

Article

# Comparative Analysis of Developmental Transcriptome Maps of *Arabidopsis thaliana* and *Solanum lycopersicum*

Aleksey A. Penin <sup>1,2,\*</sup>, Anna V. Klepikova <sup>1,3,†</sup>, Artem S. Kasianov <sup>3,4,†</sup>, Evgeny S. Gerasimov <sup>1,2</sup> and Maria D. Logacheva <sup>1,2,3,†</sup>

<sup>1</sup> Institute for Information Transmission Problems of the Russian Academy of Sciences, Bolshoy Karetny per. 19, build. 1, Moscow 127051, Russia; a.klepikova@skoltech.ru (A.V.K.); gerasimov\_e@mail.bio.msu.ru (E.S.G.); m.logacheva@skoltech.ru (M.D.L.)

<sup>2</sup> Lomonosov Moscow State University, Leninskye Gory, Moscow 119992, Russia

<sup>3</sup> Skolkovo Institute of Science and Technology, Center for Data-Intensive Biology and Biomedicine, Nobelya Ulitsa 3, Moscow 121205, Russia

<sup>4</sup> Vavilov Institute of General Genetics, Russian Academy of Sciences, Gubkina 3, Moscow 119991, Russia; artem.kasianov@vigg.ru

\* Correspondence: alekseypenin@gmail.com; Tel.: +7-926-378-93-87

† These authors contributed equally to this work.

Received: 3 December 2018; Accepted: 4 January 2019; Published: 15 January 2019



**Abstract:** The knowledge of gene functions in model organisms is the starting point for the analysis of gene function in non-model species, including economically important ones. Usually, the assignment of gene functions is based on sequence similarity. In plants, due to a highly intricate gene landscape, this approach has some limitations. It is often impossible to directly match gene sets from one plant species to another species based only on their sequences. Thus, it is necessary to use additional information to identify functionally similar genes. Expression patterns have great potential to serve as a source of such information. An important prerequisite for the comparative analysis of transcriptomes is the existence of high-resolution expression maps consisting of comparable samples. Here, we present a transcriptome atlas of tomato (*Solanum lycopersicum*) consisting of 30 samples of different organs and developmental stages. The samples were selected in a way that allowed for side-by-side comparison with the *Arabidopsis thaliana* transcriptome map. Newly obtained data are integrated in the TraVA database and are available online, together with tools for their analysis. In this paper, we demonstrate the potential of comparing transcriptome maps for inferring shifts in the expression of paralogous genes.

**Keywords:** tomato; *Solanum lycopersicum*; RNA-seq; transcriptome atlas

## 1. Introduction

In plant science, the overwhelming majority of experiments aimed at the identification of gene functions are carried out primarily in the model species, *Arabidopsis thaliana*. The knowledge of gene function is transferred from model species to non-model species (e.g., agriculturally important plants) based on the similarity of gene sequences and phylogenetic analysis. The assumption that underlies such transfer is that orthologous genes have similar functions. However, it is well-known that this is not true in many cases [1,2]). Additionally, the accuracy of orthology inference is not high, especially in plant genomes which are shaped by multiple whole-genome and segment duplications and subsequent gene loss. Thus, it is highly desirable to use additional information that would help in developing hypotheses on gene functions. Gene expression data can serve as a source of such information.

The similarity of gene-expression profiles of homologous genes indicates the conservation of functions, while a drastic difference in expression profiles suggests functional divergence. This approach requires high-resolution data on gene-expression profiles that can be compared across species. Previously, we had developed a transcriptome map of the model plant *Arabidopsis thaliana* [3]; here, we report the transcriptome map of *Solanum lycopersicum* (tomato). Tomatoes are a representative of a large clade of eudicots—asterids, while *Arabidopsis* belongs to another large clade—rosids. Tomatoes are an important agricultural plant that are cultivated worldwide for its fruit; the gross production of tomato is >170 million tons [4]. Despite this, there is still great demand for new cultivars [5]. This requires an object-specific knowledge of gene functions.

The (meta) analysis of high-throughput gene expression data has great potential to improve functional annotations (e.g., [6–8]). In December 2018, the National Center for Biotechnology Information (NCBI) database listed 213 BioProjects containing RNA-seq data for tomato; these projects included 3986 sequence read archive (SRA) accessions. The majority of these data can be divided into three categories. The first category represents fruit development: the main part of which is available via the TomExpress database (<http://tomexpress.toulouse.inra.fr/> [9]), and the SolGenomics project (<http://tea.solgenomics.net> [10–12]), which provides an attractive and useful overview of gene expression in the tomato fruit; the second category represents studies on stress response; and the third includes studies comparing gene expression in wild-type plants and in mutants or genetically modified plants. Publicly available NCBI RNA-seq datasets that can be regarded as developmental transcriptome maps are represented by two BioProjects—PRJDB5790 and PRJNA307656 [13]. Unfortunately, both of them have an important shortcoming—they were conducted in one biological replicate. This hampers any statistical analysis of gene expression and splicing. The motivation for these studies was not genome-wide expression analysis. For example, in a study by Cárdenas et al. [13], the transcriptome map was used to study the function (including detailed expression profile and co-expression with other genes) of the gene GAME9. A global analysis of the transcriptome data was not performed. Taking into account this gap in the transcriptomic data of tomato, its taxonomic position, and agricultural importance, we have collected and sequenced tomato samples and created a tomato transcriptome atlas containing data on gene-expression profiles in 30 samples that represent different organs and developmental stages (Supplementary Table S1). The results are integrated in the TraVA (Transcriptome Variation Analysis) database (<http://travadb.org/browse/Species=Tomato/>). Here, we present the results of the analysis of these data, focusing on comparison with the *Arabidopsis* transcriptome map.

## 2. Materials and Methods

### 2.1. Sample Collection

Each sample is a pool collected from at least 10 plants in order to decrease the variance caused by inter-individual differences in gene expression. Samples were collected in two biological replicates. Plants were grown in a climate chamber (POL-EKO Aparatura, Vladislavia, Poland) under a 16 h light/8 h dark cycle at 22 °C and 50–60% relative humidity.

