OPEN ACCESS Freely available online

PLOS ONE

Transcriptome Sequencing and Differential Gene Expression Analysis of Delayed Gland Morphogenesis in *Gossypium australe* during Seed Germination

Tao Tao, Liang Zhao, Yuanda Lv, Jiedan Chen, Yan Hu, Tianzhen Zhang, Baoliang Zhou*

State Key Laboratory of Crop Genetics & Germplasm Enhancement, MOE Hybrid Cotton R&D Engineering Research Center, Nanjing Agricultural University, Nanjing, Jiangsu, People's Republic of China

Abstract

The genus *Gossypium* is a globally important crop that is used to produce textiles, oil and protein. However, gossypol, which is found in cultivated cottonseed, is toxic to humans and non-ruminant animals. Efforts have been made to breed improved cultivated cotton with lower gossypol content. The delayed gland morphogenesis trait possessed by some Australian wild cotton species may enable the widespread, direct usage of cottonseed. However, the mechanisms about the delayed gland morphogenesis are still unknown. Here, we sequenced the first Australian wild cotton species (*Gossypium australe*) and a diploid cotton species (*Gossypium arboreum*) using the Illumina Hiseq 2000 RNA-seq platform to help elucidate the mechanisms underlying gossypol synthesis and gland development. Paired-end Illumina short reads were *de novo* assembled into 226,184, 213,257 and 275,434 transcripts, clustering into 61,048, 47,908 and 72,985 individual clusters with N50 lengths of 1,710 bp, 1544 BP and 1,743 bp, respectively. The clustered *Unigenes* were searched against three public protein databases (TrEMBL, SwissProt and RefSeq) and the nucleotide and protein sequences of *Gossypium raimondii* using BLASTx and BLASTn. A total of 21,987, 17,209 and 25,325 *Unigenes* were annotated. Of these, 18,766 (85.4%), 14,552 (84.6%) and 21,374 (84.4%) *Unigenes* could be assigned to GO-term classifications. We identified and analyzed 13,884 differentially expressed *Unigenes* by clustering and functional enrichment. Terpenoid-related biosynthesis pathways showed differentially regulated expression patterns between the two cotton species. Phylogenetic analysis of the terpene synthases family was also carried out to clarify the classifications of TPSs. RNA-seq data from two distinct cotton species provide comprehensive transcriptome annotation resources and global gene expression profiles during seed germination and gland and gossypol formation. These data may be used to further elucidate various mechanisms and help promote the usage of cottonseed.

Citation: Tao T, Zhao L, Lv Y, Chen J, Hu Y, et al. (2013) Transcriptome Sequencing and Differential Gene Expression Analysis of Delayed Gland Morphogenesis in *Gossypium australe* during Seed Germination. PLoS ONE 8(9): e75323. doi:10.1371/journal.pone.0075323

Editor: Jinfa Zhang, New Mexico State University, United States of America

Received May 1, 2013; **Accepted** August 14, 2013; **Published** September 20, 2013

Funding: This work was financially supported in part by grants from the National Natural Science Foundation of China (31271771, 30571184) and the Priority Academic Program Development of Jiangsu Higher Education Institutions. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: Dr. Tianzhen Zhang, co-author of this paper, is a PLOS ONE Editorial Board member. The authors confirm that this does not alter their adherence to all the PLOS ONE policies on sharing data and materials.

* E-mail: baoliangzhou@njau.edu.cn

Introduction

Cotton is a globally appreciated, remarkable economic crop, as cotton produces a natural textile fiber. In addition, worldwide cottonseed production has the potential to provide protein for half a billion people annually if cottonseed could be directly consumed as a food. However, the presence of gossypol within the pigment glands of cultivated cotton limits the usage of cottonseed due to its toxicity to humans and non-ruminant animals. Gossypol is a yellowish phenolic compound that occurs naturally in certain species of cotton plants of the family Malvaceae and contributes to the self-defense mechanisms of cotton, protecting the plant from pests, diseases and abiotic stresses [1,2]. Gossypol is synthesized in cotton roots and transported and stored within pigment glands of cotton above ground [3]. This important compound also has antitumor activity and possess contraceptive properties, which makes it unique and commercially valuable [4,5].

Many efforts have been made by geneticists and breeders to eliminate gossypol within cottonseeds. However, gossypol content is highly related to insect resistance. The glanded and glandless cotton species exhibit great differences in the amount of insect feeding [6]. Thus, breeding a high-yielding "glandless-seed" and "glanded-plant" cultivar has become an

area of interest for researchers. Interestingly, some wild diploid cotton species in Australia, such as *G. australe*, *G. bickii* and *G. sturtianum*, possess a unique characteristic, namely, that the pigment glands only appear after seed germination; thus, the dormant seeds of these species lack gossypol [7]. This distinguishing characteristic, known as the delayed gland morphogenesis trait, has the potential to enable the large-scale, direct usage of cottonseed. Various efforts have been made to introduce this unique characteristic of wild Australian cotton species into cultivated tetraploid cotton [8,9], but the cultivars with the delayed gland morphogenesis trait have not been developed by now.

Inheritance studies have been carried out to elucidate the genetics of cotton gland and gossypol formation. Various results were obtained from these studies due to the differences between diverse cotton species. Previous studies have shown that in lines of Hopi cotton, the glandless trait is controlled by recessive genes, gl_2 and gl_3 [10,11]. However, in Hai 1, the dominant gene GL_2^e is mainly responsible for this trait [12,13]. The diversity of glandless trait inheritance indicates the complexity of gland formation and regulation across different cotton species. Further studies are needed to better understand the mechanisms underlying gland development.

Terpenes comprise the largest class of natural products and participate mainly in secondary or primary metabolism in processes such as sterol and carotene biosynthesis. Plants accumulate terpenes, some of which are released for various purposes such as plant defense against herbivores, to attract pollinators and in response to stress [14]. Sesquiterpenoids are the most commonly found terpenes that accumulate within pigment glands of cotton species, including gossypol, and can be classified as phytoalexins due to their potential role in plant resistance [15,16]. Isopentenyl diphosphate (IPP) is the common precursor of all terpenes and is synthesized in plastids via the cytosol-localized mevalonic (MEV) pathway and the MEP/DOXP pathway. Geranyl diphosphate (GPP), farnesyl diphosphate (FPP) and geranylgeranyl diphosphate (GGPP) are the precursors of monoterpenes, sesquiterpenes and diterpenes, respectively [17].

Various genes associated with gossypol and glands within the terpenoid biosynthesis pathways have been cloned. The cadinene enzyme was first purified from a glandless cotton mutant by Davis et al. as a soluble hydrophobic monomer with a molecular mass of 64 to 65 kD [18]. Chen et al. first cloned and functionally characterized a (+)-δ-cadinene synthase (CDN1-XC14/U23205) from the A-genome diploid cotton *G. arboreum*. Two major subfamilies of the *Gossypium* cadinene synthase multigene family, namely CDN1-A and CDN1-C, were proposed according to sequence similarities [19-25]. The next step in gossypol biosynthesis involves hydroxylation of (+)-δ-cadinene to 8-hydroxy-(+)-δ-cadinene, which is catalyzed by (+)-δ-cadinene-8-hydroxylase, a cytochrome P450 mono-oxygenase (CYP706B1); the gene encoding this enzyme was cloned and characterized by Luo et al. [26]. Some transcription factors are important regulatory molecules involved in gland and gossypol formation, such as GaWRKY1 [27], MYC2 [28], RanBP2 zinc finger protein [29] and others, indicating that active binding events occur during gland and gossypol development.

Next-generation sequencing (NGS) technology has recently been widely employed in diverse studies to provide a comprehensive overview of the genomes and transcriptomes of certain species. Since NGS technology has the advantage of producing massive amounts of data at a low cost, deep-sequencing technology is currently undergoing rapid development. Three sequencing platforms have been employed in the majority of sequencing projects, namely Roche 454, Illumina Hiseq/Miseq and ABI SOLiD. Since the development of NGS technology, many genomes have been sequenced, including plants such as grapevine [30], tomato [31], potato [32] and others. Such studies provide large quantities of valuable information to help further elucidate complex mechanisms that occur within certain species. RNA-seq is a revolutionary tool for transcriptome profiling that uses deep-sequencing technologies. RNA-seq can be used for various purposes, such as transcript quantification, comprehensive annotation of transcriptomes, reannotation of genomes, identification of novel transcripts and alternative splicing events [33-35] and detection of polymorphisms at the transcriptome level [36,37]. RNA-seq can be performed with or without a reference genome, which makes this technique a perfect alternative for analyzing non-model species that lack fully described genomic sequences.

Recently, the genome sequence of *G. raimondii* (2n = 2x = D_5D_5 = 26), which is believed to be one of the ancestors of currently cultivated allotetraploid cotton, has been accomplished [38,39], providing cotton geneticists worldwide with a valuable resource to better explore the biological networks of this important crop. In this study, we analyzed the first wild Australian cotton species (*G. australe*), which possesses the unique delayed gland morphogenesis trait as well as *Verticillium* wilt disease- and stress-resistance characteristics [40], along with an A-genome diploid cotton species (*G. arboreum*), using the Illumina Hiseq 2000 RNA-seq platform. The paired-end (PE) reads were used for *de novo* assembly due to the differences between the three chromosome sets (A, D, G). The objective of this study was to perform a comprehensive comparison of two highly diverse cotton species during seed germination and to identify transcripts that may be important for gland and gossypol formation. The results of this study may be useful for further elucidating seed developmental mechanism, as well as the formation of glands and gossypol, at the whole-transcriptome level.

Materials and Methods

Plant Material and RNA Extraction

Plants of diploid cotton (*Gossypium arboreum* L. cv. Jianglinzhongmian and *G. australe* F. Muell) were grown in a greenhouse at Nanjing Agricultural University, China. Delinting treatment was applied to mature seeds using H_2SO_4 at a concentration of 80%. The sundried seeds were then sterilized with 70% ethanol for 30s and 30% H_2O_2 for 1 h, followed by washing with sterile water. The seed coats were removed from

the sterilized seeds by soaking the seeds in sterile water for 18 h, followed by germination in the dark at 28°C. After germination for 5 h, 15 h or 30 h, samples of *G. arboreum* L. (2n = 2x = A_2A_2 = 26) and *G. australe* F. Muell (2n = 2x = G_2G_2 = 26) were immediately frozen and stored at -70°C.

