of four genes encoding Chlorophyll A-/B-binding proteins were validated by quantitative RT-PCR (Figure 6b). In the subterranean young pod-preferred transcripts, only the molecular term motor activity was significantly enriched (Figure 7b).

Candidate genes related to embryo abortion in aerial young pod

It is yet unknown why aerially developing pods can not swell normally compared with those that penetrate into soil. We compared developing embryos between aerial and subterranean young pods through staining of paraffin sections and found aborting embryos in aerially developing pods (Figure 1). In this study, we found two putative senescence-associated genes (AHTC10019122 and AHTC10020355) and one late embryogenesis-abundant (LEA) gene (AHTC10026674), which were significantly up-regulated in the aerial young pod (Table 3). The transcript AHTC10019122 possessed the maximum number of read counts (33 386 reads in aerial pods) among all transcripts. Sharply decreased expression levels of the LEA gene were observed in three different developmental stages: AP (328) > SP1 (16) > SP2 (2). Although we identified the DEGs based on stringent criteria, we have yet associated functional

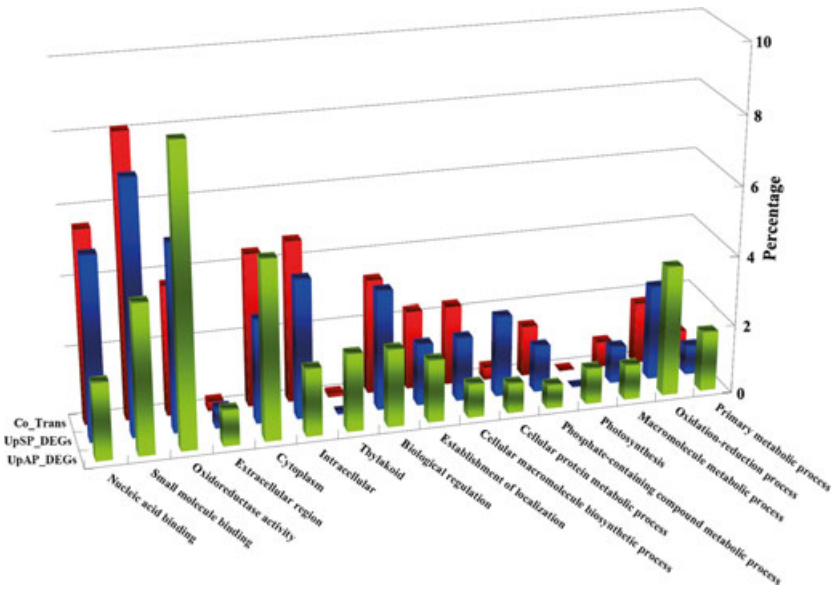

Figure 5 Gene ontology (GO) categories of differentially expressed genes compared with co-expressed transcripts on the basis of expression. UpAP-DEGs, up-regulated transcripts in AP library; upSP-DEGs, up-regulated transcripts in subterranean libraries (SP1 and SP2); co-Expr, transcripts expressed in all three libraries. To simplify the display, the figure showed only part of GO categories having difference. Detailed GO categories are showed in Table S5.