### 2.2. RNA Extraction

RNA was extracted using the RNeasy mini kit (Qiagen, Venlo, The Netherlands) and following the manufacturer's protocol. RNA quality was controlled using capillary electrophoresis on a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA).

### 2.3. Library Preparation and Sequencing

PolyA mRNA was extracted using TruSeq RNA Sample Prep Kits v2 (Illumina, San Diego, CA USA) in 0.4 of the recommended volume, due to the small amounts of RNA in some samples. Illumina cDNA libraries were constructed with the NEBNext Ultra II RNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA) following the manufacturer's protocol in 0.5 of the recommended volume.

Sequencing of the cDNA libraries was performed using the HiSeq4000 (Illumina) instrument (4 lanes, 60 bp single read run). For several libraries, the resulting the total number of reads sequencing on HiSeq4000 was less than 20 million, and they were sequenced once more on NextSeq500 (Illumina) with a 75 bp read length.

#### 2.4. Mapping

Raw reads from the SRA (Supplementary Table S2) were downloaded using the recommended utility fastq-dump (from NCBI SRA-toolkit version 2.8.0) with the “- -split-files” option.

Reads for publicly available datasets from SRA and reads for 30 samples obtained in the frame of this study were quality- and adapter-trimmed using Trimmomatic [14] version 0.36. For single-read data, Trimmomatic was run in single-read mode, and for paired-end—in paired-end mode. Other Trimmomatic options (identical for single-read and paired-end data) were set with the following command line: “ILLUMINACLIP:common.adaptors.file:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:15 MINLEN:30”.

Genome assembly for *S. lycopersicum* was taken from the Ensembl database: assembly and corresponding annotation version SL2.50. Trimmed reads were mapped on the genome assembly using Spliced Transcripts Alignment to a Reference (STAR) [15] version 2.4.2a in the “GeneCounts” mode, and with the provided annotation file to obtain read counts for genes. In this mode, STAR v. 2.4.2 is claimed to count only uniquely mapped reads per gene. Other mapping parameters were adjusted by the following options: “- -outFilterMismatchNmax 3 - -outSJfilterCountUniqueMin 3 1 1 1 - -outSJfilterCountTotalMin 3 1 1 1 - -alignIntronMin 6”.

For the FRUIT dataset, see the results of mapping in Supplementary Figure S1.

#### 2.5. Splicing Analysis

Splicing analysis was based on the splice junctions’ output from STAR. Two filter criteria were applied to the collection of all discovered splice junctions for each dataset in the study. The splice junction passes filter 1 if it exists in two or more sequencing runs (e.g., any two fastq files). The splice junction passes filter 2 if it exists in two replicates of the sample (two fastq files from the same sample). For samples in the public DEVELOPMENT dataset (Supplementary Table S2), there were no replicates provided; thus, filter 2 was not used for this dataset. Filtering and counting were performed using custom Python script. The filtering procedure is crucial to remove possible artefacts, especially on the read-mapping stage of data processing. On the other hand, filtration that is too strict can remove rare splicing events from analysis (for example, in our development dataset 1730 (1,3%), introns annotated in SL2.50 did not pass even filter 1). The splice junctions are listed in Supplementary Table S3.

#### 2.6. Expression Analysis

In order to quantify gene expression levels, we calculated total gene read (TGR) values. To avoid library size bias, TGR values were normalized between samples using size factors, as described by Anders and Huber [16]. A gene was considered as expressed if, in each biological replicate, the gene has a normalized TGR value of 5 or higher (weak threshold) or 16 or higher (strong threshold, [17]). For the completeness of the discovery of expressed genes, three publicly available datasets were used (Supplementary Table S2) with the same thresholds. Differentially expressed genes were detected using the R package “DESeq2” [18] with the following thresholds: a false discovery rate (FDR) <0.05 and a fold change  $\geq 2$ . The differential expression (DE) score was defined as the number of pairwise comparisons in which a given gene was differentially expressed [3].

#### 2.7. Detection of Stably Expressed Genes

For the assessment of expression stability, only genes expressed in all samples under a weak threshold were considered. Using normalized TGR values, the mean and standard deviation of expression were calculated for each gene expressed in all samples. The coefficient of variation (CV)

was calculated as the standard deviation divided by the mean. Genes with CV less than 0.3 were considered stably expressed across all samples.

### 2.8. Gene Ontology Enrichment Analysis

Overrepresented Gene Ontology (GO) categories in gene lists in comparison with all genes of *S. lycopersicum* were found using the PANTHER Classification System Version 13.1 [19,20] statistical overrepresentation test with default settings (including FDR <0.05) and fold enrichment  $\geq 2$ . PANTHER Pathways and the PANTHER protein class were also checked for overrepresentation.

### 2.9. Shannon Entropy

Shannon entropy (H) values were used for expression pattern width assessment, and were calculated for genes expressed in at least one sample under a weak threshold, as done by Schug et al. [21]. To avoid bias due to overrepresentation of certain parts and organs, the samples were grouped using hierarchical clustering: samples with a distance ( $1 - \text{Pearson } r^2$ ) <0.3 were grouped (the sample combination is described in Supplementary Table S4), and the gene expression levels were averaged.

### 2.10. Pseudo-Euclidean Distance

For cross-species TGR normalization, 7460 pairs of orthologous genes were taken. 79 samples in two biological replicates for 7460 *Arabidopsis* genes and 30 samples in two biological replicates for 7460 *Solanum* genes were combined in a single dataset, with 109 samples for 7460 genes. The TGR of this dataset was normalized by size factor as in “DESeq” [16], and the size factor values for each sample were stored and used for the normalization of expression of all genes in the *Arabidopsis* and *Solanum* datasets separately.

Only genes that were expressed in at least one sample were considered for analysis in both *Arabidopsis* and *Solanum*. For all genes, the normalized TGR was incremented by 1 to obtain non-zero values of median expression for tissue-specific genes; for each gene, the normalized TGR in all samples were divided by the median, which provides relative expression profiles.

Although the *Solanum* transcriptome map was collected in agreement with the *Arabidopsis* atlas, several samples were hard to match directly; such samples were grouped. A list of sample combinations is provided in Supplementary Table S5.

