De novo-based transcriptome profiling of male-sterile and fertile watermelon lines

Sun-Ju Rhee1*, Taehyung Kwon2*, Minseok Seo3,4, Yoon Jeong Jang1, Tae Yong Sim1, Seoae Cho4, Sang-Wook Han1*, Gung Pyo Lee1*

1 Department of Integrative Plant Science, Chung-Ang University, Ansung, Republic of Korea, 2 Department of Agricultural Biotechnology and Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea, 3 Interdisciplinary Program in Bioinformatics, Seoul National University, Kwan-ak Gu, Seoul, Republic of Korea, 4 CHO&KIM Genomics, C-1008, H Business Park, 26, Beobwon-ro 9-gil, Songpa-gu, Seoul, Republic of Korea

* These authors contributed equally to this work.

Abstract

The whole-genome sequence of watermelon (Citrullus lanatus (Thunb.) Matsum. & Nakai), a valuable horticultural crop worldwide, was released in 2013. Here, we compared a de novo-based approach (DBA) to a reference-based approach (RBA) using RNA-seq data, to aid in efforts to improve the annotation of the watermelon reference genome and to obtain biological insight into male-sterility in watermelon. We applied these techniques to available data from two watermelon lines: the male-sterile line DAH3615-MS and the male-fertile line DAH3615. Using DBA, we newly annotated 855 watermelon transcripts, and found gene functional clusters predicted to be related to stimulus responses, nucleic acid binding, trans-membrane transport, homeostasis, and Golgi/vesicles. Among the DBA-annotated transcripts, 138 de novo-exclusive differentially-expressed genes (DEDEGs) related to male sterility were detected. Out of 33 randomly selected newly annotated transcripts and DEDEGs, 32 were validated by RT-qPCR. This study demonstrates the usefulness and reliability of the de novo transcriptome assembly in watermelon, and provides new insights for researchers exploring transcriptional blueprints with regard to the male sterility.

Introduction

Watermelon [Citrullus lanatus (Thunb.) Matsum. & Nakai], a member of the Cucurbitaceae family, is an important crop worldwide, with annual production of approximately 110 million tons in 2013 (FAO, http://faostat.fao.org/). The first reference genome sequence of the East Asian watermelon was released in 2013 [1], based on next-generation sequencing (NGS) techniques. According to the genome announcement, watermelon has a diploid genome (2n = 2x = 22) of ~425 Mb, with 11 chromosomes and 23 440 transcripts. Completion of the reference genome has allowed members of the Cucurbitaceae family to be analyzed using RNA-seq. Two common RNA-seq assembly methods are widely used: de novo-based approach (DBA) and reference-based approach (RBA) [2–4]. Although both approaches can be applied to transcriptome
studies, they are selectively employed under different purposes and conditions, and often generate distinct results.

Since the process of RBA is more efficient than that of DBA, RBA is generally preferred when there is a well-established reference genome. On the other hand, use of DBA is unavoidable in efforts to identify transcripts in the absence of a reference genome. DBA can be also applied to incomplete reference genomes, for example in unknown species, because it is flexible to spliced transcription sites or unexpected structural variations [5, 6]. Novel transcripts may also be discovered by DBA, when reliable follow-up validation is performed. Although DBA is one of the best ways to identify transcripts without a reference genome, it carries a high computational burden and is prone to problems with uneven sequencing coverage derived from differential expression of genes, chimeric transcripts, and repeated sequences [7].

In view of the pros and cons of RBA and DBA, both approaches should be considered for RNA-seq analysis with reference genomes. The reference genomes of most non-model organisms, including watermelon, often exhibit missing expressed genes, trans-spliced genes, assembly errors, and deletions [8–10]. As the accuracy of RBA results depend on the completeness of the reference genome, the completeness of the reference genome used for RBA must be considered beforehand. The watermelon reference genome was recently published and watermelon transcriptome studies have employed RBA using it [11–17]. However, the draft version of the reference genome has not been updated yet, and thus, the reference genome as well as watermelon genome analysis should be continuously reviewed and developed by diverse genome research. To supplement the results of the previous RBA study [13], and given the advantages of DBA for an incomplete reference genome, we reasoned that it would be worthwhile to use DBA to improve transcriptomic analysis in watermelon.

In the crop industry, male sterility is an important trait for hybrid watermelon breeding as it renders emasculation unnecessary. There are three types of male sterility: cytoplasmic male sterility (CMS), genic male sterility (GMS), and cytoplasmic genic male sterility (CGMS) [18]. CMS is maternally inherited and controlled by the mitochondrial or plastid genome, GMS is inherited via the nuclear genome, and CGMS is induced by interaction of the mitochondria and nucleus when the restoration of fertility genes influences the CMS system.

Five watermelon male-sterile mutants have been reported to date, such as glabrous male-sterile (gms) [19–21] and male-sterile dwarf (ms-dw) [22], ms-1 [23], ms-2 [24], and ms-3 [25]. Although several attempts have been made to identify the genetic mechanisms underlying male sterility [19, 21–26], only one of these was a transcriptome study, and it was based on RBA such that unannotated transcripts were unable to be characterized [13].

In this study, we used DBA to re-analyze RNA-seq data from our previous RBA study, as a complementary approach to gain perspectives on gene annotation and to detect novel transcripts related to male sterility in watermelon. We identified several previously unreported transcripts including de novo-exclusive differentially-expressed genes (DEDEGs) from the comparison between RBA and DBA. We also characterized the functional networks between those newly annotated transcripts to provide an outline of the undiscovered transcriptome. Finally, we successfully validated the presence and differential expression of novel transcripts through RT-qPCR, demonstrating the efficacy and the legitimacy of DBA in complementing RBA to analyze the underpinnings of male sterility in watermelon.

**Materials and methods**

**Watermelon samples and RNA-seq experiments**

The two watermelon lines employed in this study were the genic male sterile (GMS) line DAH3615-MS (MS), msmtp, and the fertile, near-isogenic line DAH3615 (MF), Msms, which is
derived from the ms-1 Chinese male sterile line [23]. The plant materials and total RNA isolation for production of raw RNA-seq data in this study are described in our previous report [13]. Raw RNA-seq data used in this article are available in the GEO database under accession number GSE69073.

