Available online freely at www.isisn.org

Bioscience Research

Print ISSN: 1811-9506 Online ISSN: 2218-3973

Journal by Innovative Scientific Information & Services Network

RESEARCH ARTICLE BIOSCIENCE RESEARCH, 2019 16(3): 2641-2654. **OPEN ACCESS**

A proposed RNA-seq analysis workflow to study heat-stress genes in *Arabidopsis thaliana*

Heba Zaki[1], Mohammad Nassef [2], Ahmed Farouk Al-Sadek[1] and Amr Ahmed Badr [2]

[1]Agricultural Research Center, Giza, **Egypt**
[2]Department of Computer Science, Faculty of Computers and Information, Cairo University, **Egypt.**

***Correspondence:** hhamoda2010@gmail.com **Accepted:** 25 July 2019 **Published online:** 03 Aug 2019

The degree to which certain genes are relevant to biological inquiries can be an open question. State-of-the-art computational methods can help understanding functional associations between gene expression patterns and subsequent biological experiments. Therefore, the basic scientific need to correlate significant differential expression levels to biological variation phenomena is highly demanded. RNA-sequencing (RNA-seq) methods employ next-generation sequencing (NGS) technology towards scanning RNA molecules in samples and quantify their amount. Development of crops that can overcome environmental stresses, while maintaining productivity, proved to be a basic necessity for agricultural productivity. Arabidopsis thaliana is an ideal model organism for studying biologically relevant questions about global gene regulation in response to stresses. The main purpose of this study is to identify differentially expressed genes in *A. thaliana* under heat-stress conditions. A workflow for RNA-seq analysis is proposed to identify these genes using; edgeR and Fisher criterion (FC) analysis methods. The identified candidate genes are validated via two popular references; DRASTIC and TAIR10. Results suggest that these two methods can be combined to perform differential expression analysis within RNA-Seq data, without strong assumptions. Comparative evaluation of the proposed methods demonstrates successful identification of stress-related genes, with improved prediction accuracy. This shows that presented workflow and the differential analysis methods can be applied to identify differentially expressed genes from RNA-seq data for other organisms. Finally, literature based verification for the top 5% detected genes shared between FC and edgeR methods is demonstrated. Suitable justification is given to help discover newly response-related genes to heat phenomenon.

Keywords: RNA-seq, gene expression, differential expression analysis, stress genes, Arabidopsis thaliana.

INTRODUCTION

To increase crop productivity subject to shrinking cultivable land and natural resources has become a vital goal for agricultural scientists and economists alike. However, environmental stress factors like drought, salinity, high and low temperatures, high light, together with biotic factors like pests and diseases can reduce agricultural gains significantly. Notably, these factors can affect the quality and yield of harvest production gradually. Investigation of the molecular mechanisms that underlie stress resistance is considered a first step towards the goal of producing crops that demonstrate resistance to abiotic stress. Towards understanding plant stress responses, it is much needed to understand the mechanisms of stress responsive genes regulation (Shameer, K. et al., 2009).

Global transcriptomic analysis can be performed using mircoarrays or next-generation sequencing. They help understand the functional associations depending on proper expression patterns of genes to coordinate subsequent

biological experiments (Moreno-Risueno, M. et al., 2010; Less, H. et al., 2011; Friedel, S. et al., 2012). The classical transcriptomic data analysis workflows concentrate on estimating the adjustments in expression of each single gene. This is called differential expression (DE) analysis that utilizes theory testing to quantify the statistical significance of a watched expression change. This significance is based on matching between-sample (condition) variation and within-sample (replicate) variation (De la Fuente, A., 2010).

Recent advances in NGS reduced sequencing costs to a great deal, making it more feasible to create a large volume of sequencing data effectively and efficiently. Sequencing data forms a large number of short sequence fragments repository that needs to be processed through a set of phases before performing relative abundance estimation (Quinn, T.P. et al., 2018).

Even with sequencing cost reduction, RNA-seq experiments can still be expensive for small-budget research projects. As a result of constraining RNA-seq studies to just a few libraries, there is often limited replication. Hence, it is imperative to estimate biological variation as reliably as possible from small number of replicate libraries. This issue can be amplified by the fact that different genes or transcripts may have distinctive degrees of biological variation (McCarthy, D. J. et al., 2012).

Combining discovery and quantification steps in single high-throughput sequencing yields a powerful method of sequencing RNA called RNA-seq.