For each pair of *Arabidopsis* and *Solanum* genes, the pseudo-Euclidean distance was calculated in 100 repeats as follows:

1. For each sample of *Arabidopsis* and *Solanum*, one of the biological replicates was randomly taken.
2. For each pair of samples, the residuals of median-normalized TGR were calculated. In the case of a group of samples, the residuals were counted for all possible pairs of *Arabidopsis* and *Solanum* samples, and a minimum value of residuals was chosen.
3. All residual values were summed, and a squared root of the sum was calculated to obtain the pseudo-Euclidean distance.

Then, 100 replicates of the pseudo-Euclidean distance were averaged and used as the expression distance measure for the pair of genes.

### 2.11. Orthology Assessment

For the detection of orthogroups, the OrthoFinder version 2.2.6 [22] software with default parameters was used. Proteins from the longest isoforms of the TAIR10 version of the *A. thaliana* annotation and the SL2.50 version of the *S. lycopersicum* annotation were used as OrthoFinder input.

### 2.12. Data Availability

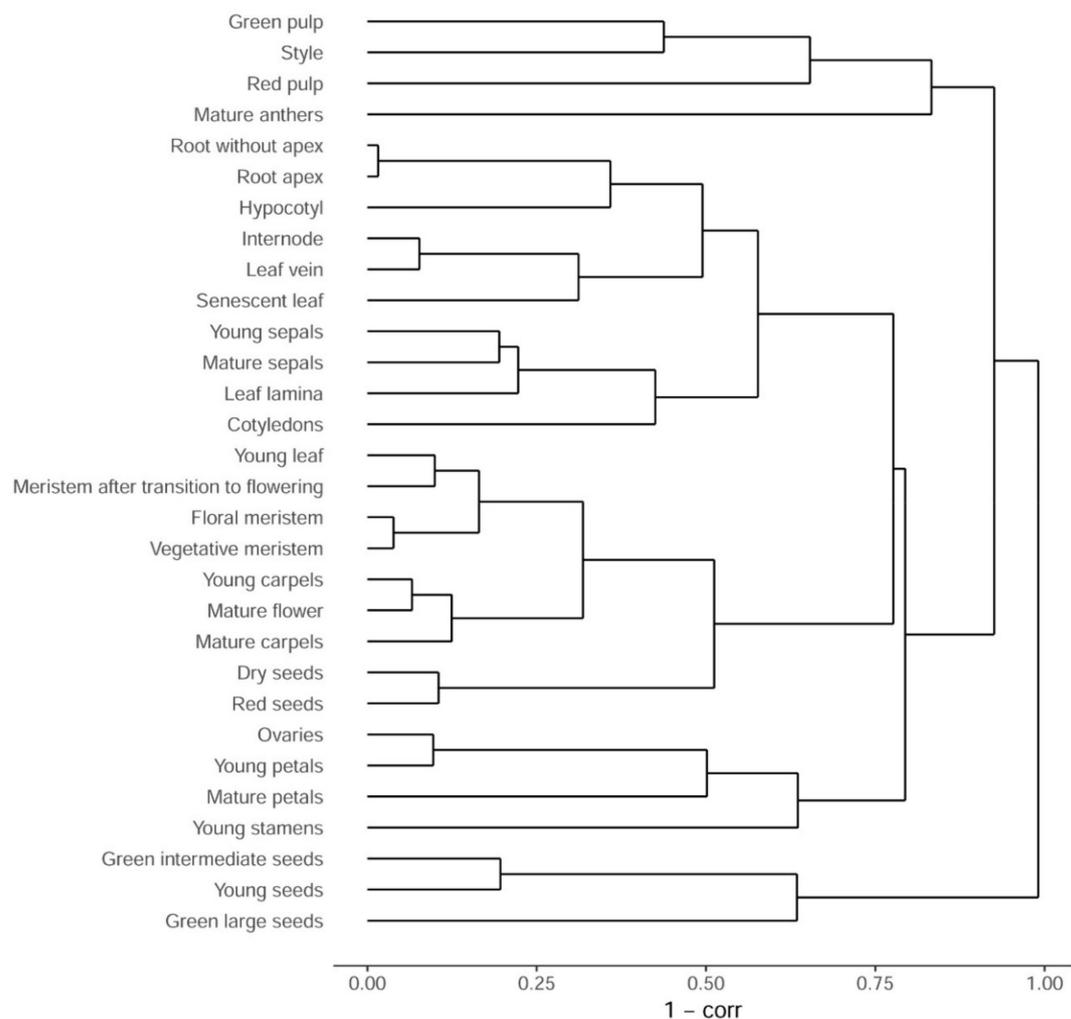
The raw data of the tomato transcriptome map have been deposited into the NCBI Sequence Read Archive (project ID: PRJNA507622).

### 3. Results and Discussion

#### 3.1. Sampling and Primary Analysis

The choice of samples for the tomato transcriptome map was based on clustering of *A. thaliana* transcriptome data from Klepikova et al. [3]. We selected the *Arabidopsis* samples that had the most dissimilar expression profiles based on the clustering tree of samples, and collected tomato samples that corresponded to these *Arabidopsis* samples (for example, anthers and senescent leaves). Assuming that expression profiles in homologous organs and/or corresponding developmental stages are similar in *Arabidopsis* and tomato, this approach would result in a set of tomato samples representing the maximum diversity of expression profiles.

The samples were sequenced with at least 20 million sequence reads were generated for each sample and read length of 75 and 60 bp (see Materials and Methods). Initial quality analysis showed a high congruence of the biological replicates: Pearson  $r^2$  correlation values for all replicates were between 0.79 and 1.0, with a mean value of 0.96 (median 0.98) (Supplementary Table S6), and a clustering tree of the replicates also indicated consistency of the data (Supplementary Figure S2). A hierarchical clustering tree of the samples reflected an organ- and age-specific structure (Figure 1). Most samples which are not replicates have highly divergent expression profiles ( $1-r^2 > 0.3$ ). This shows that the initial assumption was true and that our map indeed represents samples which are the most diverse in terms of expression profiles.



**Figure 1.** Hierarchical clustering of samples as represented by a clustering tree. Distance between samples is measured as  $1 - \text{Pearson } r^2$  correlation coefficient.