DBA and RBA data processing

For DBA, prior to assembly, Illumina adapter sequences were removed using Trimmomatic [27]. Clean transcripts were assembled using Trinity (r20140717) [28]. After generating transcript contigs, RNA-seq reads were mapped to the constructed transcriptome reference using Bowtie 2 [29], and RSEM [30] to align and quantify reads. Isoform and gene count matrices were generated using abundance_estimates_to_matrix implemented in Trinity. Finally, contigs were annotated using Trinotate (r20140708) (https://trinotate.github.io/), and only plant-originated transcripts (Viridiplantae) were used for downstream analyses. Among those isoform transcripts, the longest was selected as the representative sequence for each gene.

To compare DBA and RBA, RBA results from a previous study [13] were used. The watermelon reference genome (cv. 97103) version 1 from the Cucurbit Genomics Database [1] was employed. In addition, Trimmomatic [27], Tophat2 [31], and HTSeq-count [32] were used to quantify the abundance of mapped reads and to annotate watermelon genes. UniProtKB gene identification was used to compare gene lists between DBA and RBA.

Statistical analysis to identify DEGs in MF and MS

Considering our 2 x 2 factorial experimental design, analysis of variance (ANOVA) was used for RNA-seq and qPCR analysis. First, a negative binomial-assumed two-way analysis of deviance (ANODEV) model was employed for RNA-seq analysis as follows:

$$\log(E(\text{Expression}_{ij})) = \mu + \text{breed}_i + \text{tissue}_j$$

where $i =$ MF and MS lines, $j =$ floral bud and flower.

The effect of breeding line (MF or MS) and tissue on the detection of male-sterility-related genes was tested statistically using edgeR implemented in R [33]. Significance cutoff was used at the FDR adjusted $P$-value $\leq 0.01$. Likewise, a two-way ANOVA model was employed for the qPCR experiment because the value of $\Delta ct$ is commonly used to derive relative gene expression—usually satisfied according to the assumption of a normal distribution. The statistical model used was as follows:

$$-\Delta ct_y = \mu + \text{breed}_i + \text{tissue}_j + e_y, e_y \sim N(0, \sigma^2)$$

The value of $\Delta ct$ was calculated based on the difference of time it took to reach the threshold between control and targeted genes, which is negatively correlated with RNA-seq gene expression. To make the direction of $\Delta ct$ represent gene expression, the $-\Delta ct$ value was employed in the analysis.

Functional terms and network analysis of significantly enriched terms

The Database for Analysis, Validation, and Integrated Discovery (DAVID) was used to characterize specific gene lists [34, 35]. Three categories of functional terms from the GO database were employed: BP, MF, and CC. In addition, since most already-annotated genes derived from Arabidopsis thaliana, the Arabidopsis gene annotation was used as a background. Significance was considered at a $P$-value $\leq 0.001$ for newly annotated transcripts and a $P$-value $\leq 0.01$ for DEDEGs. Generally, transcripts are described in diverse biological terms; therefore,
functional terms can be classified based on M:N relationships. Significantly enriched terms were first combined, then a gene association matrix was generated (Terms x Genes). Binary values were used in this matrix, i.e., 0 means that a gene does not have a specific function, and 1 means that a gene does have a specific function. Using this matrix, correlation-based network analysis was conducted and the FDR adjusted $P$-value $< 0.01$ was considered as a significant relationship. Finally, identified relationships were visualized in network format using \textit{qgraph} package implemented in R [36]. Spring layout was also used to classify similar terms based on the strength of their connections.

**RT-qPCR for technical validation**

Primers for a total of 33 randomly selected candidates [14] newly annotated transcripts (Tables G and H in S1 File) and 19 DEDEGs (Tables I and J in S1 File) were designed using Primer3 [37]. Three biologically replicated samples of floral buds and mature flowers were collected from individual MF and MS plants, respectively. cDNA was synthesized by SuperScript RTaseIII (Thermo Fisher Scientific, USA) and oligo (dT)$_{15}$ using 1 $\mu$g total RNA isolated from each sample. Watermelon 18S rRNA was used as an internal control to normalize mRNA. Reagents used for qPCR were 10 $\mu$l PCR pre-mix, 1 $\mu$l evergreen fluorescence dye (SolGent, Korea), 1 $\mu$l cDNA, and 500 nM of each primer (except for the 18S control experiment in which 250 nM of each primer was used). PCR conditions were as follows: 95$^\circ$C for 12 min, 40 cycles at 95$^\circ$C for 10 s, and 60$^\circ$C for 30 s.

**Results**

**Summary of sequencing and \textit{de novo} transcriptome assembly statistics**

We previously produced raw RNA-seq data for four watermelon samples: flower and floral bud tissue samples of male-sterile and male-fertile lines (2 breeding lines $\times$ 2 tissues) [13]. The four samples contained read numbers ranging from 25 299 088 to 29 490 814. Approximately 80% of the total reads in all samples met Q30 quality control criteria (Table A in S1 File). Here, \textit{de novo} assembly was performed after removing poor-quality reads and adapter sequences. A total of 50 581 312 reads were used to define transcripts, resulting in 138 811 transcripts (Table B in S1 File). The average length of assembled transcripts was 1100 bp and the N50 was 2032 bp. For transcripts containing multiple isoforms that differed because of splicing events, the longest isoform was chosen to represent each gene. In all, 94 496 candidate genes were assembled; the average length of assembled genes was 773 bp and the N50 was 1327 bp (Table B in S1 File).

**Gene annotation and discovery of novel genes**

The 94 496 assembled transcripts were queried against the Swiss-Prot database using BLASTP and BLASTX [38]. Prior to further analyses, we selected only plant-originated watermelon transcripts by filtering annotated genes from species belonging to the kingdom \textit{Viridiplantae}; 11 072 and 14 398 plant-originated transcripts were annotated in DBA by BLASTP and BLASTX, respectively (E-value $< 10^{-5}$). The BLASTP search originally annotated fewer transcripts (11 072) than the BLASTX annotation (14 398), but removal of duplicated transcripts based on UniProtKB ID reduced the difference between the two BLAST annotations (BLASTP, 7135; BLASTX, 8045).