Total RNA was extracted from these six samples according to the modified CTAB-sour phenol extraction method [41]. Each RNA sample was treated with RNase-free DNase I (Takara Bio, Dalian, China) after extraction to remove residual DNA. The RNA quality and purity were assessed according to the $OD_{260/230}$ ratio and the RNA integrity number (RIN) using a Qubit® 2.0 Fluorometer (Invitrogen, Carlsbad, CA, U.S.) and an Agilent 2100 Bioanalyzer (Agilent Technologies, U.S.).

cDNA Library Preparation for Illumina Sequencing

The cDNA libraries of the six high-quality RNA samples (RIN > 8) were prepared following the manufacturer's instructions in the Illumina® TruSeq™ RNA Sample Preparation Kit (Illumina Inc. San Diego, CA, U.S.) using the Low-Throughput Protocol. Poly-T oligo-attached magnetic beads were used to purify the poly-A-containing mRNA molecules. The mRNA was fragmented into 200–500 bp pieces using divalent cations at an elevated temperature (94°C for 6 min). The cleaved RNA fragments were copied into first-strand cDNA using SuperScript II Reverse Transcriptase (Life Technology Inc., CA, U.S.) and random hexamer-primers with the following program: 25°C for 10 min, 42°C for 50 min and 70°C for 15 min. Second-strand cDNA was synthesized using DNA polymerase I and RNase H. These cDNA fragments were then end-repaired with the addition of a single 'A' base, followed by ligation of the adapters. The products were then purified following the instructions in the MinElute PCR Purification Kit (Qiagen, Düsseldorf, Germany) and eluted in 10 µL of Qiagen EB buffer. The eluted fragments were assessed by size on a 2% agarose gel to select fragments in the range of 400 bp ± 50 bp and retrieved using a MinElute Gel Extraction Kit (Qiagen, Düsseldorf, Germany). PCR of the selected fragments was performed using PCR Master Mix and Primer Cocktail in a Sample Preparation Kit (Illumina Inc.) using the following program: 98°C for 30 s; 15 cycles of 98°C for 10 s, 60°C for 30 s, 72°C for 30 s; 72°C for 5 min; hold at 4°C. The PCR products were purified using a MinElute PCR Purification Kit (Qiagen) in a final sample volume of 30 µL. The tagged cDNA libraries were loaded onto flow cell channels at a concentration of 2–4 pM and used for 2 × 100-bp paired-end sequencing on a single lane of the Illumina HiSeq 2000 Sequencing Platform (Illumina Inc., CA, U.S.). The samples were demultiplexed, and the indexed adapter paired reads were trimmed using CASAVA v1.8.2 software (Illumina Inc.).

Data Preprocessing and *De novo* Transcriptome Assembly

The raw FASTQ format data sets generated from CASAVA v1.8.2 were first assessed for quality using FASTQC v0.10.1 [42] and FASTX toolkit v0.0.13 [43]. Reads contaminated with Illumina adapters were detected and removed using Trimmomatic software (Released Version 0.22, http://www.usadellab.org/cms/index/php?page=trimmomatic). Poor-quality reads (Phred score < 20) were trimmed from both ends with SolexaQA packages v2.0 [44]; only the reads with lengths ≥ 25 bp on both sides of the paired-end format were subjected to further analysis. All sequencing data have been deposited in SRA (www.ncbi.nlm.nih.gov/sra). The accession number is SRR927415.The remaining quality paired-end reads were *de novo* assembled into transcripts with the Trinity program (Released on 2012-10-05) [45]. In-house Perl scripts were written to extract the longest transcript in each cluster as a unigene for downstream analysis. Representative extracted transcripts were then searched against human, bacterial and rRNA sequence contamination using the web-based version of DeconSeq (http://edwards.sdsu.edu/cgi-bin/deconseq/deconseq.cgi) [46] with default parameters.

Function Annotation and Classification of Assembled Transcriptomes

Annotations of the distinct *Unigenes* were performed using the BLASTx search program in the stand-alone NCBI-BLAST package v2.2.26+ [47]. The assembled contig sets were compared against Uniprot/Swissprot (released on 11-2012), Uniprot/TrEMBL (released on 11-2012) [48] and RefSeq-Plant (released on 11-2012 with plant data sets only) [49] protein databases with an expect E-value cutoff ≤ 1e-6. *Unigenes* were also searched against the recently published CDS and protein sequences within the *G. raimondii* genome project hosted by JGI (ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Graimondii) using BLASTx and BLASTn.

The BLASTx results were then combined and imported into Blast2GO software v2.6.2 [50] for gene ontology (GO) term analysis, describing biological process, molecular function and cellular component. The top 20 Blast hits with a cutoff E-value of 1e-6 and similarity cut-off of 55% were determined for GO annotation. The obtained annotations were enriched and refined using ANNEX; level 2 of the GO annotations are presented. GO-slim terms analysis was also performed using Blast2GO to obtain a broad overview of the ontology distributions. The Plant-slims developed by the *Arabidopsis* Information Resource was specifically chosen to implement the GO-slim step.

Moreover, the enzyme commission numbers (EC) of the corresponding GO annotated sequences were also obtained with an E-value cutoff of 1e-6. KEGG pathways were assigned to the assembled *Unigenes* using the online KEGG Automatic Annotation Server (KAAS, http://www.genome.jp/tools/kaas) [51]. The KEGG Ortholog assignments and pathway maps were obtained using the bidirectional best hit method (BBH) on the KAAS website.

Transcriptome Quantification

In many comparative analysis pipelines, including variant calling, isoform quantitation and differential gene expression, the first committed step is aligning the reads back to a reference genome or transcriptome. Here, a newly modified Burrows-Wheeler transform (BWT) aligner, Bowtie2 v 2.0.1 [52], was applied for this purpose. The quality trimmed paired-end reads were aligned back to the assembled transcriptome

with Bowtie2 and alignment results were converted to BAM format using SAMtools [53].

Normalizing and quantifying gene expression levels from ambiguous alignment results are statistical challenges when performing high-throughput RNA sequencing. The recently developed software package "eXpress" [54] v1.2.1 was used to accurately quantify the abundance of transcript-level sequences and to calculate the FPKM (fragments per kilobase of transcript per million mapped reads); only transcripts with an FPKM ≥ 1 were considered to be expressed.

Differential Gene Expression Analysis

Differentially expressed genes were called via edgeR package v3.0.8 [55]. The raw counts generated from the eXpress program were imported into edgeR to determine the significance of transcript-level expression. False discovery rate (FDR) was used to determine the threshold of the *P*-value in multiple tests. FDR ≤ 0.001 and the absolute value of | log2Ratio| ≥ 1 were considered to be the cutoff threshold to determine the significance of expression. GO enrichment analysis of differentially expressed genes was performed using Blast2GO software. A *P*-value cutoff value of 0.05 during the Fisher's exact test was used for GO enrichments against the annotated *Unigenes*.

Significantly regulated genes were also assessed by applying the Clustering algorithm. Hierarchical clustering was performed using Cluster v3.0 software [56]. Gene expression values were extracted from the edgeR-normalized FPKM data sets. Matrix distance for expression heatmap was calculated with Euclidean distance and complete-linkage methods after original FPKM values were log-transformed and centered. A heatmap was constructed using TreeView v1.1.6 [57] and MeV v4.8.1 [58]. The expression patterns of the self-defined clusters were plotted with R (2.15) scripts.

Phylogenetic Analysis of Terpene Synthase Genes

Terpene synthase-related genes of assembled "A" and "G" transcriptomes were predicted using the *getorf* program in the EMBOSS software package [59]. TPSs of other plant species were downloaded from NCBI. The *G. raimondii* protein data set was used to obtain the TPS protein sequences of the D genome. The hmmsearch program in HMMER3.0 [60] was used to search for amino acid sequences that contain the Pfam Terpene Synthase domains PF03936 and PF01397 [61]. MUSLE was used for multiple sequence alignments, and a Maximum Likelihood Tree was drawn with MEGA v5.1 [62]. The Java program FigTree v1.4.0 (http://tree.bio.ed.ac.uk/software/figtree/) was used to modify and generate the final phylogenetic tree.

Real-time Quantitative (qRT-PCR) Validation

The RNA sequencing samples that were isolated were also used to perform real-time quantitative (qRT-PCR) analysis. First-strand cDNA was synthesized using M-MLV Reverse Transcriptase (Promega, U.S.). Gene-specific primers were designed according to the comparison of the three assembled unigene sequences using Jalview [63], and Primer Premier 5.0 (Premier Biosoft International, Palo Alto, CA, U.S.) was applied to determine the primer sequences. *Histone3* (AF024716) was used as an internal control. The qRT-PCR was carried out using iQ SYBR Green Supremix (Bio-Rad, USA) according to the manufacturer's instructions. The thermal cycle conditions for PCR were as follows: 94°C for 3 min, 30 cycles including 94°C for 15 sec, 60°C for 30 sec and 72°C for 30 sec. The relative expression levels were calculated using the $2^{-\Delta\Delta Ct}$ method [64].

Results and Discussion

Illumina Sequencing and *De novo* Assembly

Mature seeds of *Gossypium arboreum* L. (2n = 2x = A_2A_2 = 26) and *Gossypium australe* F. Muell (2n = 2x = G_2G_2 = 26) were first delinted using highly concentrated H_2SO_4 and treated with ethanol and H_2O_2 to break dormancy. Preliminary tests were carried out, and three targeted germination stages (i.e., 5 h-G1/A1, 15 h-G2/A2 and 30 h-G3/A3) were subjected to RNA sequencing (See Material and Methods) using the Illumina HiSeq 2000 Platform. Morphological characteristics of seed germination were observed using a stereo microscope (Figure **1**). The pigment glands of *G. australe* could be observed after the seeds were germinated for more than 24 h, which is consistent with the results of previous studies [65].

Three cDNA libraries of each cotton species were bar code tagged and sequenced on one lane of the flow cell. A total of 142,880,698 (**G1 & G2 & G3**) and 252,661,798 (**A1 & A2 & A3**) raw paired-end reads with a length of 101 bp, corresponding to *G. australe* and *G. arboreum*, respectively, were generated, resulting in 35 GB and 62 GB, respectively. The raw reads were then trimmed with Illumina adapters using various techniques, and low quality bases were filtered out. The statistics of both raw and trimmed sequencing data are summarized in Table **1**. The manually selected insert library size was approximately 380 bp.