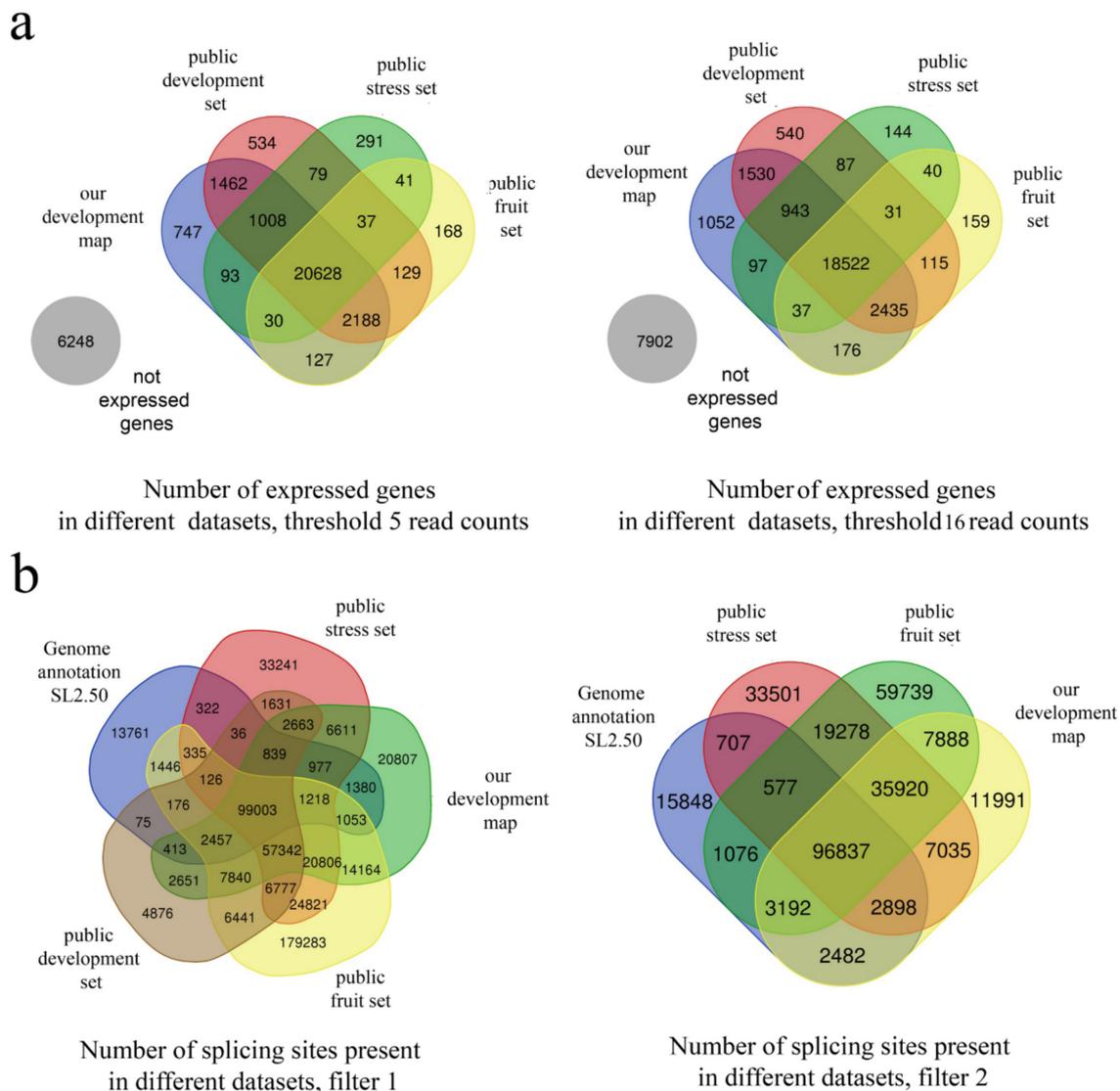
Annotation SL2.50 of the *S. lycopersicum* genome contains 33,810 coding genes. We used two thresholds to define genes as expressed in a certain sample: five normalized read counts in each of two replicates of the sample (weak threshold), and 16 normalized read counts for the strong threshold (as defined by Su et al. [17]). Using the weak threshold, 26,283 (78%) of genes were expressed in at least one sample (24,792 (73%) using strong threshold, Supplementary Table S7). In all samples with weak and strong thresholds, 13,517 (40%) and 11,669 (35%) genes were expressed, respectively (Supplementary Table S7). The lowest number of expressed genes (17,208 (51%) and 15,348 (45%) for weak and strong thresholds, respectively) was observed in the Sol.FL.r sample (red pulp), while the greatest number (20,805, 62% and 18,564, 55%) was observed in the Sol.SD.y sample (young seeds) (Supplementary Figure S3).

The splicing analysis demonstrates that the current annotation of the tomato genome lacks many splice sites. Our dataset reveals a high number of new splice sites. In contrast, only 10% of 123,617 previously known splice sites are not found in our data. Regarding new splice sites, even at the most stringent threshold, the number of new sites is twice as much as the number of annotated sites. The results of splicing are summarized in Table 1.

**Table 1.** Analysis of splicing sites based on newly obtained data.

|                         | Without Filtering | Filter 1 (Identification in Two Samples) | Filter 2 (Identification in Two Replicates) |
|-------------------------|-------------------|--|---|
| Introns, total          | 375,650           | 240,224                                  | 168,243                                     |
| Not annotated but found | 266,580           | 132,884                                  | 62,834                                      |
| Annotated but not found | 14,547            | 16,277                                   | 18,208                                      |

To assess the completeness of the transcriptome map in terms of the representation of expressed genes, we used three publicly available datasets that represent different biological processes and organs. The complete list of samples is presented in Supplementary Table S2. The first dataset—DEVELOPMENT—includes 19 samples (floral bud, leaf, petal, root, and different parts of the fruit at five stages of fruit maturity) in one replicate with a sequencing depth of 14–26 million reads [13]. The second dataset—STRESS—includes two sets of samples from biotic stress (*Cladosporium fulvum* infection-treated and control plants [23] and PRJNA419151). Each set is a time series collected in three replicates, and the sequencing depth ranges between 9.7–31 million reads. The third dataset—FRUIT—is a detailed expression atlas of the developmental dynamics of the tomato fruit [12]. It includes 49 samples in three replicates and 84 samples in four replicates. The sequencing depth is moderate, ranging from 3.6 to 25 million reads. Out of 483 samples, 183 have more than 10 million, and 367 have more than 7 million reads. The total number of genes expressed in the samples from these three datasets is 27,562 (under the threshold 5+5 reads); in our transcriptome map, we registered the expression of 26,283 genes (i.e., >95%). The same pattern is retained under a stronger threshold—16+16 reads: out of 25,908 genes expressed in these three datasets, and 24,792 genes are observed in our map. The expression of ~1000 genes is registered only in our dataset (Figure 2a); we assume that this is because several samples are unique in our map, e.g., meristems.