To detect novel transcripts, we compared the annotations derived from DBA and RBA after removing duplicated annotations using UniProtKB ID [39]. BLASTP analysis suggested that 6280 of 7135 nonduplicated transcripts (88.0%) were commonly identified between DBA and
RBA, and detected 855 putative novel genes (1132 transcripts with duplicated UniprotKB ID) (Fig 1A). A similar pattern was also observed in the BLASTX annotations (S1 Fig): a large
proportion of transcripts (6673, 82.9%) were common between both DBA and RBA, and 1372 genes were newly identified by DBA.

Despite the difference in employment of query sequence (BLASTP, predicted protein coding sequence; BLASTX, conceptual translation of sequence), the concordance between BLASTP and BLASTX analysis demonstrates robust annotation. Although the number of transcripts presented was larger in the BLASTX results (14398) than in the BLASTP results (11072), the proportion of unique gene annotation was higher with BLASTP (64.4%) than BLASTX (55.9%), as was the proportion of DBA-RBA common gene annotation (88.0% in BLASTP versus 82.9% in BLASTX). The majority (10888 of 11072) of de novo assembled contigs annotated in BLASTP were also annotated in BLASTX. Additionally, BLASTP annotation appeared to represent a relatively conservative annotation result in terms of its skewness towards lower E-values (Fig 1C). For this reason, we chose the BLASTP annotation for the downstream analyses. In all, 11072 transcripts (7135 UniProtKB identified genes) were annotated based on BLASTP from DBA using the DAH3615 watermelon lines. Investigation of the annotation sources of those transcripts (Fig 1B) revealed that a majority of transcripts (74.6%) were most closely related to genes from plants, especially from Arabidopsis thaliana (59.2%).

Functional network analysis of novel transcripts
To investigate the functional features of newly annotated transcripts, we performed enrichment analysis of functional terms based on three gene ontology (GO) sub-categories: ‘biological process’ (BP), ‘molecular function’ (MF), and ‘cellular component’ (CC). Eighteen, 28, and 8 functional terms were significantly enriched in BP, MF, and CC, respectively (enrichment test P-value ≤ 0.001, Tables C-E in S1 File). As similar biological terms were repeatedly detected across the three GO categories, we conducted network analysis of functional terms to classify analogous terms. This revealed five large clusters grouped as transmembrane transporter, homeostasis, stimulus, nucleic acid binding, and Golgi and vesicles (Fig 2). Three of the five clusters, nucleic acid binding-related, transmembrane transporter-related and homeostasis-related terms, were significantly related (in a correlation test based on a gene association matrix, FDR adjusted P-value ≤ 0.01), whereas two clusters, stimulus-related and Golgi and vesicle-related clusters, were independently observed. In three highly correlated clusters, diverse terms related to biological functions that controls internal homeostasis against external stimulus through transmembrane ion transport signals were significantly detected together. As these significantly enriched functional terms are fundamental in plants and other organisms, these newly annotated transcripts suggest its importance on plant viability, especially on stimulus and regulation of homeostasis.

Identification of differentially-expressed genes (DEGs) by DBA
After removing non-expressed transcripts, we statistically analyzed 10829 BLASTP-annotated transcripts, to not only identify DEGs associated with male sterility, but also to detect any de novo-exclusive DEGs (DEDEGs) that might be identified as DEGs in only DBA. Two-way analysis of deviance (ANODEV) was conducted for each transcript, taking into account the existence of both sterility and tissue-type variation. After removing duplicated UniProtKB IDs, 443 DEGs (508 transcripts with duplicated UniprotKB ID) were detected between the male-fertile DAH3615 (MF) and the male-sterile DAH3615-MS (MS) lines (FDR adjusted P-value ≤ 0.01).

Comparing the lists of DEGs identified via DBA and RBA, 138 nonduplicated DEGs (representing 140 transcripts) were identified as DEDEGs (Fig 3A). The gene expression pattern of these transcripts was visualized as a heatmap (Fig 3B), which revealed drastic differential
De novo transcriptome analysis of watermelon

**Node annotation**
- GO Biological process
- GO Molecular function
- GO Cellular component

**Stimulus-related terms**
- 2 response to abiotic stimulus
- 14 detection of stimulus
- 15 positive regulation of circadian rhythm
- 16 cellular response to abiotic stimulus

**Nucleic acid binding-related terms**
- 24 ATP binding
- 25 purine ribonucleoside diphosphate binding
- 26 nucleotide binding
- 27 nucleoside phosphate binding
- 28 adeny1 ribonucleotide binding
- 29 adeny1 nucleotide binding
- 32 carbohydrate derivative binding
- 33 nucleoside binding
- 34 purine ribonucleotide binding
- 35 purine nucleotide binding
- 36 ribonucleotide binding
- 37 small molecule binding
- 38 purine nucleoside binding
- 39 purine ribonucleoside binding
- 40 ribonucleoside binding
- 41 protein serine/threonine kinase activity
- 42 ion binding

**Transmembrane transporter-related terms**
- 19 ATPase activity, coupled to transmembrane movement of substances
- 20 ATPase activity, coupled to movement of substances
- 21 hydrolase activity, acting on acid anhydrides, catalyzing transmembrane movement of substances
- 22 P-P bond-hydrolysis-driven transmembrane transporter activity
- 23 primary active transmembrane transporter activity
- 30 active transmembrane transporter activity
- 31 ATPase activity, coupled pyrophosphatase activity
- 44 pyrophosphatase activity
- 45 efflux transmembrane transporter activity
- 46 hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides

**Homeostasis-related terms**
- 1 transmembrane transport
- 3 regulation of pH
- 4 regulation of intracellular pH
- 7 monovalent inorganic cation homeostasis
- 8 ion transmembrane transport
- 9 ion transport
- 10 homeostatic process
- 11 ion homeostasis
- 12 cellular monovalent inorganic cation homeostasis
- 13 cation homeostasis
- 17 inorganic ion homeostasis
- 43 inorganic cation transmembrane transporter activity

**Golgi and vesicle-related terms**
- 47 Golgi apparatus
- 48 Golgi-associated vesicle
- 49 cytoplasmic, membrane-bounded vesicle
- 50 membrane-bounded vesicle
- 51 Golgi membrane
- 52 Golgi apparatus part
- 53 coated vesicle
- 54 cytoplasmic vesicle part

**Others**
- 4 response to acid chemical
- 16 response to oxygen-containing compound
expression of DEDEGs between MS and MF groups. Of these genes, 20 genes were up-regulated in MF, and showed no expression in MS (Fig 3C).