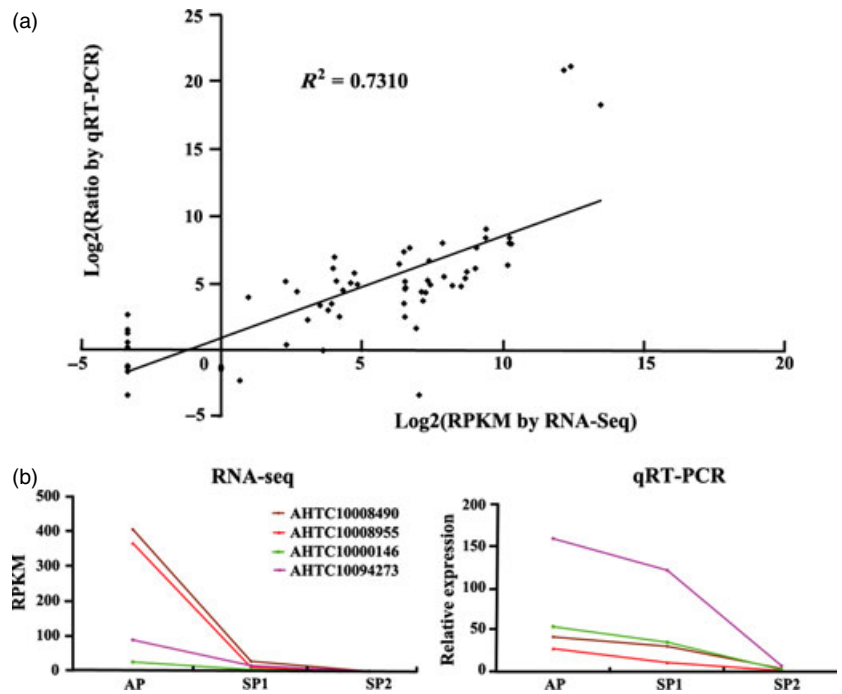

Figure 6 Validation of the RNA-seq results by qRT-PCR. (a) Comparison of expression levels measured by RNAseq and qRT-PCR for the selected 20 transcripts in three libraries (AP, SP1 and SP2). (b) Comparison of expression levels of four genes encoding Chlorophyll A-/B-binding proteins between qRT-PCR and RNAseq.

genes with physiological or morphological variations. The genes that might lead to embryo abortion and inhibit aerial pod swelling need to be confirmed in further functional genomics studies.

Discussion

Low-costs and high-throughput NGS technologies are becoming useful not only for de novo genome assembly, development of molecular markers and genome diversity studies, but also enable discovery of novel genes and investigation of gene expression patterns especially by using the RNA-seq approach (Varshney *et al.*, 2009). It is well known that in the absence of a high-quality reference genome, biological analysis based on NGS data is very challenging in the case of nonmodel organisms. This study utilizes NGS technology for RNA-seq in peanut, a species without a reference genome, for characterizing and comparing gene

Table 4 Representative aerial young pod-preferred pathways identified by KOBAS2.0

KEGG pathway	KEGG ID	No. of aerial pod-preferred transcript	No. of all transcripts	*P*-value
Photosynthesis–antenna proteins	ko00196	9	10	9.93E−07
Photosynthesis	ko00195	11	17	1.51E−05
Taurine and hypotaurine metabolism	ko00430	3	4	0.0173
Fatty acid elongation	ko00062	4	7	0.0190
Glycerolipid metabolism	ko00561	9	27	0.0298

Table 5 Representative subterranean young pod-preferred pathways identified by KOBAS2.0

KEGG Pathway	KEGG ID	No. of subterranean pod-preferred transcripts	No. of all transcripts	*P*-value
TGF-beta signalling pathway	ko04350	12	16	0.00678
Mismatch repair	ko03430	12	19	0.04623
DNA replication	ko03030	16	27	0.04650
Fatty acid biosynthesis	ko00061	5	6	0.04785