**Figure 2.** Estimation of the completeness of the expression map. (a) The number of expressed genes at different thresholds in three publicly available datasets and in our dataset; (b) the number of detected splice sites under different filters in three publicly available datasets and in our dataset.

Next, we assessed the number of samples in which each gene was expressed (Supplementary Figure S4). Most of the protein-coding genes tended to be expressed in all or almost all samples (16,326, 48% (14,378, 43%) genes were expressed in more than 25 samples), while some genes were expressed in a few samples (3365, 10% (3674, 11%) genes in 1–7 samples). We also investigated whether there was a correlation between the number of samples in which a gene was expressed and the expression level (Supplementary Figure S5). The mean and median expression levels across all samples were found to be higher for more widely expressed genes (i.e., those expressed in more samples). For maximum and minimum expression levels, the most widely expressed genes also exhibited a greater expression level, but the trend was not as prominent for these genes.

Analysis of splice sites using additional publicly available datasets shows that even in our dataset, many low-frequency sites remain unidentified. In particular, the addition of the detailed transcriptome map of fruit development results in a high number of additional splice sites (Figure 2b and Table 1). However, given low coverage of the data, they may represent artefacts.

### 3.2. Comparison with *Arabidopsis thaliana* Transcriptome Map

We compared the global parameters of the tomato transcriptome map with those of the *A. thaliana* map and found that, despite the difference in number of samples, they were similar in these two species. In particular, the distribution of the number of expressed genes and Shannon entropy (H) are similar, with the only difference being that in tomato, the peak at low entropy values is almost not visible (Supplementary Figures S6 and S7). The maximum entropy is 4.16. There are 12,641 genes with  $H \geq 3.7$ ; they are highly enriched in terms of being associated with basic cellular metabolism (Supplementary Table S8). At the lower end, there are 298 genes with  $H \leq 0.15$ . They are enriched in categories such as peroxidase activity (GO:0004601) and peptidase activity (GO:0008233) in molecular function or response to stress (GO:0006950) in biological processes or the cell wall (GO:0005618) in cellular components (Supplementary Table S8). All other global parameters (see Supplementary Figures S5, S8, and S9), such as distribution of maximum and minimum expression levels, DE score, and Z-score, are also almost identical in tomato and *Arabidopsis*.

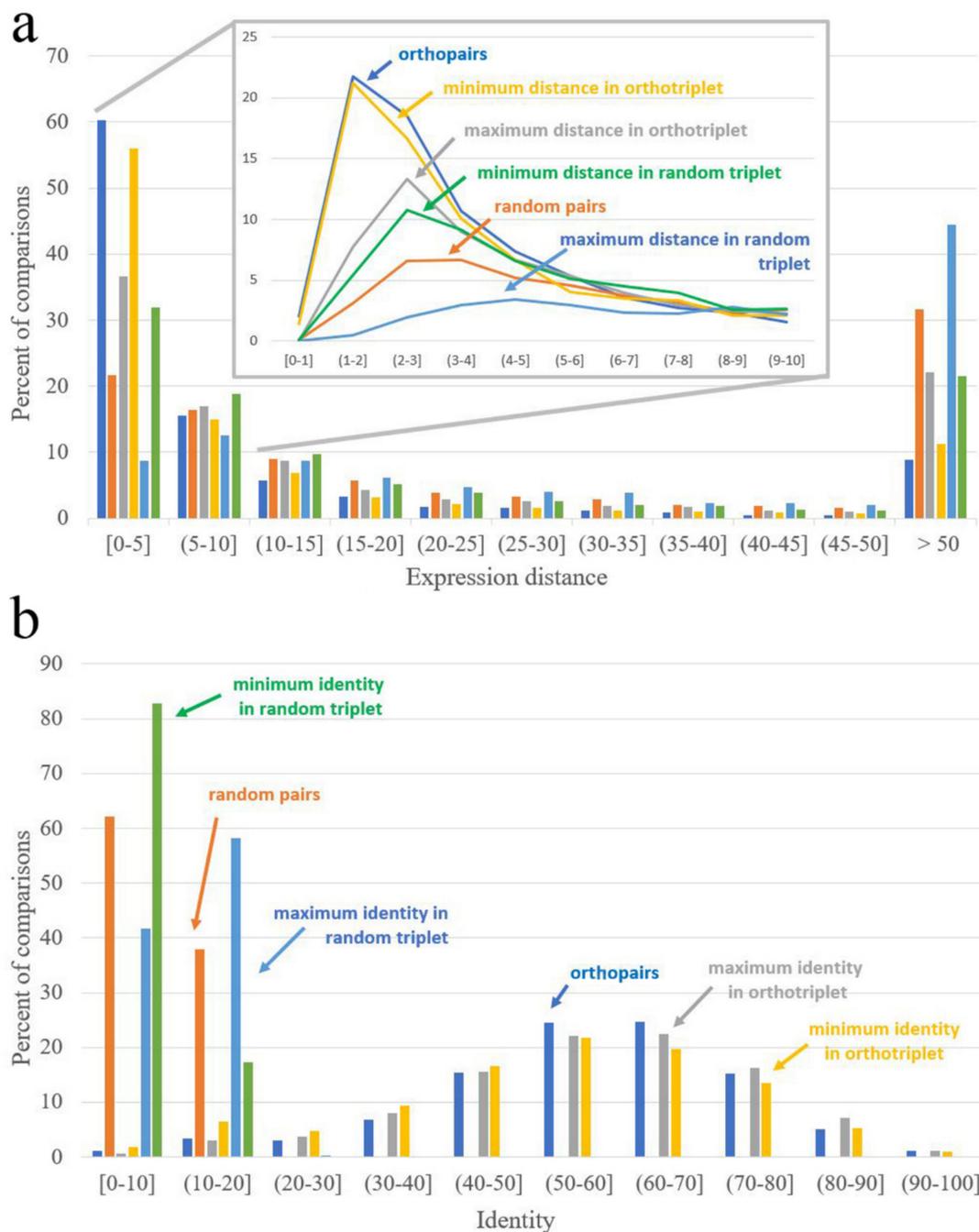
It is interesting to compare a set of genes that do not vary in expression between samples in *Arabidopsis* and in tomato. We considered only genes expressed in all samples; for each gene, a covariation was calculated. We found 123 genes with  $CV < 0.20$ , 657 with  $CV < 0.25$ , and 1527 with  $CV < 0.30$  (Supplementary Table S9). A set of genes with  $CV < 0.2$  was enriched by the categories related to transport, protein, nucleic acid localization, and kinases (Supplementary Table S10). Similar to *Arabidopsis*, the addition of publicly available RNA-seq data (sets DEVELOPMENT and STRESS) did not greatly decrease the number of stable genes (Supplementary Table S10). Unfortunately, the data from the fruit development atlas could not be used for the analysis of stable genes due to shallow sequencing depth that can lead to distortion in expression profiles (in particular, the underestimation of lowly expressed genes).