Technical variation is associated with the sequencing technology whereas biological variation refers to changes in expression levels between experimental subjects. Information is shared between genes to estimate biological variation reliably even when the number of replicates is very small. One very common issue is how to use the read counts to detect differentially expressed genes between different experimental conditions (Chen, Y. et al., 2011).

Arabidopsis thaliana has been used to study plant changes for more than fifty years and for genetic analysis. Most recently, *A. thaliana* is preferably used as a main model organism to consider distinctive parts of plant science, particularly for such branches as molecular biology, genetics and genomics (Swarbreck, D. et al., 2008).

In order to gain the maximum benefits of RNA-seq, computational techniques are required to fulfill transcriptome assembly. There are two main approaches for transforming RNA-seq raw data into transcript sequences; either the approach of genome-guided or via de novo assembly. The genome-guided approach for transcriptome studies has rapidly turned into a standard method to deal with RNA-seq analysis for model organisms like *A. thaliana*. There is a number of software packages used to serve this purpose (Haas, B. et al., 2013). This research follows RNA-seq analysis workflow to identify genes expressed in *A. thaliana* in response to heat-stress factors on the plants.

A computational system for network centric transcriptome analysis was presented for detecting biologically important genes that were collected from seedling root and shoot tissues of *A. thaliana* under stress conditions (Ma, C. et al., 2014). The positive samples for training the ML-based prediction models were known stress related genes from the DRASTIC and TAIR databases.

STIF search algorithm enabled the identification of predicted sites upstream of plant stress genes (Sundar, A. et al., 2008). The dataset of 60 stress-up regulated genes were identified within five reference databases; RARGE, DRASTIC, StressLink, AtGenExpress, DATF, and TAIR which were used also for the validation study.

STIFDB provided extensive information about various stress responsive genes and stress inducible transcription factors of *A. thaliana* (Shameer, K. et al., 2009). Sequence segments of these genes were obtained from TAIR, used to access the gene-based contents. Moreover, gene expression databases like NASC, DRASTIC, RARGEMAEDA, and the StressLink Database were used to get those genes.

A complete workflow of DE and pathway analysis using the edgeR quasi-likelihood pipeline was presented (Chen, Y. et al., 2016). Computation steps of the analysis pipeline were performed using R software packages.

The edgeR and DESeq2 standard log-ratio transformation-based methods efficiently measure DE from RNA-seq data, while certain assumptions are met (Quinn, T.P. et al., 2018). Results confirmed that those methods have high precision in simulations and perform well on real data too.

Computational framework for huge datasets RNA-seq with no dependence on transcript annotations implemented exact and productive DE at alternative splicing variants identified automatically (Hu, Y., 2013). It was indicated that no need for full transcript quantification and

reconstruction.

A guideline for RNA-seq data analysis was explored with a review for all the major steps in RNA-seq data analysis (Conesa, A. et al., 2016). It demonstrated generic roadmap for experimental design and analysis using standard Illumina sequencing.

Examination test of ALDEx2 and other transformation-based methods demonstrated that ALDEx2 runs much slower than edgeR and other methods (Quinn, T.P. et al., 2018). Additionally, ALDEx2 cannot provide a well-documented simplification for mixed models.

Cufflinks, IsoEM, RSEM, and HTSeq RNA-seq expression quantification tools were examined for their performance (Chandramohan, R. et al., 2013). It was shown that Cufflinks, RSEM, and IsoEM tools have some limitations regarding correlation with Quantitative reverse transcription PCR (RT-qPCR) measurements in comparison to HTSeq tool. However, higher accuracy of the expression values can be obtained by applying the first three tools.

A machine Learning based methodology for transcriptome analysis mlDNA implemented as R package to re-analyze a set of abiotic stress expression data in *A. thaliana* (Ma, C. et al., 2014). The mlDNA demonstrated notable success in identifying stress related genes using traditional statistical testing–based DE analysis, with noticeably improved prediction accuracy. Although, this research dealt with six types of abiotic stress, it was made clear that it concentrated only on salt stress details as it gave the best work results. Moreover, it did not present the resulting set of stress related genes that were identified for each type of the mentioned six abiotic stresses to achieve credibility.

Hands on Training in RNA-seq Data Analysis were proposed for a general workflow to carry out a RNA-seq experiment (I-Hsuan. Lin., 2016). It offered comprehensive steps for mapping and analysis of given two adult female cell lines datasets. The differential gene expression analysis was performed using edgeR. However, results of that work was not validated or checked via trusted reference databases.