The quality trimmed reads (Q ≥ 20) were then *de novo* assembled into transcripts using Trinity, with a fixed k-mer of 25. We applied the "Reduce" option within the recently modified version of the Trinity software package to reduce redundancy in assembled transcriptomes. The cDNA libraries of three different stages of germination were pooled together for Trinity assembler to represent the whole transcriptome during germination for both *G. arboreum* and *G. australe*. With the purpose of detecting differentially expressed genes between *G. arboreum* and *G. australe* during germination, the six cDNA libraries were also assembled together as a reference transcriptome using Trinity. The three data sets, corresponding to *G. arboreum*, *G. australe* and *G. australe* and *G. arboreum*, were assembled into 226,184, 213,257 and 275,434 transcripts, respectively, clustering into 61,048, 47,908 and 72,985 individual clusters (Table **2**). The transcriptome assembly results may be redundant due to various alternative splicing events as well as misassemblies [66-69]. Therefore, we manually selected the longest transcript in each cluster as the representative based on custom Perl scripts, hoping to obtain a broad view of the three assembled transcriptomes while at the same time simplifying the data sets. The *Unigenes* were evaluated for GC content, N50 and contig length

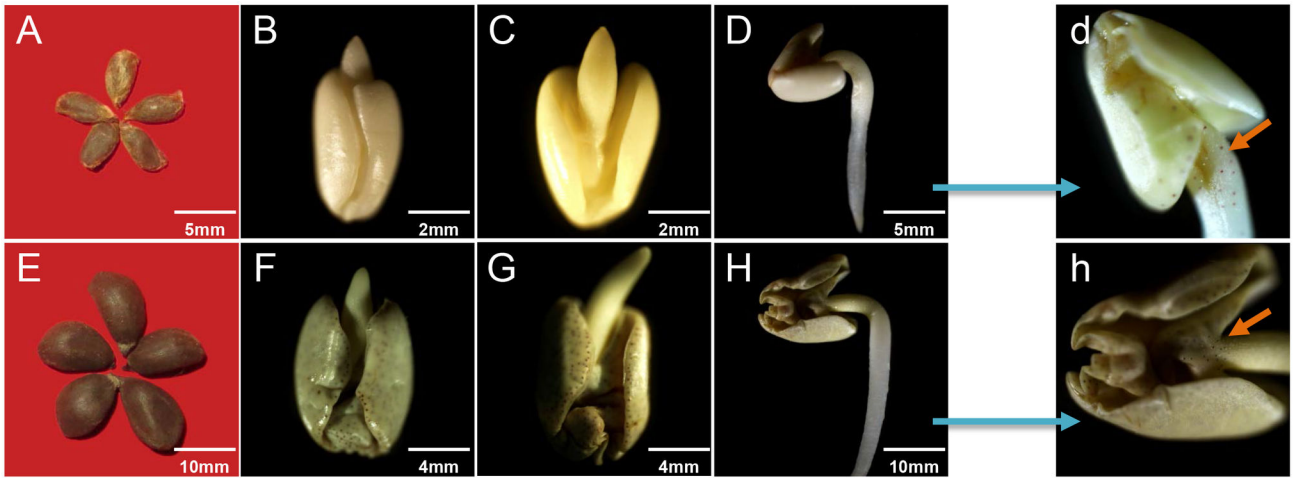

Figure 1. Stereo microscope scans of different seed germination stages. (A) and (E) are the delinted seeds of *Gossypium australe* and *Gossypium arboreum*, respectively. (B) (C) and (D) are the three germination stages of *G. australe*, i.e., 5 h, 15 h and 30 h. (F), (G) and (H) are the same three germination stages of *G. arboreum*. (d) and (h) are magnified images of (D) and (H).
doi: 10.1371/journal.pone.0075323.g001

Table 1. Statistics of transcriptome sequencing.

Library	Insert size(bp)	Raw nt (Gb)	Raw read pairs	Trimmed nt (Gb)	Trimmed read pairs (both ends≥25bp)
G1	380	6.80	13,883,940	3.03	7,292,030
G2	380	17.32	35,346,542	8.61	20,654,711
G3	380	10.88	22,209,867	5.49	13,108,013
A1	380	14.90	30,426,341	7.42	17,863,832
A2	380	32.26	65,824,214	16.36	39,236,996
A3	380	14.74	30,080,344	7.48	17,819,814

doi: 10.1371/journal.pone.0075323.t001

Table 2. Summary of *de novo* assembly.

Transcriptome assembled	Transcripts ≥ 200bp	Transcripts ≥ 500bp	No. of *Unigenes*	N50 of *Unigenes*	GC% of *Unigenes*
G. arboreum	226,184	157,539	61,048	1,710	37.67
G. australe	213,257	147,219	47,908	1,544	37.81
G. australe & G. arboreum	275,434	177,118	72,985	1,743	37.39

doi: 10.1371/journal.pone.0075323.t002

distribution based on the in-house Perl script (Table **2**). The GC contents of the three unigene sets were all approximately 37%–38%, which is considered to be normal, as cotton possesses a relatively low GC content [70-72]. The N50s of the *Unigenes* were remarkably high, achieving 1,710, 1,544 and 1,743, respectively, which may be due to the high sequencing depth. A fairly large number (34,517 of 61,084, 29,519 of 47,908 and 40,909 of 72,985) of assembled *Unigenes* were between 200 bp and 500 bp in length, indicating the presence of assembled fragments (Figure **2**).

Any rRNA sequences, as well as bacterial and human transcriptome contamination, were scanned using the web-based software DeconSeq with default parameters. A total of 13 (0.02%), 22 (0.05%) and 37 (0.05%) *Unigenes* were identified as contamination, corresponding to the three unigene sets. However, the *Unigenes* confirmed by DeconSeq were short and were likely to be assembled fragments.

Functional Annotation and Classification

We applied various approaches to validate the assemblies in order to obtain comprehensive descriptions of the assembled transcriptomes. The three assembled unigene sets (designated “A” for *G. arboreum*, “G” for *G. austral* and “A & G” for *G. arboreum* and *G. australe* for simplification, according to the genome type) were first used for homology searching against the Uniprot/Swissprot, Uniprot/TrEMBL and NCBI RefSeq Plant protein databases using the BLASTx algorithm, with an E-value cutoff of 1e-06. More than 94% of the annotated *Unigenes* in all three sets had E-values < 1e-10, indicating the reliability of the annotated results. We combined the annotation results from all three protein databases and obtained 21,987 (“A”), 17,209 (“G”) and 25,325 (“A & G”) *Unigenes* with BLASTx hits. The *Unigenes* were also searched against the CDS and protein sequences within the *G. raimondii* genome project using BLASTx and BLASTn, with an E-value cutoff of 1e-6 and 1e-10, respectively. More *Unigenes* were annotated due to sequence similarities (Figure **3**), indicating unique cotton gene models. The results show that applying diverse databases can enrich annotations, and certain species may have unique gene models that can only be annotated with closer relatives. There were also a fairly large number of sequences (39,036 of

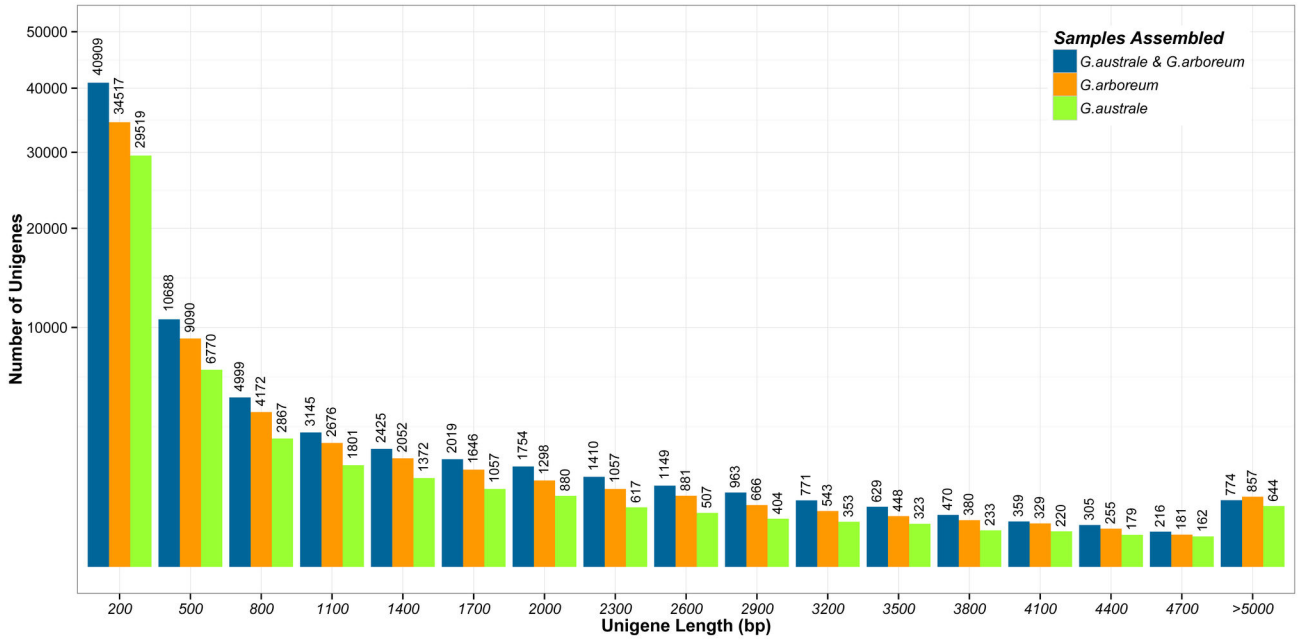

Figure 2. Length distribution of Trinity assembly for *Unigenes* of individual and combined data sets. Six data sets generated from three different seed germination stages of *G. australe* and *G. arboreum* were assembled using Trinity, length distribution of Trinity assembly from 200bp to >5000bp were presented (three data sets of *G. australe* were merged for assembly (green), as well as those of *G. arboreum* (orange), and six data sets were also assembled together (blue).

doi: 10.1371/journal.pone.0075323.g002

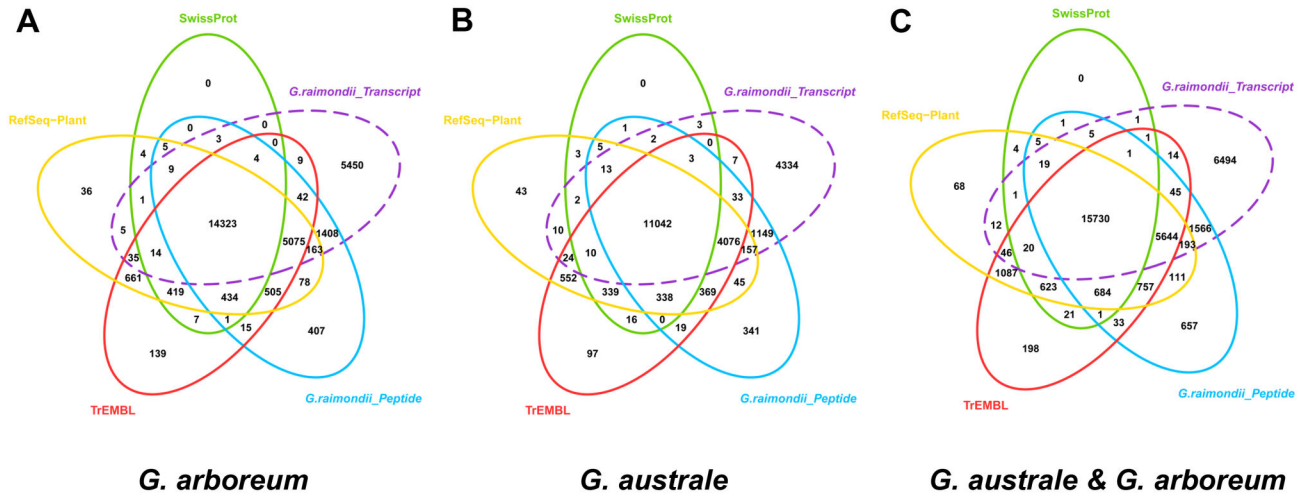

Figure 3. BLASTx and BLASTn annotation against various databases. Protein databases including Swissprot (green), TrEMBL (red) and RefSeq-Plant (yellow) were used for BLASTx annotation. The peptide (blue) and transcript (purple) sequences of *G. raimondii* (JGI) were applied for both BLASTx and BLASTn with E-value ≤ 1e-6. A, B and C represent the annotation results of *G. arboreum*, *G. australe*, *G. arboreum* & *G. australe* assemblies, respectively. The number of common annotated genes is shown in the overlapping segment of the venn diagrams.

doi: 10.1371/journal.pone.0075323.g003

61,084, 30,699 of 47,908, and 47,660 of 72,985) that were not annotated to any of the databases mentioned. These sequences may represent transcript fragments that were assembled that did not represent full-length domains, as well as noncoding RNAs or misassemblies [35,68,69].