Functional enrichment analysis to identify the functional characteristics of these genes revealed significantly enriched biological terms (enrichment test \( P \)-value \( \leq 0.01 \)) across the three GO categories (Fig 3D and Table F in S1 File). Network analysis revealed four functional clusters: terms related to homeostasis/transmembrane ion transport, wax, nucleic acid binding, and galactosidase. Homeostasis, transmembrane ion transporter activity and nucleic acid binding-related functional clusters were particularly common among newly annotated transcripts (Fig 2) and DEDEGs (Fig 3D).

Technical validation of novel transcripts and DEDEGs

Since no biological replications were used in our RNA-seq experiment, replicates were needed to validate the technique. Three biological replicates were subjected to real-time quantitative reverse transcription-polymerase chain reaction (RT-qPCR), with three technical replicates for each of four samples denoted as MF:B, MF:F, MS:B, and MS:F (B, floral bud; F, flower), for 33 randomly selected transcripts (14 newly annotated transcripts and 19 DEDEGs).

First, RT-qPCR was performed on 14 randomly selected candidates for newly annotated transcripts to investigate reliability. We presumed that the existence of the target transcripts could be demonstrated based on their relative gene expression measured against the control gene, regardless of condition. In this way, gene expression for all 14 newly annotated transcript candidates was successfully detected (S2 Fig), thus demonstrating the reliability of the discovery of novel transcripts, and supporting the use of DBA as complementary to RBA in watermelon. Next, to verify DEDEGs, 19 candidates of 138 DEDEGs (MF vs. MS, FDR adjusted \( P \)-value \( \leq 0.01 \)) were randomly selected for RT-qPCR validation. Two-way ANOVA was used to determine the statistical significance of differential expression of 19 DEDEG candidates. The relative gene expression of each transcript was also compared based on RT-qPCR, although one gene (RLF9) failed to reach the threshold. The other 18 transcripts were all significantly detected as DEGs by comparing MF and MS (Fig 4) (Bonferroni’s adjusted \( P \)-value \( \leq 0.01 \)).

Log 2-fold change (log2FC) values indicated that all of these transcripts were highly down-regulated in MS samples, providing evidence for their fertility-biased expression and association with male sterility.

Discussion

RNA sequencing (RNA-seq) is more cost- and time-effective than expressed sequence tag (EST), qPCR, or microarray analysis and can be used to directly construct de novo transcriptomes of non-model organisms [40]. Transcriptome/genome analysis through RNA-seq can be effectively accomplished through RBA when a reference genome exists, but RBA is completely dependent on the degree of completion of the reference genome. DBA, an alternative approach using de novo transcriptome assembly, can produce distinct results irrespective of presence or quality of a reference genome, but it requires a high degree of computation capacity. Further, RBA and DBA can be used independently or as a combined method.
De novo transcriptome analysis of watermelon

![Diagram](image)

- **Node annotation**
  - GO Biological process
  - GO Molecular function
  - GO Cellular component

- **Homeostasis/Transmembrane ion transport-related terms**
  1. regulation of cellular pH
  2. regulation of intracellular pH
  3. cellular monovalent inorganic cation homeostasis
  4. regulation of pH
  5. monovalent inorganic cation homeostasis
  6. ion transport
  7. transmembrane transport
  8. proton transport
  9. hydrogen transport
  10. ATPase coupled ion transmembrane transporter activity
  11. ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism
  12. active ion transmembrane transporter activity
  13. hydrogen ion transmembrane transporter activity
  14. P-P-bond-hydrolysis-driven transmembrane transporter activity
  15. primary active transmembrane transporter activity
  16. cation-transporting ATPase activity

- **Wax-related terms**
  1. wax biosynthetic process
  2. wax metabolic process

- **Nucleic acid binding-related terms**
  1. purine:ribose-5-phosphate binding
  2. ATP binding
  3. purine:ribose-5-phosphate binding
  4. purine nucleoside binding

- **Galactosidase-related terms**
  1. beta-galactosidase activity
  2. galactosidase activity
  3. vacuole

- **Others**
  1. meristem initiation
  2. phospholipid transporter activity
  3. integral component of organelle membrane
  4. intrinsic component of organelle membrane

---

[Image 1]

[Image 2]

[Image 3]
Reference genomes of major model organisms have been updated continuously based on follow-up research [41]. Since its release, the reference genome of watermelon has been served a crucial role in watermelon genome analyses; however, the current watermelon reference genome is only a draft version published at 2013 and has not been updated yet. For these reasons, we applied RBA and DBA simultaneously to uncover the watermelon transcriptome and to contribute to watermelon genome analysis.

In the watermelon RNA-seq studies using RBA, different numbers of annotated genes were observed, indicating differences in transcriptome profiling, likely attributable to various factors such as experimental conditions or tissue/breed specificity [11, 12, 14]. Thus, gene annotations in diverse tissues and breeding lineages are helpful to explore such specificities via RNA-seq analysis. Based on our previous RBA study, we speculated that complementary annotation is needed to elucidate the missing part of the previous RBA, considering the relative infancy of the watermelon reference genome. The low mapping rates to the watermelon reference genome (51.0–54.7%) compared to that of *Arabidopsis thaliana* also indicate the insufficiency of RBA using the watermelon reference with DAH3615 lines, implying that application of DBA could be helpful to complement the reference-based watermelon transcriptome in terms of providing genomic information [42, 43].