MATERIALS AND METHODS

The Proposed RNA-seq Analysis workflow is divided into two main phases: (1) Generation of Expression Matrix, and (2) Differential Expression analysis. A schematic overview of the workflow is detailed in Figure 1 to clearly describe smooth navigation between the two phases. All processes and steps of this workflow do not have dependencies on the structure of the entry datasets or the reference genome and annotations. Additionally, no certain assumptions or parameters are required to apply this workflow on other datasets.

Datasets, Software Packages and Computational Requirement

Datasets and Experiment Description

A. thaliana reference genome FASTA sequences and annotation GTF files can be downloaded from the Ensembl FTP (https://plants. ensembl.org/info /website/ftp/index.html) as shown in Table 1.

Table 1: Arabidopsis Thaliana reference genome and annotation files

File Name	Size
Arabidopsis _thaliana. TAIR10.dna. toplevel.fa.gz	36 MB
Arabidopsis_thaliana. TAIR10.41.gtf.gz	10 MB

RNA-seq FASTQ data files for *A. thaliana* under heat-stress were downloaded from the NCBI website. The experiment data that is covered in this study collected raw reads of plants in Moscow, Russia. A Third leaf was collected from 15 plants of age 21 days after heat treatment at 42°C for 1, 3, 6, 12, and 24 hours. The experiment was performed with 2 replicates for each of the mentioned 5 different time points. This experiment was SINGLE stranded – Illumina Hi-Seq 2000 – RNA-seq libraries from TRANSCRIPTOMIC PolyA RNA. Ten files that were downloaded are listed below with their accession, description, and size from (https://www.ncbi .nlm.nih.gov/sra/?term=Arabidopsis+th aliana+Heat+%2Bmoscow) as shown in Table 2.

Software Packages and Tools

The following list contains software tools and packages that were integrated with custom code to carry out the execution of the various processes along the presented workflow.

STAR (Spliced Transcripts Alignment to a Reference): 2.5.3a [March 17, 2017] version available on BA-HPC.