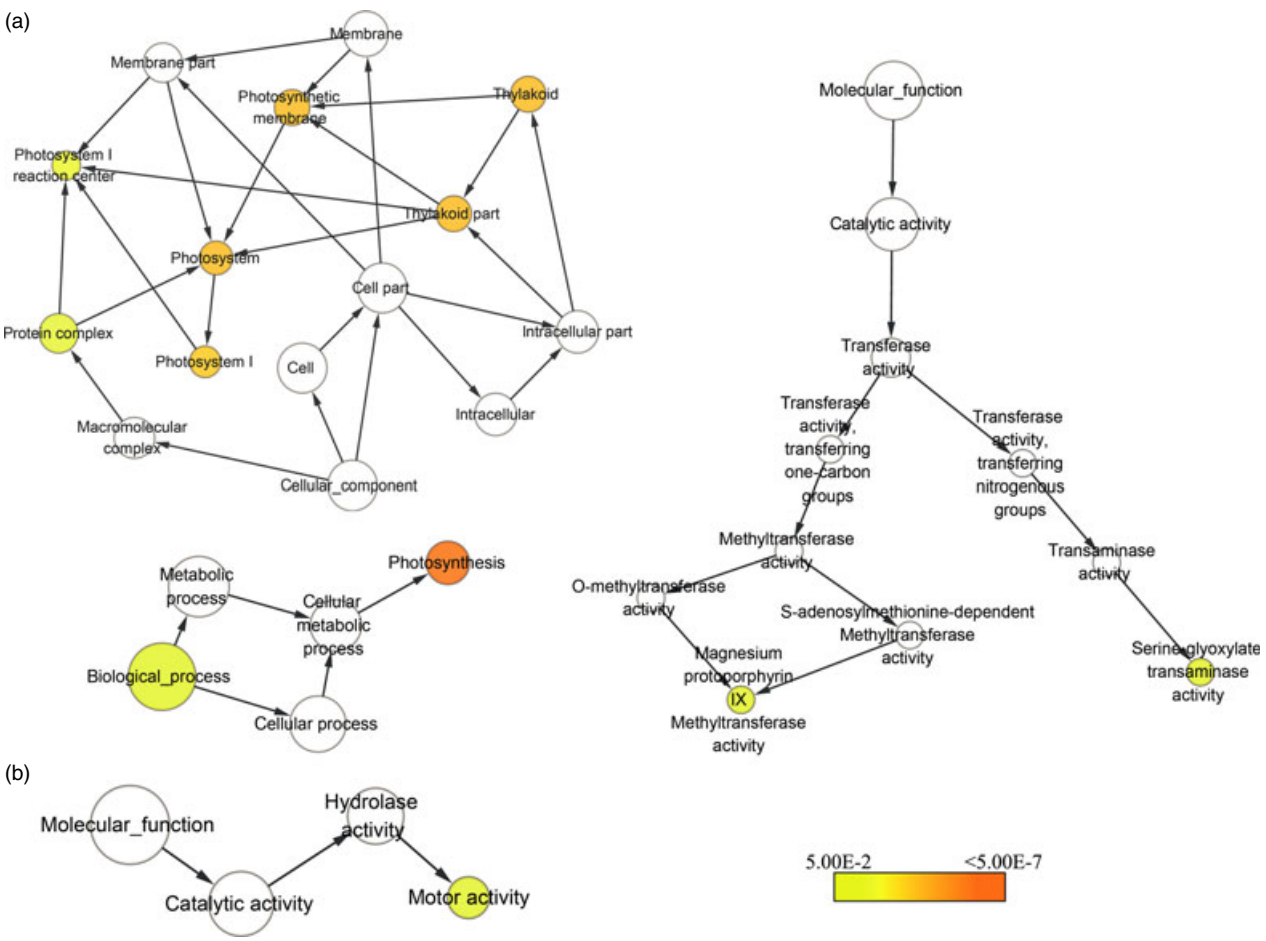

Figure 7 Significantly enriched gene ontology (GO) categories of specifically expressed transcripts in aerial (a) and subterranean (b) young pods. The transcripts showing specific expression in aerial and subterranean young pod were analysed using BinGO. Node size is proportional to the number of transcripts in each category and the significance levels are colour coded ranging from 5E−2 to <5E−7 (white, no significant difference; yellow, $P = 0.05$; orange, $P < 5E-7$).

expression profiles between aerial and subterranean pods to identify candidate genes related to embryo abortion in aerially developing pods.

It is difficult to identify an assembler that is optimal for *de novo* assembly of transcriptome data. Thus, choice of the best assembler relies on the data set and the assembly needs to be optimized. Previous studies compared the performance of a handful of the current transcriptome assemblers and indicated that TGICL and Newbler programs perform better for 454 data than other assemblers (Kumar and Blaxter, 2010; Garg *et al.*, 2011a,b). In this study, we used TGICL2.0 and Newbler (v2.6) for de novo assembly of our transcriptome data. TGICL generated a larger number of contigs with a smaller N50 relative to Newbler. In a previous study (Garg *et al.*, 2011a,b), TGICL generated a larger number of contigs than Newbler (v2.5p1), but a smaller number as compared with Newbler (v2.3), suggesting that Newbler continues to improve contiguity, but at the cost of reduced transcript diversity. It is difficult for an assembler to make a good compromise between contiguity and transcript diversity (Wang *et al.*, 2009). Therefore, more than one assembly program should be tested for assembling different types of transcriptome data from various species. Although Newbler produced a larger number of long contigs (>1 kb) than TGICL, the number of contigs having similarity to soybean orthologues with coverage of >80% was comparable for assemblies generated by TGICL and Newbler (Table 2). Furthermore, contigs yielded by TGICL hit more soybean orthologues than those from Newbler. In terms of transcriptome *de novo* assembly, TGICL performs better than Newbler. This may be because Newbler originally was designed for *de novo* genome assembly, but TGICL was developed for *de novo* transcriptome assembly especially for long EST reads (Pertea *et al.*, 2003). It is worth noting that approximately 25% of reads could not be assembled into contigs using both assemblers. Because the two programs are based on the overlap–layout–consensus strategy, the inability to assemble contigs is thus in large part related to the short overlaps. Current programs need to be improved to *de novo* assemble various lengths of transcript reads for the application of NGS to transcripts from species without a reference genome, like peanut.