Here, we compared DBA and RBA approaches on the same RNA-seq data to identify whether DBA would improve upon RBA-based annotation and provide distinct results. With regard to the low mapping rates on the reference genome, we speculated that the individual application and comparison of DBA and RBA would minimize the loss of information, and enable more straightforward observation of the watermelon transcriptome than another combined strategy, such as align-then-assemble. To minimize false positives and conservatively compare DBA with RBA, we collected only plant-derived transcripts for BLASTP annotation of the *de novo* assembled transcriptome, and discovered 855 new transcripts that represent parts of the transcriptome thus far undiscovered by RBA (Fig 1A). Since these novel findings may provide valuable information that RBA could not detect, it was necessary to validate their reliability—given that all background information is generated *ab initio*, DBA is particularly prone to the problem of false positives.

Four notable pieces of evidence support the reliability of DBA in this study. First, while 855 genes (1132 transcripts) were newly detected, 6280 (88.0%) genes from DBA were commonly identified in RBA (Fig 1A). Second, the large proportion (74.6%) of annotated transcripts were most closely related to those from plants, including *Arabidopsis thaliana* (59.2%) (Fig 1B), despite the fact that the functional annotation in the Swiss-Prot database is generally skewed towards representative mammalian organisms. Third, E-value distribution in the annotation step revealed that many annotated transcripts matched uniquely between the database and query sequences; this finding indicates that the annotated contigs were well assembled with clarity (Fig 1C). Finally, when 14 of the 855 newly identified transcripts were selected for RT-qPCR with biological replicates, gene expression could be confirmed for all transcripts (S2 Fig). This result supports the idea that the newly identified transcripts derived from DBA are
bona fide. Taken together, these results show that DBA provides distinct benefits for watermelon transcriptome research; the identification of 855 novel transcripts is valuable for complementing the available RBA genomic information of watermelon.

For an overview of the possible functional properties of the 855 newly annotated watermelon transcripts, we conducted functional term network analysis. We detected five large clusters related to transmembrane transporter, homeostasis, stimulus, nucleic acid binding, and Golgi and vesicles (Fig 2 and Tables C-E in S1 File). Genes assigned to these terms serve basic but crucial roles in cellular activities of plants. Three of the five clusters, nucleic acid binding-related, transmembrane transporter-related and homeostasis-related terms, were especially highly correlated. These serve a role in one of the most important cellular processes, cell survival through regulation of homeostasis. The fact that the functional terms of these newly annotated transcripts are relevant to basic processes occurring in the plant provides evidence of the necessity for DBA in watermelon. Thus, we conclude that the use of DBA contributes useful insights and enables diverse interpretations in further watermelon research.

Of our 855 newly annotated transcripts, DBA uniquely revealed 138 transcripts that were differentially expressed between DAH3615 and DAH3615-MS; we termed these DEDEGs. A previous transcriptome study using RBA suggested the existence of biased expression between MF and MS groups [13], and indeed a similarly biased pattern was observed among our DEDEGs (Fig 3B and 3C). Considering the phenotypic differences between male sterility and fertility and the fertility biased-expression patterns observed in both RBA and DBA analyses, this observation suggests that the DEDEGs are candidates for serving roles in male sterility along with DEGs previously discovered by RBA.

The 138 DEDEGs we observed formed four functional clusters including homeostasis/ ion transporter, wax, nucleic binding, and galactosidase-related terms (Fig 3D and Table F in S1 File). These clusters are frequently observed to function in plant sterility. Ion transporters are involved in signal transduction, cell wall metabolism, and rearrangement of cytoskeletons [44]. Such transporters are enriched in pollen and involved in pollen maturation and pollen tube elongation. Notably, Ms-cd1 mutant cabbage producing collapsed pollen showed repression of various ion transporters in floral buds [45]. Galactosidase is a cell wall modifying enzyme that is involved in microspore development [46]. The anther surface and pollen exine are composed of cutin and intra- and epi-cuticular waxes [47]. Similar to our results, wax-related genes and galactosidase-related genes have been reported in the transcriptome comparison of male-sterile and fertile lines [45–48]. Consistent with these results, we anticipate that the DAH3615-MS, which lacks pollen and exhibits small-sized stamen, is deficient in these structural proteins.

Our RNA-seq experiment was based on data from single biological samples; thus, technical validation was required to authenticate those findings. The importance of biological replicates in conducting accurate experiments that take biological variation into account cannot be ignored; however, RNA-seq is frequently been used to pre-screen and narrow the focus of transcriptomic studies, which may then be followed by RT-qPCR. We used RNA-seq analysis to select the most probable candidates for technical validation, then performed RT-qPCR on three biological replicates, and three technical replicates for 33 randomly selected transcripts (14 newly annotated transcripts and 19 DEDEGs). Thirty-two of 33 transcripts were successfully
validated (Fig 4 and S2 Fig), and functional annotations of these transcripts were produced. The rest of the newly annotated transcripts discovered by DBA are also strong candidates to be breeding line-specific genes, and further experiments with replication are needed to verify their differential expression.

Among the 18 DEDEGs successfully validated by RT-qPCR, we identified EXPA9, which encodes expansin, a cell wall-loosening enzyme located in pollen grains that participates in pollen germination to loosen the cell wall of the stigma and the style, thus helping lignin pollen tube penetration [49, 50]. Additionally, among our validated transcripts were two TBB1 and one TBB2 tubulin genes (Fig 4). Alpha-tubulin and beta-tubulin are major components of microtubules, which play roles in pollen development and pollen tube germination [51]. Pyruvate decarboxylase (PDC) transforms pyruvate into acetaldehyde and carbon dioxide [49]. It is abundant in pollen grains and is related to pollen tube germination and growth. PDC2 is the only functional PDC gene in pollen; the pdc2 knockout mutant had significantly reduced pollen tube growth compared to the wild type. PDC2 has been suggested as a strong candidate for a role in male sterility in petunia [52].