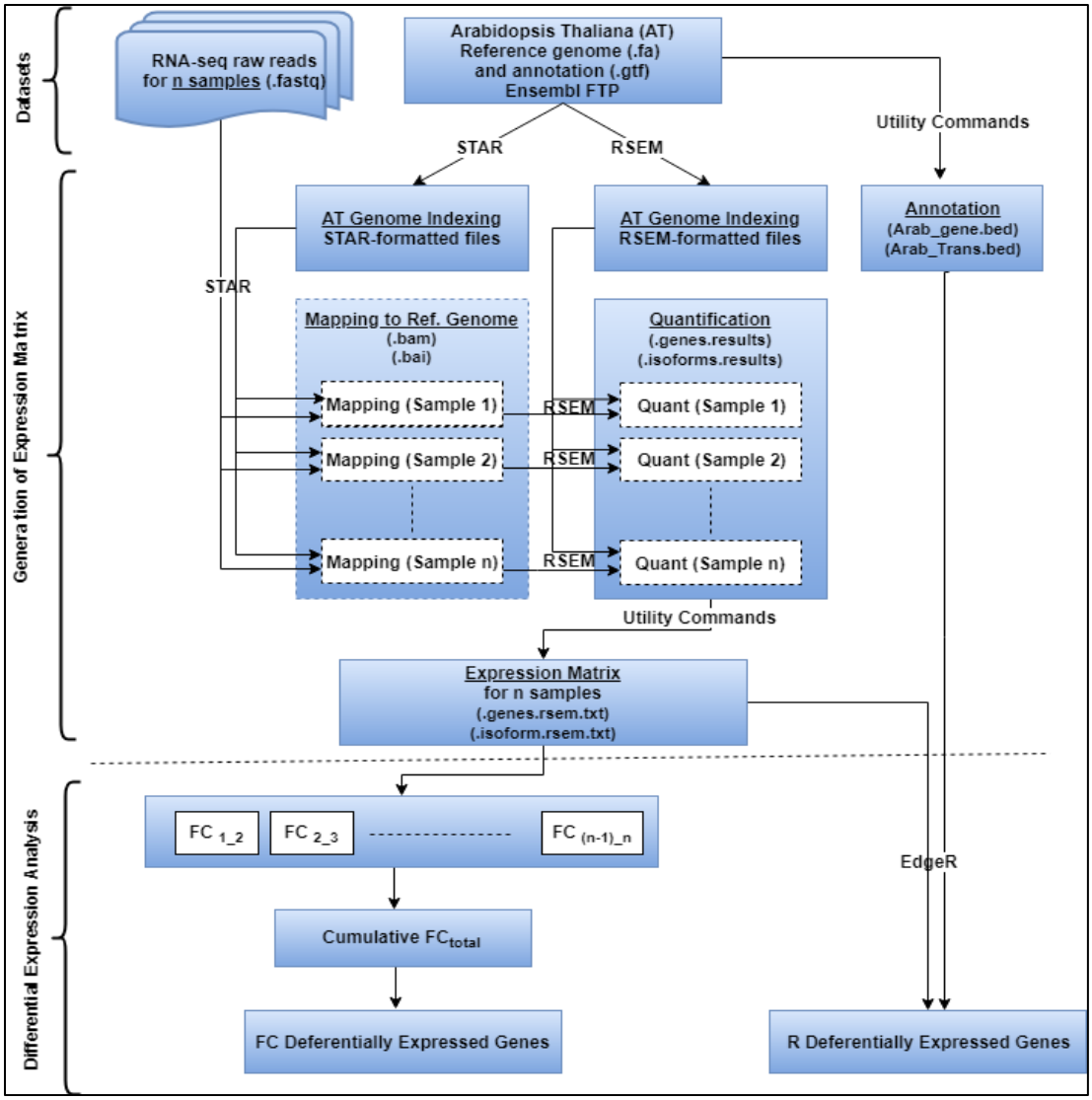

Figure 1: RNA-seq analysis Workflow

Table 2: **Data files of RNA-seq FASTQ raw reads**

Symbol	Accession	File Name	Description	Size
1h_ Rep1	SRX1881868	SRR3724768.fastq.gz	Heat Treatment 1 hour Replicate 1	532 MB
1h_ Rep2	SRX1881876	SRR3724774.fastq.gz	Heat Treatment 1 hour Replicate 2	73 MB
3h_ Rep1	SRX1881880	SRR3724778.fastq.gz	Heat Treatment 3 hours Replicate 1	1.4 GB
3h_ Rep2	SRX1881883	SRR3724782.fastq.gz	Heat Treatment 3 hours Replicate 2	1.16 GB
6h_ Rep1	SRX1881886	SRR3724785.fastq.gz	Heat Treatment 6 hours Replicate 1	1.38 GB
6h_ Rep2	SRX1881888	SRR3724786.fastq.gz	Heat Treatment 6 hours Replicate 2	1.46 GB
12h_Rep1	SRX1881889	SRR3724787.fastq.gz	Heat Treatment 12 hours Replicate 1	1.64 GB
12h_Rep2	SRX1881897	SRR3724798.fastq.gz	Heat Treatment 12 hours Replicate 2	1.63 GB
24h_Rep1	SRX1881908	SRR3724806.fastq.gz	Heat Treatment 24 hours Replicate 1	1.61 GB
24h_Rep2	SRX1881912	SRR3724814.fastq.gz	Heat Treatment 24 hours Replicate 2	1.39 GB

RSEM (RNA-seq by Expectation-Maximization): 1.2.30 [May 15, 2016] version available on BA-HPC.

edgeR: Package R version 3.3.2 [2016-10-31].

Computational Requirements

Unix-type operating systems (primarily Linux); it provides a command-line interface and is best run on a high-memory, multicore computer or in a high-performance computing environment. In general, having ~1 GB of RAM per 1 million paired-end reads is recommended. A typical configuration is a multicore server with 256 GB to 1 TB of RAM.

For the research problem presented here, the lack of required computing resources to accomplish the required work can be a challenge. In this project, the computational resources that were used were provided by The Bibliotheca Alexandrina (bibalex) (BA-HPC group, 2018). The super computer BA-HPC capabilities are used to achieve this work.

Generation of Expression Matrix

Creation of BED annotations and Genome Indices

The annotations recorded in the GTF files are converted into two BED-formatted files; one for genes and the other for transcripts, as shown in Table 3.

Table 3: Arabidopsis Thaliana annotation files in BED format

File Name	Size
Arab_Thail_gene.bed	1.8 MB
Arab_Thail_transcript.bed	0.4 MB

Then, the reference genome FASTA file is used to create the genome indices via both STAR and RSEM tools.