In this study, we produced a peanut reference transcriptome consisting of 151 533 transcripts representing about 114.87 Mb sequence and 4.10% of the peanut genome (2800 Mb). Recently, two peanut transcriptome data sets were released (Duan *et al.*, 2012), NCBI Peanut UniGene Build #3 (UniGene #3 http://www.ncbi.nlm.nih.gov/UniGene/UGOrg.cgi?TAXID=3818) and PeanutDB (http://bioinfolab.muohio.edu/txid3818v1/), representing only 34.41 Mb (1.23% of the peanut genome) and 33.97 Mb (1.21%) sequences, respectively (Table S6). A total of 5286 TACs did not match any unigenes in both of Unigene #3 and PeanutDB. These TACs represented novel transcript sequences discovered in this study. Of these, approximately 18.09% (956/5286) of TACs had high similarity to UniProt known proteins. As peanut has no complete genome sequence and a limited number of transcripts is publicly available, little is known with respect to the number of protein-encoding genes and transcripts derived from alternative splicing in peanut. This aggravates the already difficult task that is to estimate how many genes are expressed in peanut using transcriptome data. If we assume the number of protein-encoding genes in peanut is commensurate with that in soybean (55 787), our reference transcriptome, UniGene #3 and PeanutDB would likely represent at least 53%, 38% and 34% of genes in peanut, respectively (Table S6). The peanut reference transcriptome generated here represents more peanut genes than the other two data sets. In addition, the number of TACs in this study is three times higher than the number of protein-encoding genes predicted from the complete genome of soybean. This might be due to the fact that some of TACs generated here may be represent alternatively spliced forms of the same gene locus or could suggest that more transcript sequences are needed to improve connectivity.

Like estimating the number of genes, the level of transcript coverage is also an important problem for transcriptome sequencing and is more difficult in this study due to the lack of a reference genome. Furthermore, in contrast to genome sequencing, the large dynamic range of gene expression levels aggravated the already difficult estimation due to massive redundancy for coverage of some extra highly expressed genes. At the same time, transcripts of genes with baseline or rare expression levels might be underestimated. In the present study, the coverage of peanut transcripts averaged 15 RPKM. Relatively high coverage is crucial for quality and length of TACs obtained (Garg *et al.*, 2011a,b). In addition, we only used approximately 75% of transcript reads, which uniquely mapped to the assembled reference transcriptome in this study. Approximately 23%–26% of transcript reads, which are the multiple-mapped reads, remained unanalysed. These multiple-mapped reads might be duplicated genes or segmental duplications (Lu *et al.*, 2010).

RNA-seq has revolutionized our ability to extensively investigate global changes in gene expression and has been proven to be a powerful and quantitative approach for the in-depth analysis of transcriptomes at high resolution (Varshney *et al.*, 2009). In this study, the global analysis of gene expression provided a comprehensive data set (Table S1) with each gene represented by its relative expression level for peanut young pod development under light and dark conditions. In total, we detected 2194 DEGs between aerial and subterranean young pods with a combination of various criteria. We also performed histological surveys at 2-day intervals in the aerial and subterranean young pods. The differential expression levels of the TACs could have resulted from pod development under disparate conditions, especially with and without light. The DEGs are excellent candidates for future functional genomics studies to elucidate early embryo abortion in peanut aerial young pod. But we could not rule out that other biological differences probably led to different expression levels of transcripts between two conditions. GO and pathway analyses indicated categories and pathways involved in photosynthesis were significantly up-regulated and enriched in aerial green pods. This was consistent with expression in pods of pea and soybean that were green and photosynthetically active (Weber *et al.*, 2005). However, peanut aerially developing pods will not swell normally. Although previous studies showed that pod swelling was controlled by growth regulators such as auxin, kinetin, ABA and IAA (Jacobs, 1951; Ziv and Zamski, 1975; Ziv and Kahana, 1988; Shlamovitz *et al.*, 1995), we could not detect significantly up-regulated genes involved in the related pathways. Thompson *et al.* (1985) reported that embryo growth was stimulated in darkness and inhibited by light. In the present study, paraffin sections detected that the aerial pod embryo ceased growth at early stages and finally aborted. We identified two senescence-associated which were significantly up-regulated in the aerial pod. Senescence-associated genes have been believed to regulate the senescence program preceding death and identified in various species (Buchanan-Wollaston and Ainsworth, 1997; Gepstein *et al.*, 2003). They might be potential candidate genes which lead

to embryo abortion, and in turn, inhibit pod swelling. Further studies will be conducted to provide evidence for elucidating the underlying mechanism of embryo abortion and swelling in the aerial young pod.