Our study also revealed that expression of transcripts related to flowering time and organogenesis was biased towards the MF line. A transcript for the phosphatidylinositol/phosphatidylcholine transfer protein SFH3, was highly enriched in the MF line (Fig 4; logFC: −8.3); this protein is reportedly associated with early bolting and early flower formation, giving rise to variation in flower and petal size in Brassica napus [53].

Orthologs of some of the genes we identified have been reported to be responsible for inducing male sterility. A transgenic fasciclin-like arabinogalactan protein (FLA3)-overexpressing line had reduced stamen filament elongation that was both directly and indirectly associated with male sterility [54]. Based on this report, we anticipate that FLA5 also has the potential to be involved in male sterility.

Polyubiquitination (catalyzed via UBQ14) regulates various physiological functions such as sexual reproduction. Ubiquitin (Ub) and Ub-conjugated proteins are involved in early anther development in Nicotiana alata [55]. The E3 ligase-like protein and the F-box protein are related to male sterility in hybrid rice [56]. Another DEDEG showed similarity to BAG1, which encodes a protein with an ubiquitin-like domain and a BAG domain that, like heat shock-induced gene 1, a putative grape BAG protein, promotes the meristematic transition from vegetative to reproductive growth and early flowering [57, 58]. ATPase (encoded by PMA2) plays a crucial role in energy release by dephosphorylating ATP to ADP. SPLAYED (SYD), a novel SWI/SNF ATPase homolog, interacts with LEAFY, which is a well-known regulator of floral transition in Arabidopsis. As shown by a study of a syd-2 line, which exhibits male fertility and a reduction in anther dehiscence, SYD is necessary for reproductive and meristem development [59].

Another notable DEDEG was a transcript encoding a protein Stellacyanin (STEL), a blue copper protein, which was predominantly expressed in male-fertile watermelon lines. Although, there are no previous reports of a relationship between STEL and male sterility or reproductive organ development, our previous RBA results have deduced another blue copper protein to be the most significantly differentially expressed gene in watermelon male-fertile lines compared to male-sterile lines [13]. We therefore conclude that STEL could be a novel gene involved in male sterility in watermelon.

The potential links to reproductive development of our candidate genes described above serves to further validate the reliability of DBA, especially in identifying genes that might be helpful for future studies of male sterility in watermelon. Although we have technically validated only 18 of the candidate DEDEGs, the others are also likely to be strong candidates related to male sterility of watermelon—further studies should seek to validate these genes.
To sum up, we carried out DBA to complement RBA on watermelon RNA-seq data. This simultaneous application and comparison of both approaches improved upon RBA alone, as shown by the following results. A total of 855 transcripts were newly discovered using DBA, and 138 DEDEGs were identified as DBA-derived candidate male-sterility genes. The DEDEGs and their technical validation corresponded with RBA results in terms of male-fertility biased expression and genes with analogous functions. Through the functional annotation of our newly annotated transcripts, essential gene functions related to transmembrane transport, homeostasis, stimulus, nucleic acid binding, and Golgi and vesicles were established for watermelon species. Furthermore, our set of 138 putative male sterility-related genes should prove valuable for further watermelon studies. Overall, we conclude that DBA provided a distinct result that could not be discovered using RBA with the current watermelon reference genome. Within the limits of the reference genome of Citrullus lanatus, individual application of DBA and RBA can be a valuable tool to complete the transcriptome. The reliable results obtained in this watermelon genome study can be useful for further watermelon transcriptome studies, showing the value of DBA in non-model plant organisms and providing clues to male sterility in watermelon. This integration of DBA and RBA thus contributes to genome study of watermelon as well as to plant male-sterility research.

Supporting information

S1 Fig. Venn diagram showing the comparison between DBA and RBA BLASTX annotations. (PDF)

S2 Fig. RT-qPCR results for 14 newly annotated transcripts. (PDF)

S1 File. Supplementary tables A–J. (DOCX)

Acknowledgments

This study was supported by the Golden Seed Project (213006051SBV20); the Ministry of Agriculture, Food, and Rural Affairs (MAFRA); the Ministry of Oceans and Fisheries (MOF); the Rural Development Administration (RDA); and the Korean Forest Service (KFS) of the Republic of Korea.

Author Contributions

Conceptualization: Sang-Wook Han, Gung Pyo Lee.
Data curation: Sun-Ju Rhee, Taehyung Kwon, Minseok Seo.
Formal analysis: Minseok Seo, Seoae Cho.
Funding acquisition: Gung Pyo Lee.
Investigation: Sun-Ju Rhee, Taehyung Kwon, Yoon Jeong Jang, Tae Yong Sim.
Methodology: Taehyung Kwon, Minseok Seo, Yoon Jeong Jang, Tae Yong Sim, Gung Pyo Lee.
Project administration: Gung Pyo Lee.
Resources: Seoae Cho.
Software: Minseok Seo.

Supervision: Sang-Wook Han, Gung Pyo Lee.

Validation: Sun-Ju Rhee, Yoon Jeong Jang, Gung Pyo Lee.

Visualization: Sun-Ju Rhee, Taehyung Kwon, Minseok Seo.

Writing – original draft: Sun-Ju Rhee, Taehyung Kwon, Sang-Wook Han, Gung Pyo Lee.

Writing – review & editing: Sun-Ju Rhee, Taehyung Kwon, Minseok Seo, Yoon Jeong Jang, Sang-Wook Han, Gung Pyo Lee.

References

1. Guo S, Zhang J, Sun H, Salse J, Lucas WJ, Zhang H, et al. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. Nat Genet. 2013; 45(1):51–8. https://doi.org/10.1038/ng.2470 PMID: 23179023

2. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010; 11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106 PMID: 20979621

3. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010; 28(5):511–5. https://doi.org/10.1038/nbt.1621 PMID: 20436464

4. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics. 2010; 26(4):493–500. https://doi.org/10.1093/bioinformatics/btp692 PMID: 20022975

5. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29(7):644–52. https://doi.org/10.1038/nbt.1883 PMID: 21572440

6. Burset M, Seledtsov I, Solovyev V. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000; 28(21):4364–75. PMID: 11058137

7. Martin JA, Wang Z. Next-generation transcriptome assembly. Nat Rev Genet. 2011; 12(10):671–82. https://doi.org/10.1038/nr8368 PMID: 21897427

8. Chen G, Li R, Shi L, Qi J, Hu P, Luo J, et al. Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. BMC Genomics. 2011; 12(1):590.

9. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics. 2011; 27(8):1068–75. https://doi.org/10.1093/bioinformatics/btr085 PMID: 21330289

10. Phillippy AM, Schatz MC, Pop M. Genome assembly forensics: finding the elusive mis-assembly. Genome Biol. 2008; 9(3):R55. https://doi.org/10.1186/gb-2008-9-3-r55 PMID: 18541692

11. Fan M, Huang Y, Zhong Y, Kong G, Xie J, Niu M, et al. Comparative transcriptome profiling of potassium starvation responsiveness in two contrasting watermelon genotypes. Planta. 2014; 239(2):397–410. https://doi.org/10.1007/s00425-013-1976-z PMID: 24185372

12. Guo S, Sun H, Zhang H, Liu J, Ren Y, Gong G, et al. Comparative Transcriptome Analysis of Cultivated and Wild Watermelon during Fruit Development. PLoS One. 2015; 10(6):e0130267. https://doi.org/10.1371/journal.pone.0130267 PMID: 26079257

13. Rhee S-J, Seo M, Jang Y-J, Cho S, Lee GP. Transcriptome profiling of differentially expressed genes in floral buds and flowers of male sterile and fertile lines in watermelon. BMC Genomics. 2015; 16(1):1.

14. Grassi S, Piro G, Lee JM, Zheng Y, Fei Z, Dalessandro G, et al. Comparative genomics reveals candidate carotenoid pathway regulators of ripening watermelon fruit. BMC Genomics. 2013; 14(1):1–20. https://doi.org/10.1186/1471-2164-14-761 PMID: 24219562

15. Celik Altunoglu Y, Baloglu MC, Baloglu P, Yer EN, Kara S. Genome-wide identification and comparative expression analysis of LEA genes in watermelon and melon genomes. Physiol Mol Biol Plants. 2017; 23(1):5–21. https://doi.org/10.1007/s12298-016-0405-8 PMID: 28250580

16. Yang Y, Mo Y, Yang X, Zhang H, Wang Y, Li H, et al. Transcriptome Profiling of Watermelon Root in Response to Short-Term Osmotic Stress. PLoS One. 2016; 11(11):e0166314. https://doi.org/10.1371/journal.pone.0166314 PMID: 27861528

17. Zhu Q, Gao P, Liu S, Zhu Z, Amanullah S, Davis AR, et al. Comparative transcriptome analysis of two contrasting watermelon genotypes during fruit development and ripening. BMC Genomics. 2017; 18(1):3. https://doi.org/10.1186/s12864-016-3442-3 PMID: 28049426
18. Vedel F, Pla M, Vitart V, Gutierrez S, Chetrit P, Depaepe R. Molecular Basis of Nuclear and Cytoplasmic Male-Sterility in Higher-Plants. Plant Physiol Bioch. 1994; 32(5):601–18.

19. Ray DT, Sherman JD. Desynaptic Chromosome Behavior of the Gms Mutant in Watermelon. J Hered. 1988; 79(5):397–9.

20. Watts V, editor A marked male-sterile mutant in watermelon. Proc Amer Soc Hort Sci; 1962.

21. Watts V, editor Development of disease resistance and seed production in watermelon stocks carrying msg gene. Proceedings of the American Society for Horticultural Science; 1967: AMER SOC HORTICULTURAL SCIENCE 701 NORTH ASAPTH STREET, ALEXANDRIA, VA 22314–1998.

22. Huang HX, Zhang XQ, Wei ZC, Li OH, Li X. Inheritance of male-sterility and dwarfism in watermelon [Citrullus lanatus (Thunb.) Matsum. and Nakai]. Sci Hortic-Amsterdam. 1998; 74(3):175–81.

23. Zhang X, Wang M. A genetic male-sterile (ms) watermelon from China. Report-Cucurbit Genetics Cooperative. 1995;(13):45–6.

24. Dyutin K, Sokolov S. Spontaneous mutant of watermelon with male sterility. Tsitologiya i Genetika. 1990; 24(2):56–7.

25. Bang H, King SR, Liu W. A new male sterile mutant identified in watermelon with multiple unique morphological features. REPORT-CUCURBIT GENETICS COOPERATIVE. 2005; 28:47.

26. Watts V, editor A marked male-sterile mutant in watermelon. Proceedings of the American Society for Horticultural Science; 1962.

27. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014:btu170.

28. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Procot. 2013; 8(8):1494–512. https://doi.org/10.1038/nprot.2013.084 PMID: 23845962

29. Langmead B, Salezberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4):357–9. https://doi.org/10.1038/nmeth.1923 PMID: 22388286

30. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12(1):1.

31. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36 PMID: 23618408

32. Falchi M, Moustafa JSE-S, Takousis P, Pesce F, Bonnefond A, Andersson-Assarsson JC, et al. Low copy number of the salivary amylase gene predisposes to obesity. Nat Genet. 2014; 46(5):492–7. https://doi.org/10.1038/ng.2939 PMID: 24686848

33. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1):139–40. https://doi.org/10.1093/bioinformatics/btp816 PMID: 19910308

34. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2008; 4(1):44–57.

35. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009; 37(1):1–13. https://doi.org/10.1093/nar/gkn233 PMID: 19033633

36. Episkamp S, Cramer AO, Waldorp LJ, Schmittmann VD, Borsboom D. ograph: Network visualizations of relationships in psychometric data. J Stat Softw. 2012; 48(4):1–18.