The annotated *Unigenes* were then assigned to Gene Othology (GO) terms for functional classification. Three main categories of GO classification, i.e., biological process, molecular function and cellular component, were analyzed separately to learn as much as possible about their functional distribution. A total of 18,766 (85.4%, "A"), 14,552 (84.6%, "G") and 21,374 (84.4%, "A & G") of the annotated *Unigenes* could be assigned to one or more GO term. To simplify the functional distribution, the annotated sequences were assigned to GO-slim terms [73] of plants to obtain a "thin" version of classification. Cellular process (GO:0009987) and metabolic process (GO:0008152) within biological process, binding activity (GO:0005488) and catalytic activity (GO:0003824) within molecular function and cells (GO:0005623) and organelles (GO:0043226) within cellular component were the most representative level 2 GO terms in all three data sets (Figure **4**). All annotated sequences were then associated with enzyme codes (ECs), which returned 1,202 ("A"), 1,121 ("G") and 1,202 ("A & G") unique EC numbers.

To further identify the biological pathways that are active during seed germination, unigene sets were searched against pathway collections in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. A total of 293 ("A"), 295 ("G") and 296 ("A & G") pathways were predicted using online annotation software. The most representative pathways included biosynthesis of secondary metabolites, involving terpene backbone biosynthesis and starch and sucrose metabolism. RNA transports, as well as spliceosome pathways, were also prominent within the mapping results.

Comparative Analysis of Differential Expression during Germination

To better understand the dynamic performance between the two different transcriptomes during seed germination, abundance estimation was applied to quantify the expression levels of the six sequenced libraries. We first aligned the paired-end reads back to the combined assembled transcriptome (i.e., *G. arboreum & G. australe*) using Bowtie2. The alignment results were retrieved and pooled into the newly published quantification software "eXpress". Compared with previous quantification software such as RSEM and Cufflinks, "eXpress" excels in that it employs a sequence-bias model and specificity against *de novo* assembly workflow without the dependency of the genome background [54]. The FPKM (fragments per kilobase of transcript per million mapped reads) of each unigene were calculated and extracted from the estimation results.

To further identify genes exhibiting significant differences between the libraries, pairwise comparison was carried out and significance was confirmed using edgeR. In total, we identified 13,884 differentially expressed genes through all three germination stages and between the two cotton species, with an FDR cutoff of 0.001 and $|\log2Ratio| \geq 1$. A total of 7,146 (51.5%) DEGs were annotated using the older BLASTx procedure, leaving nearly half of the DEGs unannotated; these DEGs were considered to possibly represent novel transcripts, fragments or long noncoding RNAs. We used an FPKM cutoff of 1 to consider a gene to be expressed. The detailed relationships between expressed genes and differentially expressed genes are shown in Figure **5**. A total of 1,945 (1,144 + 801), 339 and 873 *Unigenes* were specifically regulated DEGs corresponding to the three stages of A1, A2 and A3, respectively, while 2,808 (1,918 + 890), 460 and 1,755 *Unigenes* were specifically regulated corresponding to the G1, G2 and G3 stages, respectively, and 4,700 and 3,584 *Unigenes* corresponding to *G. arboreum* and *G. australe*, respectively, were coexpressed DEGs through all three germination stages. The distributions of up- and downregulated *Unigenes* through nine pairwise comparisons are shown in Figure **6**. The number of DEGs detected in same-stage comparisons between the two cotton species was generally greater than that detected from same-species comparisons at different stages, indicating the huge differences between the two genome types and their regulatory patterns. Pigment glands appear in the third stage of *G. australe* seeds germination (G3), while seeds at the first stage (G1) do not contain glands. We also considered the middle stage of the developmental process to help elucidate the dynamic mechanisms of both seed germination and gland formation. The number of DEGs identified in the G1 vs. G3 comparison was remarkably higher than that detected between G1 and G2 and between G2 and G3, providing valuable resources for further elucidating the complicated regulatory mechanisms that occur during germination and gland development.

Clustering and Functional Enrichment of DEGs

We performed hierarchical clustering of the DEGs (Figure **7A**) using the Euclidean distance method associated with complete-linkage, hoping to further illustrate the relationships between DEGs with various expression patterns. We self-defined 16 clusters according to the cluster results, and eight main clusters, accounting for 90% of the DEGs, were plotted with expression patterns (Figure **7B**). The K3 cluster possessed the most genes (2,674); the majority of these genes (821 of 2,674) showed upregulation in *G. australe* and downregulation in *G. arboreum*. GO-enrichment was performed against all of the annotated *Unigenes* of the combined assembly. The overrepresented GO-slim terms of DEGs in biological process are shown in Table **3**. Many of the DEGs are involved in metabolism process, as well as energy and binding activities. We further analyzed the overrepresented GO functions within each main cluster; the enriched GO terms of biological process are showed in Figure **8A**. The K3 cluster contained the most overrepresented GO terms among all of the clusters. Genes involving secondary metabolic process, lipid metabolic process and generation of precursor metabolites and energy were greatly enriched in this cluster. These results suggest that not only is there a delay in gland development in *G. australe*, but the genes that exhibit opposite regulatory patterns in this species may also affect many other traits.

Pairwise comparisons between different species and between different germination stages can provide clues about the complexity of seed germination and gland formation. We therefore carried out enrichment analysis of the samples. More genes were enriched in particular GO terms than were revealed in the hierarchical clustering results, indicating the

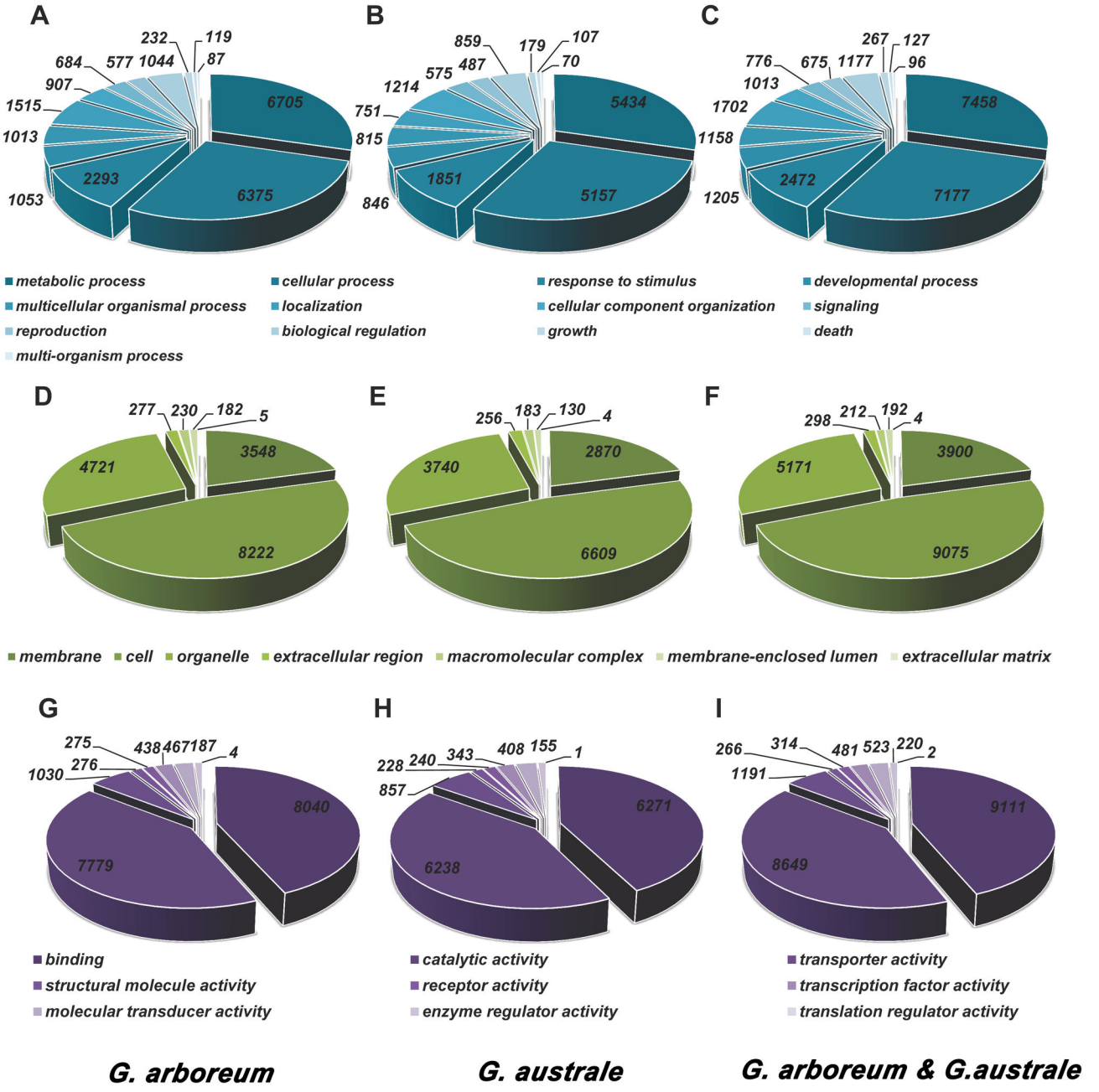

Figure 4. Distribution of GO-slim functional classification. (A) (D) and (G) represent the GO-slim classification of *G. arboreum*; (B), (E) and (H) represent *G. australe* and (C), (F) and (I) represent the combined assembly (*G. arboreum* & *G. australe*). (A) (B) and (C) are the distribution of the level 2 Biological Process within GO-slim classification; (D), (E) and (F) are the level 2 Cellular Component distribution and (G), (H) and (I) are the level 2 Molecular Function distribution. The pie charts corresponding to the detailed GO-slim classification are arranged clockwise.

doi: 10.1371/journal.pone.0075323.g004

dynamics of seed development. The main GO terms overrepresented in the G1 vs. G3 pair include photosynthesis, secondary metabolic process, generation of precursor metabolites and energy (Figure **8B**). Genes that encode enzymes for secondary metabolism, response to stimulus, lipid metabolism and carbohydrates were greatly enriched in both the cluster analysis and the pairwise comparisons. Taken together, these results reveal the dynamic processes that occur during seed germination and the development of diverse traits.