In conclusion, we sequenced and characterized the transcriptomes of peanut aerial and subterranean young pods. The use of this transcriptome resource has enabled us to characterize gene expression profiles and examine differential expression profiles, to identify functional genes related to embryo abortion in the aerially developing pod, and in turn, to aid in understanding the molecular mechanisms controlling peanut pod swelling in the absence of light. This study will provide a valuable transcriptomics resource for peanut, a less-studied crop, to facilitate future functional genomics studies.

Materials and methods

Plant materials and RNA extraction

Plants of 'Yueyou 7', a widespread cultivar in Southern China, were grown in the field in the summer (March-July, 2010), at the experimental station of Guangdong Academy of Agricultural Sciences (GAAS). Selfed flowers were identified with coloured plastic thread, and elongating aerial pegs were tied with coloured tags on the eighth day after flowering (DAF). After identification of pegs with coloured tags, we artificially covered one-third of tagged pegs with soil, while for the other two-third pegs, we put thick plastic membrane under tagged pegs to prevent them from penetrating into the soil. We collected aerial pods 8, 10, 12, 16 and 20 DAF and collected subterranean pods 2, 4, 6, 8, 10 and 12 days after soil penetration (DASP), corresponding to 10, 12, 14, 16, 18 and 20 DAF (Figure 2). Aerial pods were excised around 15 mm from the apex to obtain all the important components such as the ovules and meristem. For subterranean pods, the swelling part was collected for RNA isolation. To reduce stage-specific gene expression patterns and simplify the comparison of pod development under light and dark conditions, we pooled all samples from aerial pods for a aerial library(AP), and pooled samples from 2, 4 and 6 DASP as well as 8, 10 and 12 DASP for two subterranean libraries (SP1 and SP2), respectively (Figure 2).

Total RNA was isolated from aerial and subterranean pods using a modified CTAB-based protocol (Chang *et al.*, 1993) with high salt and further purified with the RNeasy Plant Mini Kit (Qiagen, Shanghai, China). RNA quality and quantity were determined using a NanoDrop ND-1000 Spectrophotometer (Nanodrop Technologies, Wilmington, DE) and verified for degradation using a 2100 Bioanalyser RNA Nanochip (Agilent, Palo Alto, CA). Pooled RNA samples of aerial pods were prepared with equivalent RNA (200 μg) from five time points. For subterranean pods, two pooled RNA samples were prepared with equivalent RNA (300 μg) from 2, 4 and 6 DASP, as well as 8, 10 and 12 DASP, respectively.

Light microscope observation

Material was fixed in FAA (50% alcohol:acetic acid:formaldehyde solution = 89:6:5) at room temperature. Samples were washed by 50% alcohol, dehydrated using an ethyl alcohol series, cleared in xylene and embedded in paraffin wax. The specimens were sectioned to a thickness of 8 μm. Sections were stained with acid fuchsin and fast green, examined and photographed using a Leica DMLB light microscope (Leica Microsystems GmbH, Wetzlar, Germany).

454 sequencing and *de novo* assembly

Transcript sequencing using Roche GS FLX Titanium platform was performed at Macrogene Inc. (www.macrogen.com). Library construction and sequencing followed the standard sequencing protocols recommended by Roche. Titration runs on the 454 were conducted on the three libraries. Bases were called by measuring the luminescence intensity from each well and comparing it with known standard (control). Sequences were screened for primer concatemers, weak signal and poly A/T tails. SeqClean (http://compbio.dfci.harvard.edu/tgi/software/) was employed to remove vector sequences against the UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html). Low-quality reads were eliminated based on the score value (reads with more than 30% of bases with quality score (*Q* value) of <20 and the remaining high-quality reads were filtered for short reads below 50 bp. Because repetitive DNA exacerbates the decreased connectivity for *de novo* assembly, repeat sequences were masked using RepeatMasker (http://www.repeatmasker.org/) before assembling.

To obtain the optimal assembly for a credible set of transcripts, two assembly programs, Newbler/GS *de novo* assembler (version 2.6; http://www.454.com) and TGICL (version 2.0; http://sourceforge.net/projects/tgicl) were used for *de novo* assembly of 454 reads generated in this study with different parameters. Reads were subjected to the Newbler assembler with the cDNA option and minimum read size of 45 bp using 16 CPUs. TGICL was used for *de novo* assembly of the data with minimum overlap of 40 (-l 40), minimum percentage identity of 90 or 95 for overlaps (-p 90 or 95) and maximum length of unmatched overhangs of 20 (-v 20). The soybean proteome data used for validation of assemblies generated by TGICL and Newbler was downloaded at http://www.phytozome.net/soybean.php. As there is not a reference genome available for peanut, in this study, we generated a reference transcriptome for peanut by merging *de novo* assemblies with the UGA Tifrunner transcriptome assembly (http://nespal.org/oziasakinslab/?page_id=1400). CD-HIT-EST (Li and Godzik, 2006) was used to remove redundancy and retain longest possible contigs. The remaining contigs constituted the final reference transcriptome referred as TACs.