Analysis of GO enrichment of stable genes in *Arabidopsis* and tomato reveals similar categories: GO:0051169~nuclear transport, GO:0016192~vesicle-mediated transport, GO:0015031~protein transport, GO:0008104~protein localization, GO:0006886~intracellular protein transport, GO:0006497~protein amino acid lipidation, GO:0006403~RNA localization, GO:0006397~mRNA processing GO:0004386~helicase activity, GO:0042175~nuclear envelope-endoplasmic reticulum network, GO:0016023~cytoplasmic membrane-bounded vesicle, GO:0005794~Golgi apparatus, GO:0005654~nucleoplasm, and GO:0005635~nuclear envelope.

The identification of stably expressed genes is important for further studies that utilize quantitative PCR (qPCR) for the measurement of gene expression levels. It is well-known that the genes traditionally used as a reference in qPCR experiments (glyceraldehyde 3-phosphate dehydrogenase (GAPDH), actin, tubulin, etc.) are indeed not stable across conditions and organs [24], and each species and each experimental system requires selection and validation of the optimal reference genes [25]. The set of stably expressed genes identified in our study could be used as a basis for such a selection in tomato. Notably, tomato orthologues of two genes that were identified as the most stable in *Arabidopsis* are also among the most stable (Supplementary Table S11).

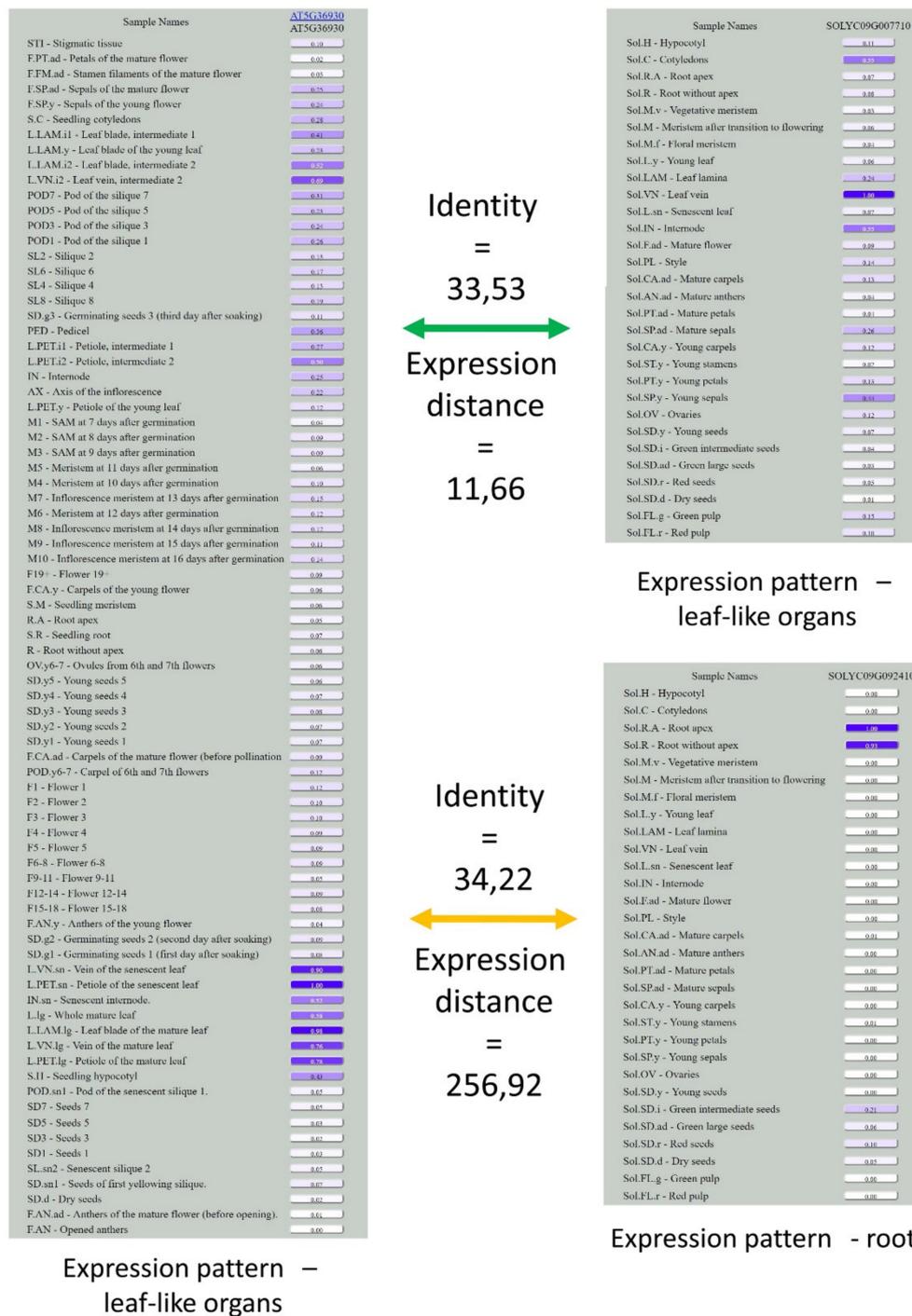
### 3.3. Analysis of Expression Patterns of Duplicated Genes