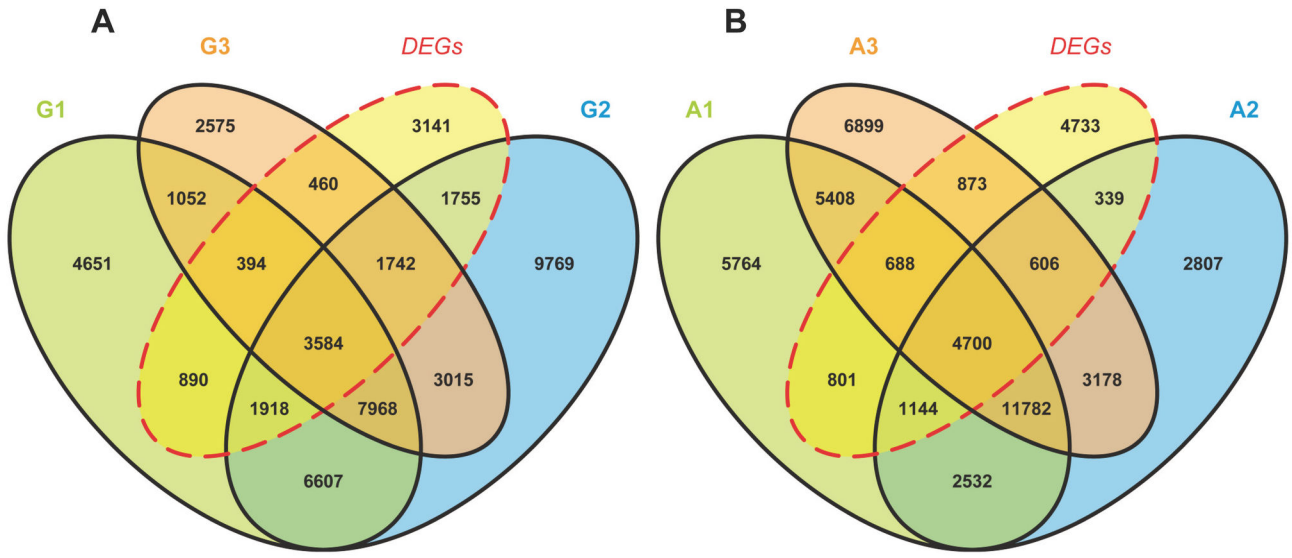

Figure 5. Venn diagrams of expressed genes and DEGs. (A) (B) are venn diagrams of *G. australe* and G. arboreum illustrating the relationship between total expressed *Unigenes* (FPKM ≥ 1) and DEGs detected by bowtie2-eXpress workflow. G1/A1, G2/A2, G3/A3 represent the number of expressed *Unigenes* in three germination stages. The individual and overlapping areas in venn diagrams represent the number of specifically expressed and co-expressed *Unigenes* between different stages.

doi: 10.1371/journal.pone.0075323.g005

To further explore the biological pathways that involve the differentially expressed genes, we performed KEGG analysis of DEGs using the online annotation software "KAAS". We detected dozens of genes related to biosynthesis of secondary metabolites, especially pathways accounting for terpenoid synthesis and starch and sucrose metabolism processes. Interestingly, when we combined the annotation results of both the whole transcriptome and the DEGs, we found that the sesquiterpenoid biosynthesis pathway, which accounts for the production of most of the gossypol composition [18-22,27], was not particularly enriched, but the expression levels of cadinene synthase genes were relatively high. These results indicate that cadinene synthases is required for the biosynthesis of gossypol. We found that the upstream biosynthesis pathways were relatively active among the DEGs, such as starch and sucrose metabolism (379 *Unigenes* assigned), the glycolysis/gluconeogenesis pathway (118 *Unigenes* assigned) and the mevalonate (MEV) and MEP/DOXP pathways (43 unigene assigned) within terpenoid backbone biosynthesis. Interestingly, we found that a fairly large number of *Unigenes* assigned to the terpenoid backbone biosynthesis pathways showed completely opposite expression patterns between *G. australe* and *G. arboreum*. Figure **9** indicates the expression levels of genes that encode enzymes in the terpenoid backbone biosynthesis pathways. In addition, we carried out hierarchical clustering of all of these genes. The first cluster is enriched in genes that exhibit opposite expression patterns in *G. australe* and *G. arboreum*; the differential expression of these genes may help explain the delayed development of gossypol and glands in *G. australe*. The results also suggest that the key genes that regulate gland formation may encode upstream regulatory factors that have a huge impact on downstream pathways. We found that only (E, E)-farnesyl diphosphate synthase (EC:2.5.1.1 2.5.1.10) is responsible for sesquiterpenoid and triterpenoid biosynthesis, which is consistent with the fact that gossypol is the product of cyclization of (E, E)-farnesyl diphosphate to (+)-delta-cadinene, which is later converted to 8-hydroxy-(+)-delta-cadinene [24,74].

Candidate Transcription Factors for Pigment Gland Formation

Transcription factors play important roles in regulation. Studies have shown that various transcription factors may be important for the formation of gossypol and the development of pigment glands [27-29]. Therefore, in this study, we were interested in describing the distribution and expression patterns of transcription factor genes in both cotton species at various stages. First, we obtained sequence data for transcription factors from 10 plant species (See Materials and Methods) from the PlantTFDB (http://planttfdb.cbi.edu.cn). The BLASTx algorithm was then performed using the assembled transcriptomes. We identified 3,725 ("A"), 2,735 ("G") and 4,185 ("A & G") possible transcription factors, 1,253 of which were differentially expressed among samples (Figure **S1A**). The largest category of differentially expressed TF genes encodes MYB/MYB-related transcription factors, followed by NAC.

The Gl_2^e gene has previously been fine mapped between two distinguished SSR markers, NAU2251b and CIR362. To better understand the mechanism of pigment gland formation, we extracted the genome sequences of *G. raimondii* between the two markers. We used FGENESH+ gene model prediction

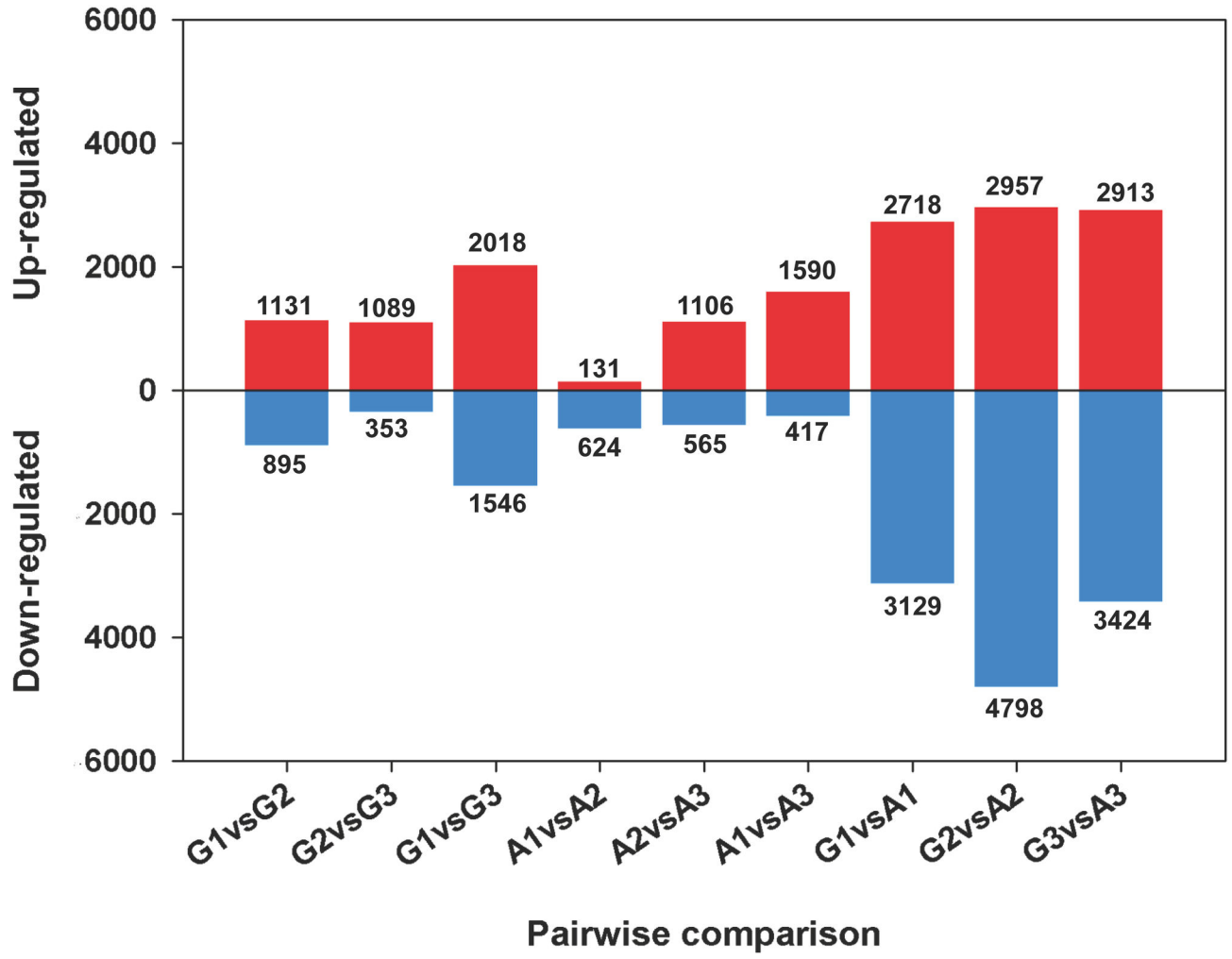

Figure 6. Bar graph of up- and downregulated genes from pairwise comparison. Three different stages of *G. australe* and *G. arboreum* seed germination were compared using the pairwise comparison method; up- and downregulated *Unigenes* are indicated.
doi: 10.1371/journal.pone.0075323.g006

software to predict the possible ORFs between the markers; this analysis returned 137 predictions. The predicted ORFs were also assigned to the transcription factor database. We detected 27 transcription factors between the markers, including ERF, MYB, GRF, NAC, Trihelix, zf-HD, bHLH, co-like, WRKY, ARR-b and others. We then used the detected TF sequences to search against the combined assembled *Unigenes* with BLASTn. This analysis returned 23 hit sequences. The regulatory patterns and expression levels of these candidate transcription factors were compared using a hierarchal clustering algorithm (Figure **S1B**).