37. Unterberger A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012; 40(15):e115–e. https://doi.org/10.1093/nar/gks598 PMID: 22730293

38. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25(17):3389–402. PMID: 9254694

39. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011; 2011:bar009. https://doi.org/10.1093/database/bar009 PMID: 21447597

40. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2008; 10(1):57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660

41. Tyner C, Barber GP, Casper J, Clawson H, Diekhans M, Eisenhart C, et al. The UCSC Genome Browser database: 2017 update. Nucleic acids research. 2016; 45(D1):D626–D34. https://doi.org/10.1093/nar/gkw1134 PMID: 27899642
42. Loraine AE, McCormick S, Estrada A, Patel K, Qin P. RNA-seq of *Arabidopsis* pollen uncovers novel transcription and alternative splicing. Plant Physiol. 2013; 162(2):109–2–109. https://doi.org/10.1104/pp.112.211441 PMID: 23590974

43. Vidal EA, Moyano TC, Krouk G, Katari MS, Tanurdzic M, McCombie WR, et al. Integrated RNA-seq and sRNA-seq analysis identifies novel nitrate-responsive genes in *Arabidopsis* thaliana roots. BMC Genomics. 2013; 14(1):701.

44. Bock KW, Honys D, Ward JM, Padmanaban S, Nawrocki EP, Hirschi KD, et al. Integrating membrane transport with male gametophyte development and function through transcriptomics. Plant Physiol. 2006; 140(4):1151–68. https://doi.org/10.1104/pp.105.074708 PMID: 16607029; PubMed Central PMCID: PMCPMC1435806.

45. Kang J, Zhang G, Bonnema G, Fang Z, Wang X. Global analysis of gene expression in flower buds of Ms-cd1 *Brassica oleracea* conferring male sterility by using an *Arabidopsis* microarray. Plant Mol Biol. 2008; 66(1–2):177–92. https://doi.org/10.1007/s11103-007-9261-9 PMID: 18040866.

46. Du K, Liu Q, Wu X, Jiang J, Wu J, Fang Y, et al. Morphological Structure and Transcriptome Comparison of the Cytoplasmic Male Sterility Line in *Brassica napus* (SaNa-1A) Derived from Somatic Hybridization and Its Maintainer Line SaNa-1B. Front Plant Sci. 2016; 7:1313. https://doi.org/10.3389/fpls.2016.01313 PMID: 27656189; PubMed Central PMCID: PMCPMC5011408.

47. Chang Z, Chen Z, Wang N, Xie G, Lu J, Yan W, et al. Construction of a male sterility system for hybrid rice breeding and seed production using a nuclear male sterility gene. Proc Natl Acad Sci U S A. 2006; 113(49):14145–50. https://doi.org/10.1073/pnas.1613792113 PMID: 27864513; PubMed Central PMCID: PMCPMC510371.

48. Omidvar V, Mohoririan I, Dalmay T, Zheng Y, Fei Z, Pucci A, et al. Transcriptional regulation of male-sterility in 7B-1 male-sterile tomato mutant. PLoS One. 2017; 12(2):e0170715. https://doi.org/10.1371/journal.pone.0170715 PMID: 28178307; PubMed Central PMCID: PMCPMC5298235.

49. Cosgrove DJ, Bedinger P, Durachko DM. Group I allergens of grass pollen as cell wall-loosening agents. P Natl Acad Sci USA. 1997; 94(12):6559–64. https://doi.org/10.1073/pnas.94.12.6559

50. Wang W, Scali M, Vignani R, Milanesi C, Petersen A, Sari-Go rla M, et al. Male-sterile mutation alters *Zea m 1* (beta-expansin 1) accumulation in a maize mutant. Sex Plant Reprod. 2004; 17(1):41–7.

51. Dai S, Li L, Chen T, Chong K, Xue Y, Wang T. Proteomic analyses of *Oryza sativa* mature pollen reveal novel proteins associated with pollen germination and tube growth. Proteomics. 2006; 6(8):2504–29. https://doi.org/10.1002/pmic.200401351 PMID: 16548068.

52. Choi D, Cho HT, Lee Y. Expansins: expanding importance in plant growth and development. Physiol Plant. 2006; 126(4):511–8. https://doi.org/10.1111/j.1399-3054.2005.00612.x

53. Dong J, Kim ST, Lord EM. Plantacyanin plays a role in reproduction in *Arabidopsis*. Plant Physiol. 2005; 138(2):778–89. https://doi.org/10.1104/pp.105.063388 PMID: 15908590; PubMed Central PMCID: PMCPMC1150396.

54. Coimbra S, Almeida J, Junqueira V, Costa ML, Pereira LG. Arabinogalactan proteins as molecular markers in *Arabidopsis thaliana* sexual reproduction. J Exp Bot. 2007; 58(15–16):4027–35. https://doi.org/10.1093/jxb/erm259 PMID: 18039740

55. Li YQ, Southworth D, Linskens HF, Mulcahy DL, Cresti M. Localization of Ubiquitin in Anthers and Pistils of Nicotiana. Sex Plant Reprod. 1995; 8(3):123–8.

56. Long Y, Zhao L, Niu B, Su J, Wu H, Chen Y, et al. Hybrid male sterility in rice controlled by interaction between divergent alleles of two adjacent genes. Proc Natl Acad Sci USA. 2008; 105(48):18871–6. https://doi.org/10.1073/pnas.0810108105 PMID: 19033192; PubMed Central PMCID: PMC2596266.

57. Kobayashi M, Takato H, Fujita K, Suzuki S. HSG1, a grape Bcl-2-associated athanogene, promotes floral transition by activating CONSTANS expression in transgenic *Arabidopsis* plant. Molecular biology reports. 2012; 39(4):4367–74. https://doi.org/10.1007/s11033-011-1224-1 PMID: 21901420

58. Doukanina EV, Chen S, van der Zalm E, Godzik A, Reed J, Dickman MB. Identification and functional characterization of the BAG protein family in *Arabidopsis thaliana*. Journal of Biological Chemistry. 2006; 281(27):18793–801. https://doi.org/10.1074/jbc.M511794200 PMID: 16636050

59. Wagner D, Meyerowitz EM. SPAYERD, a novel SWI/SNF ATPase homolog, controls reproductive development in *Arabidopsis*. Curr Biol. 2002; 12(2):85–94. https://doi.org/10.1016/S0960-9822(01)00651-0 PMID: 11818058