The most prominent feature of plant genomes is that they undergo multiple whole-genome or segmental duplications. Gene copies resulting from the duplication usually diverge in functions (alternatively, one of the copies can be lost). In cases when one gene of a model object is an orthologue of two paralogous genes from a non-model object, it is usually difficult to identify which of the co-orthologues retains ancestral function because both of them have a similar level of sequence identity. Indeed, we found that the distribution of the identity values for interspecific pairs within ortho-triplets (1 *Arabidopsis* gene–2 tomato genes and vice versa) is almost identical with the distribution of the identities for orthopairs. Even the distributions of minimal and maximal identity values are not drastically different (Figure 3 and Table 2).



**Figure 3.** Distribution of identity and expression distance for orthopairs, random pairs, interspecific pairs from ortho-triplets, and interspecific pairs from random triplets. (a) The distribution of expression distance; (b) the distribution of identity.

This means that in most cases of ortho-triplets, both paralogs from one species are equally similar to a single gene from the other species. In contrast, the similarity of expression profiles greatly differs for interspecific pairs—i.e., for most ortho-triplets, there is a pair with low expression distance (close to the distance typical for ortho-pairs) and a pair with high distance (see e.g., Figure 4 and Table 2).



**Figure 4.** Expression patterns in ortho-triplet AT5G36930, Solyc09g092410, and Solyc09g007710. Identities are almost equal, while expression distances differ ~20-fold. Note that while Solyc09g007710, similar to AT5G36930, is expressed in leaves, Solyc09g092410 has a drastically different expression profile, with a maximum expression level in roots.

In terms of function, this means that one of the co-orthologues in the ortho-triplet usually retains the ancestral function, while the other acquires a new function. Presumably, this occurs by the divergence of the regulatory elements of the paralogs after duplication. At the same time, sequence similarity at the level of protein-coding sequencing remains the same for both co-orthologues, and does not allow for conclusions on the function to be made.

**Table 2.** Medians of distributions of identity and expression distance.

| Expression Distance<br>Distance = 0 Corresponds<br>to Identical Expression Patterns |       | Identity<br>Identity = 100 Corresponds<br>to Identical Sequences |       |
|---|-------|--|-------|
| Orthopairs  | 3.68  | Orthopairs   | 58.48 |
| Random pairs  | 17.42 | Random pairs   | 8.40  |
| Minimal distance in interspecific<br>pairs from ortho-triplets                      | 4.12  | Maximal identity in interspecific<br>pairs from ortho-triplets   | 58.78 |
| Maximal distance in interspecific<br>pairs from ortho-triplets                      | 8.52  | Minimal identity in interspecific<br>pairs from ortho-triplets   | 55.01 |
| Minimal distance in interspecific<br>pairs from random triplets                     | 9.76  | Maximal identity in interspecific<br>pairs from random triplets  | 10.96 |
| Maximal distance in interspecific<br>pairs from random triplets                     | 37.64 | Minimal identity in interspecific<br>pairs from random triplets  | 6.18  |

### 3.4. Integration of the Solanum Transcriptome Map into the Database TraVA

Our expression data were integrated into the public database, Transcriptome Variation Analysis (TraVA, [3]). It allows for graphical representation of the expression profiles of single or multiple genes (delimited by spaces), both in absolute and relative to maximum values. A user can also choose a type of read count normalization. The results of differential expression analysis were also included in TraVA. A user can select a sample of interest from a list of samples and receive a colored visualization of fold changes for the analyzed gene, or genes between the selected sample and all other samples. The results of all types of analyses can be downloaded as Excel files.

## 4. Conclusions

The transcriptome map includes expression data for over 95% of the genes annotated in the tomato genome, allowing for the analysis of differential expression between organs and stages. Our previously developed *Arabidopsis* transcriptome map that has a similar sample structure allows one to conduct side-by-side comparison of gene expression profiles, and to reveal the cases of conservation and shift of function in co-orthologues.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2073-4425/10/1/50/s1>, Figure S1: Mapping of reads on FRUIT dataset, Figure S2: Hierarchical Clustering Tree, Figure S3: The number of genes expressed in samples, Figure S4: Distribution of genes by number of samples in which each gene was expressed, Figure S5: The distribution of gene expression level on number of samples in which a gene is expressed for minimum, mean, median and maximum expression levels of each gene, Figure S6: Distribution of Shannon entropy H, Figure S7: Distribution of Shannon entropy H in tomato and *Arabidopsis*, Figure S8: Distribution of DE Score, Figure S9: Distribution of Z-score for selected samples, Table S1: The description of samples collected and sequenced in our transcriptome map, Table S2: Publicly available data used for completeness test, Table S3: Splice junctions detected in analyzed datasets (separate file), Table S4: The sample combination for Shannon entropy calculation, Table S5: A sample combination for pseudo-Euclidean distance calculation, Table S6: Pearson squared correlation coefficient for samples reported in this study, Table S7: Genes expressed in all or at least 1 sample, Table S8: GO and other terms enrichment of genes with low and high H, Table S9: Number of stable genes, Table S10: GO and other terms enrichment of genes with CV < 0.2, Table S11: Orthopairs where both genes are stable.

**Author Contributions:** Conceptualization, A.A.P. and M.D.L.; Data curation, A.V.K.; Formal analysis, A.V.K., A.S.K. and E.S.G.; Funding acquisition, A.A.P.; Investigation, A.A.P. and A.V.K.; Methodology, A.A.P.; Project administration, A.A.P.; Software, A.V.K., A.S.K. and E.S.G.; Supervision, A.A.P.; Writing—original draft, A.A.P., A.V.K. and M.D.L.; Writing—review & editing, A.A.P., A.V.K. and M.D.L.