Functional annotation and analysis of aerial and subterranean transcriptomes

To deduce the putative function, transcripts were subjected to BLASTX analysis against UniprotKB database with a cut-off of 1E−5. The putative functions of query transcripts were defined by the first subject hits. If the first hit is defined as uncharacterized protein, the next characterized subject hit with significance $\leq$ 1E−5 was used to define the query transcript. An in-house Perl script was used to perform gene ontologies (GO) annotation based on UniProtKB GOA file (ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz). KOBAS (KEGG Orthology Based Annotation System, v2.0) was used to identify biochemical pathways and to calculate the statistical significance of each pathway (Xie *et al.*, 2011). The UniProtKB accession numbers assigned to peanut transcripts were submitted to KOBAS for searching known pathways in the KEGG database.

Differential gene expression analysis

All reads from three libraries (AP, SP1 and SP2) were, respectively, mapped onto the nonredundant reference transcriptome to quantify the abundance of transcripts assemble using SSAHA2 (v2.5.3) (Wellcome Trust Sanger Institute, Cambridge, UK) with

default parameters except for '-best 1'. The coverage of each transcript was determined in terms of number of reads per kilobase per million (RPKM). The MA-plot-based method with random sampling model (DEGseq) proposed by Wang *et al.* (2009). The independence of DEGs was further assessed using a single statistic analysis, R value, which was developed by Stekel *et al.* (2000) for comparison gene expression from multiple cDNA libraries.

qRT-PCR analysis

Twenty pairs of primers were designed to generate amplicons for validating the RNA-seq data (Figure 6a). Aliquots of total RNA extracted for sequencing as described earlier were used for quantitative real-time PCR (qRT-PCR) experiments according to the manufacturer's instructions (Roche, Shanghai, China). All assays for a particular gene were performed in triplicate synchronously under identical conditions. All qRT-PCR experiments were run in a 25 μl volume with the Roche LightCyler 480 system (Roche). The *actin* gene was used as a reference in all qRT-PCR experiments. Relative quantification analyses of all target genes were performed using the E (Efficiency)-method from Roche Applied Science (Tellmann and Geulen, 2006). The expression level of each target gene was normalized to the level of the reference gene *actin*. The relative expression values were then validated for the RNA-seq data.

GC content analysis

GC content analysis was conducted using in-house Perl scripts. The gene indices for *Arabidopsis* (v15.0), soybean (v16.0), *Lotus Japonicus* (v6.0), *Medicago truncatula* (v11.0) and rice (v19.0) were downloaded from The Gene Index Projects (http://compbio.dfci.harvard.edu/tgi/plant.html).

Acknowledgements

This research was funded by grants from National Natural Science Foundation of China (No. 30971819 and 30900907), Natural Science Foundation of Guangdong Province (No. 10151064001000002), Science and Technology Planning Project of Guangdong Province (2011B010500019), Pearl River Science and Technology Nova of Guangzhou (No. 2011J2200035) and supported by the earmarked fund for Modern Agro-industry Technology Research System (No. nycycx-19, CARS-14). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We declare no conflict of interests.

References

Balasubramanian, V. and Yayock, J.Y. (1981) Effect of gypsum and moisture stress on growth and pod-fill of groundnut (*Arachis hypogaea* L.). *Plant Soil*, **62**, 209–219.

Bi, Y.P., Liu, W., Xia, H., Su, L., Zhao, C.Z., Wan, S.B. and Wang, X.J. (2010) EST sequencing and gene expression profiling of cultivated peanut (*Arachis hypogaea* L.). *Genome*, **53**, 832–839.

Buchanan-Wollaston, V. and Ainsworth, C. (1997) Leaf senescence in *Brassica napus*: cloning of senescence related genes by subtractive hybridisation. *Plant Mol. Biol.* **33**, 821–834.

Chang, S., Puryear, J. and Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **11**, 113–116.