Phylogenetic Analysis of Plant Terpene Synthases

Terpenoids, the largest family of secondary metabolic compounds, function in various plant defense and attraction reactions. Previous studies have demonstrated that (+)-delta-cadinene synthase is mainly responsible for the accumulation of sesquiterpenes in cotton. In this study, we further explored the cotton terpene synthase family. The *Getorf* program, available in the EMBOSS software package, was used to extract the predicted ORF sequences in the transcriptome assemblies of *G. arboreum* and *G. australe*. We used HMMER3.0 to further determine possible terpene synthase gene sequences. The protein sequences predicted in the *G. raimondii* genome project were also used to extract full-length terpene synthase-like sequences. Validation of the extracted sequences was carried out by searching against the Non-redundant (NR) databases in NCBI using BLASTp. Moreover, sequences of other plant species such as *Vitis vinifera, Populus trichocarpa, Selaginella moellendorffii, Physcomitrella patens, Arabidopsis thaliana, Oryza sativa* and others were also obtained from NCBI. Terpene synthases genes (TPSs) can be generally divided into seven subfamilies, i.e., TPS-a to TPS-g. Chen et al., defined another TPS-h subfamily specifically for the TPSs identified in *Selaginella moellendorffii*. TPS-e/f subfamilies are mainly found in vascular plants. The

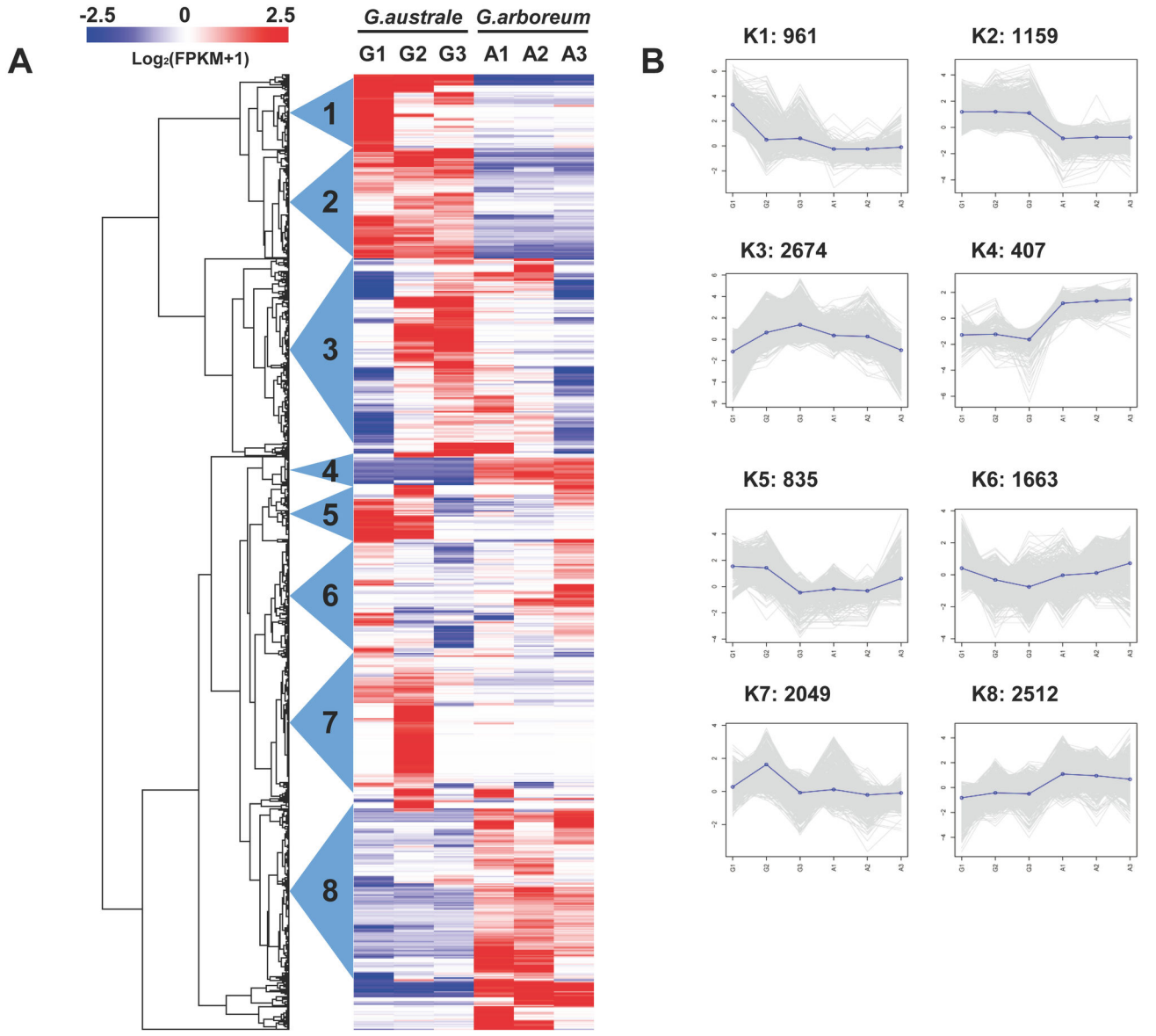

Figure 7. Hierarchical clustering of DEGs. (A) Heatmap plot of DEGs using the hierarchical clustering method; eight main clusters are shown; expression values of six individual germination stages are presented after being centered and log-transformed; decreased (blue) and increased (red) expression of DEGs are distinguished from different species and stages; (B) Expression patterns of genes in the eight main clusters, namely K1-K8, corresponding to the hierarchical heatmap.

doi: 10.1371/journal.pone.0075323.g007

TPS-d group is the only group that was not detected within the cotton TPS family; this group is only found in gymnosperms [17]. A phylogenetic tree of all aligned TPSs is shown in Figure **S2** and the distribution of TPSs identified in three species i.e. *G. raimondii*, *G. australe* and *G. arboreum* are presented in Table 4. The most dominant subfamily in *G. raimondii* is TPS-a, and 44 TPSs were identified in the genome sequences. TPS-a is mainly responsible for the synthesis of sesquiterpenes and is considered to be the most diverse TPSs among the whole family. Seven and five TPSs were identified as TPS-a in *G. arboreum* and *G, australe*, respectively. Many TPSs showed tissue specific expression and thus may explain the limited number of TPSs detected in the two species. TPS-b subfamily is mainly responsible for the synthesis of monoterpenes and ranks the second in *Gossypium* TPSs subfamilies. Interestingly, only one TPS-g like sequence can be found in all three cotton species, indicating the importance of this unique TPS gene. TPS-g subfamily is thought to be mainly responsible for the synthesis of monoterpenes and the lack of RRx $_8$W motif makes it distinguished from TPS-b subfamily.

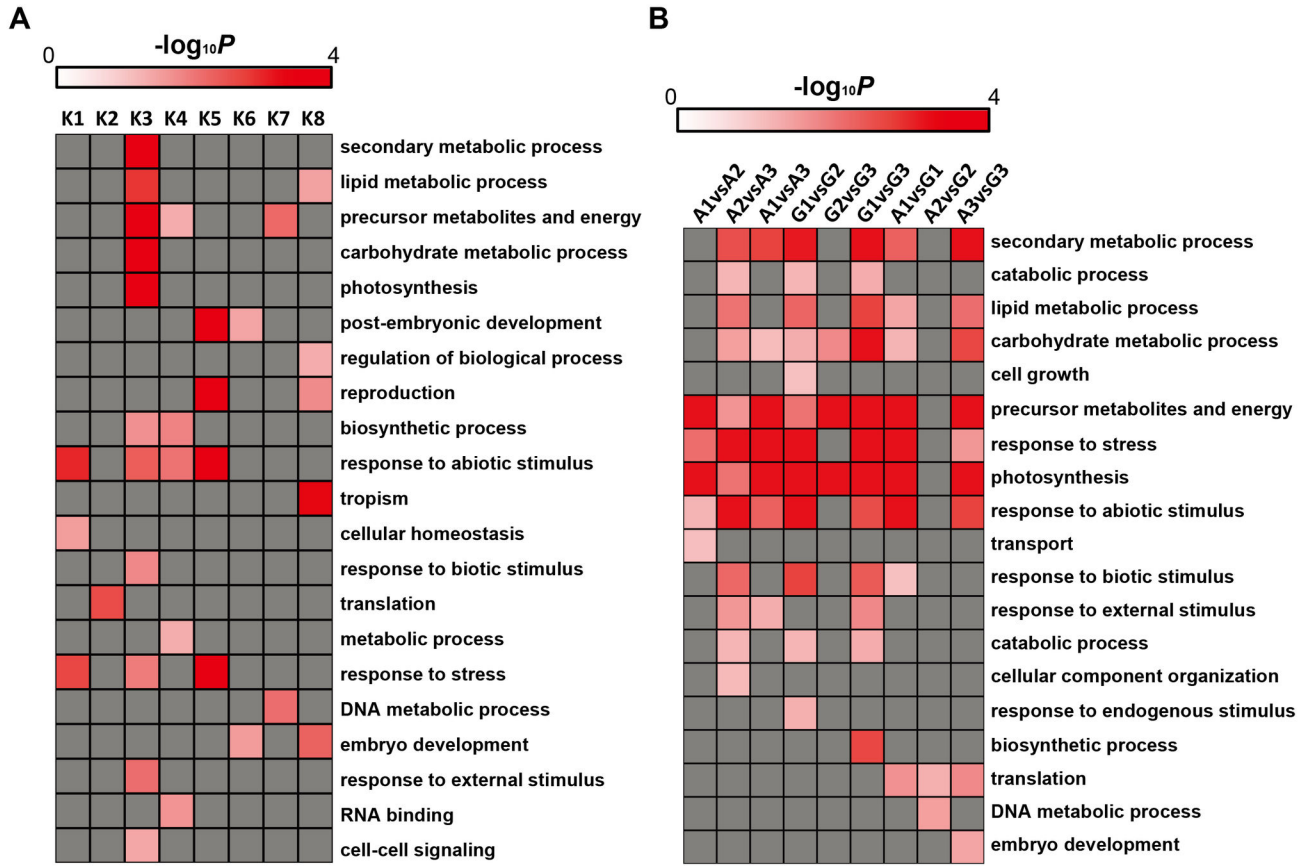

Figure 8. GO-term function enrichment analysis of different clusters, and pairwise comparisons. (A) (B) represent the GO-term enrichment of eight main clusters and nine pairwise comparisons; the significances of the most represented GO-slims in each main cluster and comparison pair are indicated using log-transformed P-value (red), the dark grey areas represented the missing values; (B) GO-term enrichment of nine pairwise comparisons.

doi: 10.1371/journal.pone.0075323.g008

Table 3. Overrepresented GO-terms of DEGs.