**Funding:** This research was funded by the Russian Science Foundation, grant number 17–14–01315, TraVA database was updated with support of budgetary subsidy to IITP RAS (Laboratory of Plant Genomics).

**Acknowledgments:** The authors express gratitude to D.D. Sokoloff for the help with tomato meristem collection.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kyojuzuka, J.; Konishi, S.; Nemoto, K.; Izawa, T.; Shimamoto, K. Down-regulation of *RFL*, the *FLO/LFY* homolog of rice, accompanied with panicle branch initiation. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 1979–1982. [[CrossRef](#)] [[PubMed](#)]
2. Kramer, E.M. Patterns of gene duplication and functional evolution during the diversification of the AGAMOUS subfamily of MADS box genes in angiosperms. *Genetics* **2004**, *166*, 1011–1023. [[CrossRef](#)] [[PubMed](#)]
3. Klepikova, A.V.; Kasianov, A.S.; Gerasimov, E.S.; Logacheva, M.D.; Penin, A.A. A High-resolution map of the *Arabidopsis thaliana* developmental transcriptome based on RNA-seq profiling. *Plant J.* **2016**, *88*, 1058–1070. [[CrossRef](#)] [[PubMed](#)]
4. FAO. Production of Tomatoes. FAOSTAT of the United Nations. 2016. Available online: <http://www.fao.org/> (accessed on 9 November 2018).
5. Bai, Y.; Lindhout, P. Domestication and breeding of tomatoes: What have we gained and what can we gain in the future? *Ann. Bot.* **2007**, *100*, 1085–1094. [[CrossRef](#)] [[PubMed](#)]
6. Leale, G.; Baya, A.E.; Milone, D.H.; Granitto, P.M.; Stegmayer, G. Inferring unknown biological function by integration of GO annotations and gene expression data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *15*, 168–180. [[CrossRef](#)] [[PubMed](#)]
7. Wren, J.D. A global meta-analysis of microarray expression data to predict unknown gene functions and estimate the literature-data divide. *Bioinformatics* **2009**, *25*, 1694–1701. [[CrossRef](#)] [[PubMed](#)]
8. Ma, M.; Liu, Z.L. Comparative transcriptome profiling analyses during the lag phase uncover *YAP1*, *PDR1*, *PDR3*, *RPN4*, and *HSF1* as key regulatory genes in genomic adaptation to the lignocellulose derived inhibitor HMF for *Saccharomyces cerevisiae*. *BMC Genomics* **2010**, *11*, 660. [[CrossRef](#)] [[PubMed](#)]
9. Zouine, M.; Maza, E.; Djari, A.; Lauvernier, M.; Frasse, P.; Smouni, A.; Pirrello, J.; Bouzayen, M. TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks. *Plant J.* **2017**, *92*, 727–735. [[CrossRef](#)]
10. Fernandez-Pozo, N.; Zheng, Y.; Snyder, S.I.; Nicolas, P.; Shinozaki, Y.; Fei, Z.; Catala, C.; Giovannoni, J.J.; Rose, J.K.C.; Mueller, L.A. The tomato expression atlas. *Bioinformatics* **2017**. [[CrossRef](#)]
11. Pattison, R.J.; Csukasi, F.; Zheng, Y.; Fei, Z.; Van der Knaap, E.; Catala, C. Comprehensive tissue-specific transcriptome analysis reveals distinct regulatory programs during early tomato fruit development. *Plant Physiol.* **2015**, *168*, 1684–1701. [[CrossRef](#)]
12. Shinozaki, Y.; Nicolas, P.; Fernandez-Pozo, N.; Ma, Q.; Evanich, D.J.; Shi, Y.; Xu, Y.; Zheng, Y.; Snyder, S.I.; Martin, L.B.B.; et al. High-resolution spatiotemporal transcriptome mapping of tomato fruit development and ripening. *Nat. Commun.* **2018**, *9*. [[CrossRef](#)]
13. Cárdenas, P.D.; Sonawane, P.D.; Pollier, J.; Vanden Bossche, R.; Dewangan, V.; Weithorn, E.; Tal, L.; Meir, S.; Rogachev, I.; Malitsky, S.; et al. GAME9 regulates the biosynthesis of steroidal alkaloids and upstream isoprenoids in the plant mevalonate pathway. *Nat. Commun.* **2016**, *7*. [[CrossRef](#)]
14. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [[CrossRef](#)] [[PubMed](#)]
15. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21. [[CrossRef](#)]
16. Anders, S.; Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **2010**, *11*, R106. [[CrossRef](#)] [[PubMed](#)]
17. Su, Z.; Łabaj, P.P.; Li, S.; Thierry-Mieg, J.; Thierry-Mieg, D.; Shi, W.; Wang, C.; Schroth, G.P.; Setterquist, R.A.; Thompson, J.F.; et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **2014**, *32*, 903–914. [[CrossRef](#)]
18. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [[CrossRef](#)]
19. Thomas, P.D. PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **2003**, *13*, 2129–2141. [[CrossRef](#)]
20. Mi, H.; Muruganujan, A.; Thomas, P.D. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **2012**, *41*, D377–D386. [[CrossRef](#)]

21. Schug, J.; Schuller, W.-P.; Kappen, C.; Salbaum, J.M.; Bucan, M.; Stoeckert, C.J. Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.* **2005**, *6*, R33. [[CrossRef](#)]
22. Emms, D.; Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **2015**, *16*, R157. [[CrossRef](#)] [[PubMed](#)]
23. Xue, D.Q.; Chen, X.L.; Zhang, H.; Chai, X.F.; Jiang, J.B.; Xu, X.Y.; Li, J.F. Transcriptome analysis of the Cf-12-mediated resistance response to *Cladosporium fulvum* in tomato. *Front. Plant Sci.* **2017**, *5*, 2012. [[CrossRef](#)] [[PubMed](#)]
24. Czechowski, T. Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol.* **2005**, *139*, 5–17. [[CrossRef](#)] [[PubMed](#)]
25. Gutierrez, L.; Mauriat, M.; Pelloux, J.; Bellini, C.; Van Wuytswinkel, O. Towards a systematic validation of references in real-time RT-PCR. *Plant Cell* **2008**, *20*, 1734–1735. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).