Duan, X., Schmidt, E., Li, P., Lenox, D., Liu, L., Shu, C., Zhang, J. and Liang, C. (2012) PeanutDB: an integrated bioinformatics web portal for *Arachis hypogaea* transcriptomics. *BMC Plant Biol.* **12**, 94.

Feng, Q.L., Stalker, H.T., Pattee, H.E. and Isleib, T.G. (1995) *Arachis hypogaea* plant recovery through in vitro culture of peg tips. *Peanut Sci.* **22**, 129–135.

Garg, R., Patel, R.K., Jhanwar, S., Priya, P., Bhattacharjee, A., Yadav, G., Bhatia, S., Chattopadhyay, D., Tyagi, A.K. and Jain, M. (2011a) Gene discovery and tissue-specific transcriptome analysis in chickpea with massively parallel pyrosequencing and web resource development. *Plant Physiol.* **56**, 1661–1678.

Garg, R., Patel, R.K., Tyagi, A.K. and Jain, M. (2011b) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* **18**, 53–63.

Gepstein, S., Sabehi, G., Carp, M.J., Hajouj, T., Nesher, M.F., Yariv, I., Dor, C. and Bassani, M. (2003) Large-scale identification of leaf senescence-associated genes. *Plant J.* **36**, 629–642.

Golombek, S.D. and Johansen, C. (1997) Effect of soil temperature on vegetative and reproductive growth and development in three spanish genotypes of peanut (*Arachis hypogaea* L.). *Peanut Sci.* **24**, 67–72.

Guo, B., Chen, X., Dang, P., Scully, B.T., Liang, X., Holbrook, C.C., Yu, J. and Culbreath, A.K. (2008) Peanut gene expression profiling in developing seeds at different reproduction stages during Aspergillus parasiticus infection. *BMC Dev. Biol.* **8**, 12.

Guo, B., Chen, X., Hong, Y., Liang, X., Dang, P., Brenneman, T., Holbrook, C. and Culbreath, A. (2009) Analysis of gene expression profiles in leaf tissues of cultivated peanuts and development of EST-SSR markers and gene discovery. *Int. J. Plant Genomics*, **2009**, 715605.

Jacobs, W.P. (1951) Auxin relationships in an intercalary meristem. Further studies on the gynophore of Arachis hypogaea L. *Am. J. Bot.* **38**, 307–310.

Kumar, S. and Blaxter, M.L. (2010) Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics*, **11**, 571.

Lee, T.A., Ketring, D.L. and Powell, R.D. (1972) Flowering and Growth Response of Peanut Plants (*Arachis hypogaea* L. var. Starr) at Two Levels of Relative Humidity. *Plant Physiol.* **49**, 190–193.

Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S.L., Kebrom, T.H., Provart, N., Patel, R., Myers, C.R., Reidel, E.J., Turgeon, R., Liu, P., Sun, Q., Nelson, T. and Brutnell, T.P. (2010) The developmental dynamics of the maize leaf transcriptome. *Nat. Genet.* **42**, 1060–1067.

Lu, T., Lu, G., Fan, D., Zhu, C., Li, W., Zhao, Q., Feng, Q., Zhao, Y., Guo, Y., Li, W., Huang, X. and Han, B. (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.* **20**, 1238–1249.

Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nigam, S.N., Dwivedi, S.L., Ramraj, V.M. and Chandra, S. (1997) Combining ability of response to photoperiod in peanut. *Crop Sci.* **37**, 1159–1162.

Pandey, M.K., Monyo, E., Ozias-Akins, P., Liang, X., Guimaraes, P., Nigam, S.N., Upadhyaya, H.D., Janila, P., Zhang, X., Guo, B., Cook, D.R., Bertioli, D.J., Michelmore, R. and Varshney, R.K. (2012) Advances in *Arachis* genomics for peanut improvement. *Biotechnol. Adv.* **30**, 639–651.

Pattee, H.E., Stalker, H.T. and Moss, J.P. (1988) Embryo rescue in wide crosses in *Arachis*. 2. embryo development in cultured peg tips of *Arachis hypogaea*. *Ann. Bot.* **61**, 103–112.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J. and Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.