GO IDs	Function description of biological process	*P*-value
GO:0015979	Photosynthesis	1.74E-12
GO:0006091	Generation of precursor metabolites and energy	3.34E-06
GO:0009628	Response to abiotic stimulus	6.20E-06
GO:0006629	Lipid metabolic process	4.38E-05
GO:0008152	Metabolic process	3.73E-04
GO:0019748	Secondary metabolic process	6.87E-04
GO:0009058	Biosynthetic process	9.37E-04
GO:0006950	Response to stress	1.63E-03
GO:0005975	Carbohydrate metabolic process	2.28E-03
GO:0006412	Translation	3.45E-03
GO:0044249	Cellular biosynthetic process	3.45E-03
GO:0009059	Macromolecule biosynthetic process	3.45E-03
GO:0034645	Cellular macromolecule biosynthetic process	3.45E-03
GO:0050896	Response to stimulus	6.31E-03

doi: 10.1371/journal.pone.0075323.t003

qRT-PCR Validation

We also used RNA samples isolated for RNA sequencing to perform qRT-PCR analysis. The cadinene synthase genes are mainly responsible for sesquiterpenoid accumulation and have an impact on gossypol synthesis. We therefore employed three of these genes (i.e., comp62156_c1_seq1, comp78894_c1_seq1 and comp73507_c0_seq1), as well as two randomly chosen differentially expressed genes that encode transcription factors, for qPCR validation. The results of qPCR and RNA-seq were consistent. Both experiments showed that the expression patterns of cadinene synthase genes in *G. australe* vs. *G. arboreum* were quite different (Figure **S3**). We then analyzed the tissue-specific expression patterns of comp78894_c1_seq1 using qRT-PCR. This gene exhibited the highest expression levels among all cadinene synthase genes that were detected and may be the gene that is primarily responsible for the accumulation of gossypol in *G. australe* and *G. arboreum*. The results show that this unigene was specifically expressed in roots, which provides powerful evidence for the notion that gossypol is synthesized and

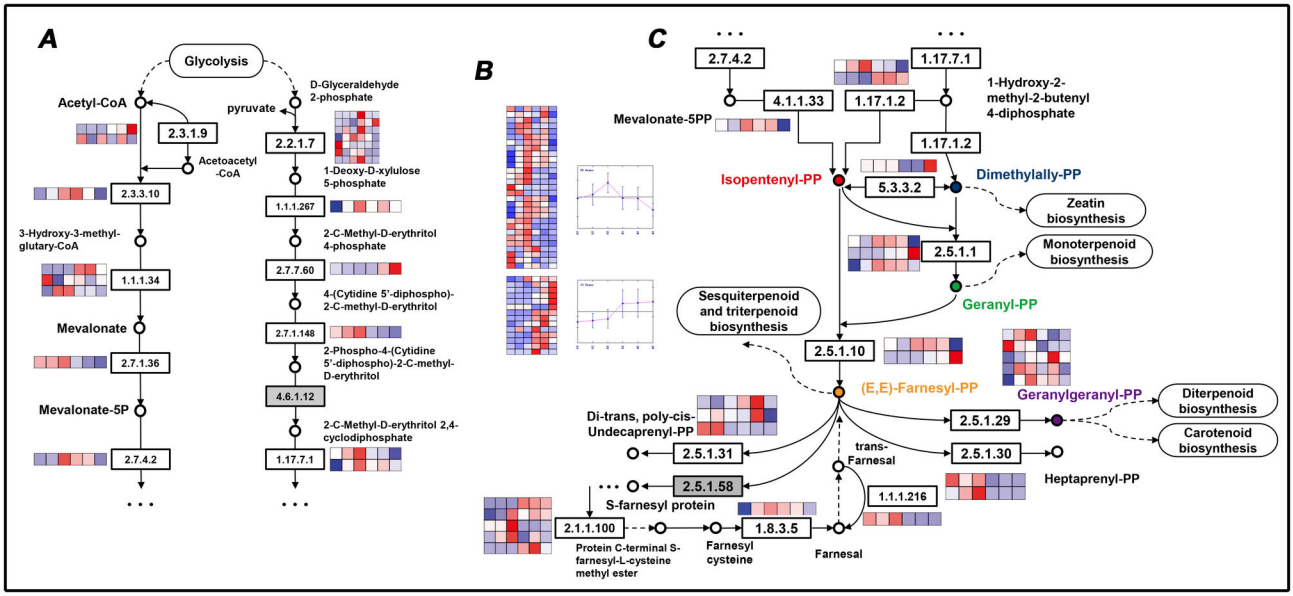

Figure 9. Differentially expressed *Unigenes* assigned to terpenoid backbone biosynthesis. (A) Expression patterns of enzymes within the MEV and MEP/DOXP pathways. (B) Two clusters of expression patterns for all enzymes assigned to terpenoid backbone biosynthesis. (C) Expression patterns of enzymes within downstream biosynthesis pathways (IPP, GPP, FPP, GGPP). The expression heatmaps are arranged in the following order: G1, G2, G3, A1, A2, A3. The log-transformed expression values range from -1 to 1.

doi: 10.1371/journal.pone.0075323.g009

accumulated in roots and is then transported to the aboveground parts of *Gossypium* [3].

There are two possible explanations for the observation that the seeds of *G. australe* lack gossypol. First, certain transportation processes may be blocked in this species, thus preventing the gossypol from getting to the ovules, resulting in the production of gossypol-free seeds. Second, as glands are the storage organs of gossypol, the delayed gland morphogenesis observed in *G. australe* may lead to the production of glandless seeds. Therefore, even if gossypol is synthesized in this species, it would not be able to accumulate at the proper destination. More interestingly, the E genome species *Gossypium stocksii* exhibits a unique characteristic, namely, dormant seeds of this species are covered with glands but contain no gossypol. This observation suggests that the relationship between glands and gossypol development is quite complex [75]. Zhu et al. examined the anatomical structures of several Australian wild cotton species using scanning electron microscopy. Their results showed that although glands were invisible to the naked eye in these species, special cells comprising lysigenous cavities, referred to as the “gland primordia”, were observed in the glandless, gossypol-free seeds of these species. The disintegration of these cells during germination leads to the appearance of glands [76,77]. Further studies will be needed to further explore the mechanisms underlying both the accumulation and transportation of gossypol and gland formation.

Table 4. Statistics of TPSs classification in three species.

TPS subfamilies	*G. raimondii*	*G. australe*	*G. arboreum*
TPS-a	44	5	7
TPS-b	25	2	2
TPS-c	6	0	1
TPS-d	0	0	0
TPS-e/f	5	1	1
TPS-g	1	1	1

doi: 10.1371/journal.pone.0075323.t004

Conclusions

The presence of glands and gossypol are two related but distinguishable characteristics of cotton. The delayed gland morphogenesis trait in *G. australe* makes this plant an ideal model for studying gland and gossypol formation. Our results show that the upstream pathways of terpenoid compound synthesis are delayed in *G. australe*, resulting in a delay in gossypol synthesis and gland appearance. We also identified candidate genes that are related to this process. The results provide evidence for key genes that regulate gossypol synthesis and gland formation. Some of the genes encoding upstream regulatory factors that exhibited large differences in expression levels may be responsible for these processes. In addition, the data provide us with powerful resources to further

elucidate the biological processes that occur during seed germination, gland formation and gossypol synthesis.

Supporting Information

Figure S1. Distribution of differentially expressed transcription factors and expression heatmap of candidate TFs. (A) Distribution of differentially expressed transcription factors during seed germination. (B) Expression heatmap of candidate transcription factors. The expression heatmaps are arranged in the following order: G1, G2, G3, A1, A2, and A3. The log-transformed expression values range from -1 to 1.
(TIF)

Figure S2. Phylogenetic analysis and subfamily classification of Terpene Synthase genes (TPSs). TPSs can be classified into seven main subfamilies, i.e., TPS-a to TPS-g. Terpene synthase genes derived from *G. raimondii*, *G. arboreum*, *G. australe* and other plant terpene synthase genes were used to generate the phylogenetic tree. The bootstrap value was set to 1000.
(TIF)

Figure S3. Relative expression values of chosen *Unigenes*. Expression values of all stages were compared to that of A1 for relative comparison purposes; the expression pattern results were consistent between qRT-PCR and RNA-seq analysis. Tissue-specific expression validation of comp78894_c1_seq1 was carried out. The relative expression levels (compared to Roots) of Roots (R), Stems (S), Leaves (L) and 10-DPA Ovules are shown.
(TIF)

Acknowledgements

We thank Mr. Yu Chen for helping with the preparation of plant materials and Mr. Kai Wang for suggestions and help with data analysis. We are also grateful to Dr. CL Brubaker (Plant Indusry of CSIRO, Australia) for providing seeds of *Gossypium australe*.

Author Contributions

Conceived and designed the experiments: BLZ TZZ. Performed the experiments: TT LZ. Analyzed the data: TT YDL JDC. Contributed reagents/materials/analysis tools: YH. Wrote the manuscript: TT.

References

1. Adams R, Geissman TA, Edwards JD (1960) Gossypol, a pigment of cottonseed. Chem Rev 60: 555-574. doi:10.1021/cr60208a002. PubMed: 13681414.
2. Scheffler JA, Romano GB (2008) Modifying gossypol in cotton (*Gossypium hirsutum* L.): a cost effective Method for small seed samples. J Cotton Sci 12: 202-209.
3. Smith F (1961) Biosynthesis of gossypol by excised cotton roots. Nature 192: 888-889. doi:10.1038/192888a0.
4. Bell AA, Stipanovic RD, O'Brien DH, Fryxell PA (1978) Sesquiterpenoid aldehyde quinones and derivatives in pigment glands of *Gossypium*. Phytochemistry 17: 1297-1305. doi:10.1016/S0031-9422(00)94578-3.
5. Blackstaffe L, Shelley MD, Fish RG (1997) Cytotoxicity of gossypol enantiomers and its quinone metabolite gossypolone in melanoma cell lines. Melanoma Res 7: 364–372. doi: 10.1097/00008390-199710000-00002. PubMed: 9429219.
6. Bottger G, Sheehan ET, Lukefahr M (1964) Relation of Gossypol Content of Cotton Plants to Insect Resistance1, 2. J Econ Entomol 57: 283-285.
7. Fryxell PA (1965) A revision of the Australian species of *Gossypium* with observations on the occurrence of Thespesia in Australia (Malvaceae). Aust J Bot 13: 71-102. doi:10.1071/BT9650071.
8. Dilday R (1986) Development of a cotton plant with glandless seeds, and glanded foliage and fruiting forms. Crop Sci 26: 639-641. doi: 10.2135/cropsci1986.0011183X002600030046x.
9. Zhu SJ, Ji D (2001) Inheritance of the delayed gland morphogenesis trait in Australian wild species of *Gossypium*. Chin Sci Bull 46: 1168-1174. doi:10.1007/BF02900595.
10. McMichael SC (1959) Hopi cotton, a source of cottonseed free of gossypol pigments. Agron J 51: 630-630. doi:10.2134/agronj1959.00021962005100100025x.
11. McMichael SC (1960) Combined effects of glandless genes gl2 and gl3 on pigment glands in the cotton plant. Agron J 52: 385-386. doi: 10.2134/agronj1960.00021962005200070005x.
12. Carvalho Ld, Vieira RdM (2000) Expression of the *Gossypium barbadense* GL2E gene in *Gossypium hirsutum* annual cotton. Rev Oleaginosas Fibrosas 4: 39-44.
13. Dong C, Ding Y, Guo W, Zhang T (2007) Fine mapping of the dominant glandless Gene Gl 2 ^{e} in Sea-island cotton (*Gossypium barbadense* L.). Chin Sci Bull 52: 3105-3109. doi:10.1007/s11434-007-0468-6.
14. Gershenzon J, Dudareva N (2007) The function of terpene natural products in the natural world. Nat Chem Biol 3: 408-414. doi:10.1038/nchembio.2007.5. PubMed: 17576428.
15. Threlfall D, Whitehead I (1991) Terpenoid phytoalexins: aspects of biosynthesis, catabolism and regulation. Ecol Chem Biochemistry Plants Terpenoids 31.
16. Chappell J (1995) Biochemistry and molecular biology of the isoprenoid biosynthetic pathway in plants. Annu Rev Plant Biol 46: 521-547. doi: 10.1146/annurev.pp.46.060195.002513.
17. Chen F, Tholl D, Bohlmann J, Pichersky E (2011) The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J 66: 212-229. doi:10.1111/j.1365-313X.2011.04520.x. PubMed: 21443633.
18. Davis EM, Tsuji J, Davis GD, Pierce ML, Essenberg M (1996) Purification of (+)-δ-cadinene synthase, a sesquiterpene cyclase from bacteria-inoculated cotton foliar tissue. Phytochemistry 41: 1047-1055. doi:10.1016/0031-9422(95)00771-7. PubMed: 8728715.
19. Chen X-Y, Chen Y, Heinstein P, Davisson VJ (1995) Cloning, expression, and characterization of (+)-δ-cadinene synthase: a catalyst for cotton phytoalexin biosynthesis. Arch Biochem Biophys 324: 255-266. doi:10.1006/abbi.1995.0038. PubMed: 8554317.
20. Chen X-Y, Wang M, Chen Y, Davisson VJ, Heinstein P (1996) Cloning and heterologous expression of a second (+)-δ-cadinene synthase from *Gossypium arboreum*. J Nat Prod 59: 944-951. doi:10.1021/np960344w. PubMed: 8904844.
21. Davis E, Chen Y, Essenberg M, Pierce M (1998) cDNA sequence of a (+)-delta-cadinene synthase gene (accession No. U88318) induced in *Gossypium hirsutum* L. by bacterial infection. Plant Physiol 116: 1192.
22. Meng Y-L, Jia J-W, Liu C-J, Liang W-Q, Heinstein P et al. (1999) Coordinated Accumulation of (+)-δ-Cadinene Synthase mRNAs and Gossypol in Developing Seeds of *Gossypium hirsutum* and a New Member of the cad 1 Family from *G. arboreum*. J Nat Prod 62: 248-252. doi:10.1021/np980314o. PubMed: 10075752.
23. Cai Y, Mo J, Zeng Y, Ren W, Xu Y et al. (2003) Cloning of cDNAs of differentially expressed genes in the development of special pigment gland of cotton by suppression subtractive hybridization. J Beijing Univ Forest 25: 6-10.
24. Martin GS, Liu J, Benedict CR, Stipanovic RD, Magill CW (2003) Reduced levels of cadinane sesquiterpenoids in cotton plants

expressing antisense (+)-δ-cadinene synthase. Phytochemistry 62: 31-38. doi:10.1016/S0031-9422(02)00432-6. PubMed: 12475616.
25. Townsend BJ, Poole A, Blake CJ, Llewellyn DJ (2005) Antisense suppression of a (+)-δ-cadinene synthase gene in cotton prevents the induction of this defense response gene during bacterial blight infection but not its constitutive expression. Plant Physiol 138: 516-528. doi: 10.1104/pp.104.056010. PubMed: 15849309.
26. Luo P, Wang YH, Wang GD, Essenberg M, Chen XY (2001) Molecular cloning and functional identification of (+)-δ-cadinene-8-hydroxylase, a cytochrome P450 mono-oxygenase (CYP706B1) of cotton sesquiterpene biosynthesis. Plant J 28: 95-104. doi:10.1046/j.1365-313X.2001.01133.x. PubMed: 11696190.
27. Xu Y-H, Wang J-W, Wang S, Wang J-Y, Chen X-Y (2004) Characterization of GaWRKY1, a cotton transcription factor that regulates the sesquiterpene synthase gene (+)-δ-cadinene synthase-A. Plant Physiol 135: 507-515. doi:10.1104/pp.104.038612. PubMed: 15133151.
28. Hong G-J, Xue X-Y, Mao Y-B, Wang L-J, Chen X-Y (2012) Arabidopsis MYC2 interacts with DELLA proteins in regulating sesquiterpene synthase gene expression. Plant Cell Available: 24: 2635-2648 PubMed: 22669881.
29. Xie Y-F, Wang B-C, Li B, Cai Y-F, Xie L et al. (2007) Construction of cDNA library of cotton mutant (Xiangmian-18) library during gland forming stage. Colloids Surf B Biointerfaces 60: 258-263. doi:10.1016/j.colsurfb.2007.06.020. PubMed: 17689935.
30. Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C et al. (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463-467. doi:10.1038/nature06148. PubMed: 17721507.
31. Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K et al. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485: 635-641. doi:10.1038/nature11119. PubMed: 22660326.
32. Xu X, Pan S, Cheng S, Zhang B, Mu D et al. (2011) Genome sequence and analysis of the tuber crop potato. Nature 475: 189–195. doi: 10.1038/nature10158. PubMed: 21743474.
33. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5: 621-628. doi:10.1038/nmeth.1226. PubMed: 18516045.
34. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57-63. doi:10.1038/nrg2484. PubMed: 19015660.
35. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511-515. doi:10.1038/nbt.1621. PubMed: 20436464.
36. Bancroft I, Morgan C, Fraser F, Higgins J, Wells R et al. (2011) Dissecting the genome of the polyploid crop oilseed rape by transcriptome sequencing. Nat Biotechnol 29: 762-766. doi:10.1038/nbt.1926. PubMed: 21804563.
37. Harper AL, Trick M, Higgins J, Fraser F, Clissold L et al. (2012) Associative transcriptomics of traits in the polyploid crop species *Brassica napus*. Nat Biotechnol, 30: 798–802. PubMed: 22820317.
38. Paterson AH, Wendel JF, Gundlach H, Guo H, Jenkins J et al. (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. Nature 492: 423-427. doi:10.1038/nature11798. PubMed: 23257886.
39. Wang K, Wang Z, Li F, Ye W, Wang J et al. (2012) The draft genome of a diploid cotton *Gossypium raimondii*. Nat Genet.
40. McFadden H, Beasley D, Brubaker CL (2004) Assessment of *Gossypium sturtianum* and *G. australe* as potential sources of Fusarium wilt resistance to cotton. Euphytica 138: 61-72. doi:10.1023/B:EUPH.0000047076.38747.81.
41. Jiang J-X, Zhang T-Z (2003) Extraction of total RNA in cotton tissue with CTAB-acidic phenolic method. J Cotton Sci 15: 2.
42. Andrews S (2010) FASTQC. A quality control tool for high throughput sequence data.
43. Gordon A (2011) FASTX-toolkit. Computer program distributed by the author. Available: http://hannonlab.cshl.edu/fastx_toolkit/index.html.
44. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11: 485. doi:10.1186/1471-2105-11-485. PubMed: 20875133.
45. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29: 644-652. doi:10.1038/nbt.1883. PubMed: 21572440.
46. Schmieder R, Edwards R (2011) Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLOS ONE 6: e17288. doi:10.1371/journal.pone.0017288. PubMed: 21408061.
47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410. doi:10.1016/S0022-2836(05)80360-2. PubMed: 2231712.
48. Uniprot, SwissProt UT (2012). http://www.ebi.ac.uk/uniprot/database/download.html.
49. RefSeq (2012). ftp://ftp.ncbi.nlm.nih.gov/refseq/.
50. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21: 3674-3676. doi: 10.1093/bioinformatics/bti610. PubMed: 16081474.
51. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35: W182-W185. doi:10.1093/nar/gkm321. PubMed: 17526522.
52. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359. doi:10.1038/nmeth.1923. PubMed: 22388286.
53. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J et al. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078-2079. doi:10.1093/bioinformatics/btp352. PubMed: 19505943.
54. Roberts A, Pachter L (2012) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods, 10: 71–3. PubMed: 23160280.
55. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139-140. doi:10.1093/bioinformatics/btp616. PubMed: 19910308.
56. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95: 14863-14868. doi:10.1073/pnas.95.25.14863. PubMed: 9843981.
57. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. Bioinformatics 20: 3246-3248. doi:10.1093/bioinformatics/bth349. PubMed: 15180930.
58. Saeed AI, Sharov V, White J, Li J, Liang W et al. (2003) TM4: a free, open-source system for microarray data management and analysis. BioTechniques 34: 374–378. PubMed: 12613259.
59. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European molecular biology open software suite. Trends Genet 16: 276-277. doi: 10.1016/S0168-9525(00)02024-2. PubMed: 10827456.
60. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29-W37. doi: 10.1093/nar/gkr367. PubMed: 21593126.
61. Bateman A, Coin L, Durbin R, Finn RD, Hollich V et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-D141. doi: 10.1093/nar/gkh121. PubMed: 14681378.
62. Tamura K, Peterson D, Peterson N, Stecher G, Nei M et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739. doi:10.1093/molbev/msr121. PubMed: 21546353.
63. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview, version 2 a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191
64. Livak KJ, Schmittgen TD (2001) Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 2< sup>− ΔΔCT</sup> Method. Methods 25: 402-408. doi:10.1006/meth.2001.1262. PubMed: 11846609.
65. Zhu S-J, Ji D-F, Wang R-H, Wang H-M (1999) Studies on the Gossypol Trend of the Cotyledon during Seed Germination and the Relationship between Gossypol and Gland Formation in the Wild Species of Gossypium in Australia. J Cotton Sci 11: 169-173.
66. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X et al. (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Res 21: 1193-1200. doi:10.1101/gr.113779.110. PubMed: 21536722.
67. Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods 8: 469-477. doi:10.1038/nmeth.1613. PubMed: 21623353.
68. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet, 12: 671–82. PubMed: 21897427.
69. Zhao Q-Y, Wang Y, Kong Y-M, Luo D, Li X et al. (2011) Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a

comparative study. BMC Bioinformatics 12: S2. doi: 10.1186/1471-2105-12-S1-S2. PubMed: 22373417.
70. Arumuganathan K, Earle E (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep 9: 208-218. doi:10.1007/BF02672069.
71. Han Z, Wang C, Song X, Guo W, Gou J et al. (2006) Characteristics, development and mapping of *Gossypium hirsutum* derived EST-SSRs in allotetraploid cotton. TAG Theoretical Appl Genet 112: 430-439. doi: 10.1007/s00122-005-0142-9. PubMed: 16341684.
72. Grover CE, Kim H, Wing RA, Paterson AH, Wendel JF (2007) Microcolinearity and genome evolution in the AdhA region of diploid and polyploid cotton (*Gossypium*). Plant J 50: 995-1006. doi:10.1111/j.1365-313X.2007.03102.x. PubMed: 17461788.
73. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M et al. (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258–D261. doi:10.1093/nar/gkh036. PubMed: 14681407.
74. Benedict CR, Liu J, Stipanovic RD (2006) The peroxidative coupling of hemigossypol to (+)-and (−)-gossypol in cottonseed extracts. Phytochemistry 67: 356-361. doi:10.1016/j.phytochem.2005.11.015. PubMed: 16403543.
75. Ding L, Zhu S-J, Hu D-Y, Ji D-F (2004) Observation on the anatomical structure of pigment glands and analysis of the gossypol content in *Gossypium stocksii*. Acta Agron Sinica 30: 100-103.
76. Brubaker C, Benson CG, Miller C, Leach DN (1996) Occurrence of terpenoid aldehydes and lysigenous cavities in the'glandless' seeds of Australian *Gossypium* species. Aust J Bot 44: 601-612. doi:10.1071/BT9960601.
77. Zhu S-J, Ji D-F, Wang R-H, Wang H-M (1998) Observation on the anatomical structure of the glandless seed and glanded plant trait in the 5 wild species of *Gossypium* in Australia. J Cotton Sci 10: 81-87.