C. and Jackson, S.A. (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
Sexton, P.J., Bennett, J.M. and Boote, K.J. (1997) The effect of dry pegging zone soil on pod formation of florunner peanut. *Peanut Sci.* **24**, 19–24.
Shlamovitz, N., Ziv, M. and Zamski, E. (1995) Light, dark and growth regulator involvement in groundnut (*Arachis hypogaea* L.) pod development. *Plant Growth Regul.* **16**, 37–42.
Smith, B.W. (1950) *Arachis hypogaea*: aerial flower and subterranean fruit. *Am. J. Bot.* **37**, 802–815.
Stalker, H.T. and Wynne, J.C. (1983) Photoperiodic response of peanut species. *Peanut Sci.* **10**, 59–62.
Stekel, D.J., Git, Y. and Falciani, F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res.* **10**, 2055–2061.
Tellmann, G. and Geulen, O. (2006) LightCycler 480 Real-Time PCR system: innovative solutions for relative quantification. *Biochemica*, **4**, 16–17.
Thompson, L.K., Ziv, M. and Deitzer, G.F. (1985) Photocontrol of peanut (*Arachis hypogaea* L.) embryo and ovule development *in Vitro*. *Plant Physiol.* **78**, 370–373.
Tirumalaraju, S.V., Jain, M. and Gallo, M. (2011) Differential gene expression in roots of nematode-resistant and -susceptible peanut (*Arachis hypogaea*) cultivars in response to early stages of peanut root-knot nematode (*Meloidogyne arenaria*) parasitization. *J. Plant Physiol.* **168**, 481–492.
Underwood, C.V., Taylor, H.M. and Hoveland, C.S. (1971) Soil physical factors affecting peanut pod development. *Agronomy J.* **63**, 953–954.
Varaprasad, P.V., Craufurd, D.Q. and Summerfield, R.J. (1999) Sensitivity of peanut to timing of heat stress during reproductive development. *Crop Sci.* **539**, 1352–1357.
Varaprasad, P.V., Craufurd, D.Q. and Summerfield, R.J. (2000) Effect of high air and soil temperature on dry matter production, pod yield and yield components of groundnut. *Plant Soil*, **222**, 231–239.
Varshney, R.K., Nayak, S.N., May, G.D. and Jackson, S.A. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol.* **27**, 522–530.
Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.
Wang, L., Feng, Z., Wang, X., Wang, X. and Zhang, X. (2010) DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.
Weber, H., Borisjuk, L. and Wobus, U. (2005) Molecular physiology of legume seed development. *Annu. Rev. Plant Biol.* **56**, 253–279.
Weber, A.P., Weber, K.L., Carr, K., Wilkerson, C. and Ohlrogge, J.B. (2007) Sampling the *Arabidopsis* transcriptome with massively parallel pyrosequencing. *Plant Physiol.* **144**, 32–42.
Xie, C., Mao, X., Huang, J., Ding, Y., Wu, J., Dong, S., Kong, L., Gao, G., Li, C. Y. and Wei, L. (2011) KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic Acids Res.* **39**, W316–W322.
Zamski, E. and Ziv, M. (1975) Pod formation and its geotropic orientation in the peanut, *Arachis hypogaea* L., in relation to light and mechanical stimulus. *Ann. Bot.* **40**, 631–636.
Zhang, J., Liang, S., Duan, J., Wang, J., Chen, S., Cheng, Z., Zhang, Q., Liang, X. and Li, Y. (2012) *De novo* assembly and Characterisation of the Transcriptome during seed development, and generation of genic-SSR markers in Peanut (*Arachis hypogaea* L.). *BMC Genomics*, **13**, 90.
Ziv, M. and Kahana, O. (1988) The role of the peanut (*Araschis hypogaea*) ovular tissue in the photo-morphogenetic response of the embryo. *Plant Sci.* **57**, 159–164.
Ziv, M. and Zamski, E. (1975) Geotropic responses and pod development in gynophore explants of peanut (*Arachis hypogaea* L.) cultured *In Vitro*. *Ann. Bot.* **39**, 579–583.

Supporting information

Additional Supporting information may be found in the online version of this article:

Figure S1 Aerial and subterranean pods for sample collection.
Figure S2 Frequency distribution of 454 sequencing read lengths.
Figure S3 Distribution of contig lengths generated by TGICL and Newbler.
Figure S4 Comparison of cumulative percentages of *G. max* orthologues covered by contigs from various assemblies.
Table S1 Gene expression levels of aerial and subterranean pods in reads per million (RPM) and reads per kilobase of transcript per million mapped reads (RPKM).
Table S2 KEGG pathways represented in aerial and subterranean young pod transcriptomes.
Table S3 List of differentially expressed genes between aerial and subterranean young pods.
Table S4 Putative function of differentially expressed genes.
Table S5 Function classification of differentially expressed genes based on Gene Ontology (GO) system.
Table S6 Comparison of the reference transcriptome assembly generated in this study with the peanut Unigene Build #3 in NCBI and the PeanutDB (v1.0).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.