STAR: 15 STAR-formatted files generated. Those indices will be used into the step of mapping raw reads to reference genome.

RSEM: 7 RSEM-formatted files generated which will be used into the step of expression quantification.

Mapping raw reads to reference genome

The step of mapping of the RNA-seq reads aims to find matches between the reference genome and the sequences of the sampled short reads. STAR is one of the most popular RNA-seq mappers. It is adjusted to identify non-canonical splice junctions or map long-reads (Conesa, A. et al., 2016). Using indices files generated by STAR in the previous step, each *individual* read with the reference genome is mapped. BAM files are generated sorted by coordinates. Moreover, *'Log.final.out'* file is generated which shows some statistics of mapping process such as *Number of input reads, Mapping speed (Millions of reads per hour), Percentage of uniquely mapped reads, Average mapped length, Percentage of unmapped reads, etc*.... These statistics are useful for quality control and some of them are shown in Table 4.

Table 4: some mapping statistics in *'Log.final.out'*

Sample Name	Number of input reads	Uniquely mapped reads %
1h_ Rep1	5442791	92.97
1h_ Rep2	761850	93.76
3h_ Rep1	14381738	79.94
3h_ Rep2	11927547	92.74
6h_ Rep1	14136645	88.9
6h_ Rep2	14925137	93.35
12h_ Rep1	93.35	91.1
12h_ Rep2	16758327	94.15
24h_ Rep1	16442117	93.09
24h_ Rep2	14216929	85.93

Expression Quantification

In essence, quantification involves “counting” the number of times a sequence aligns to a specific part of the reference. The counts are represented as a matrix of describing the estimated frequency of each transcript is presented for each sample under study. RSEM software implements the Expectation Maximization (EM) algorithm which estimates the related abundances of the transcripts (Quinn, T.P. et al., 2018). In the previous step, STAR was used to output genomic alignments in transcriptomic coordinate *'Aligned.to Transcriptome.out.bam'*. This file and the generated indices using RSEM were passed in to quantify the gene and transcript expression levels for each mapped read. RSEM generates two result files for each sample or replicate representing the expected RNA-seq fragments assigned to all genes and isoforms exist in that sample as the example file of 12 hr rep 2 is shown in Table 5 and Table 6.

Building Expression Matrix

This section describes the preparation of gene-level and transcript-level expression

matrices. All *'rsem.genes. results*' files for all replicates are merged side-by-side, and then the columns containing the *(expected_ count)* information are selected, and placed into ONE final output file (Gene-level:*' genes.rsem. txt*'). The same is done for *'rsem.isoforms.results'* files for all replicates and ONE final output file (Transcript-level: *'isoforms.rsem.txt*'). Table 7 shows the abundance estimation by RSEM for all replicates at Gene-level.

Table 5: some results from file *'rsem.genes.results'* for sample '12h_Rep2'

Gene Id	Transcript Id(s)	Length	Effective Length	Expected Count
AT1G01010	AT1G01010.1	1688	1639	106
AT1G01020	AT1G01020.1,AT1G01020.2,AT1G01020.3,AT1G01020.4,AT1G01020.5,AT1G01020.6	1196.51	1147.51	328
AT1G01030	AT1G01030.1,AT1G01030.2	1905	1856	47
AT1G01040	AT1G01040.1,AT1G01040.2	6161.97	6112.97	1233.49
AT1G01046	AT1G01046	207	158	18
AT1G01050	AT1G01050.1,AT1G01050.2	994	945	734
AT1G01060	AT1G01060.1,AT1G01060.2,AT1G01060.3,AT1G01060.4, AT1G01060.5,AT1G01060.6,AT1G01060.7,AT1G01060.8	2618.03	2569.03	37
AT1G01070	AT1G01070.1,AT1G01070.2	1536.49	1487.49	87

Table 6: some results from file *'rsem.isoforms.results'* for sample '12h_Rep2'

Transcript Id	Gene Id	Length	Effective Length	Expected Count
AT1G01010.1	AT1G01010	1688	1639	106
AT1G01020.1	AT1G01020	1329	1280	0
AT1G01020.2	AT1G01020	1087	1038	198.38
AT1G01020.3	AT1G01020	1420	1371	115.69
AT1G01020.4	AT1G01020	1397	1348	13.93
AT1G01020.5	AT1G01020	1306	1257	0
AT1G01020.6	AT1G01020	944	895	0
AT1G01030.1	AT1G01030	1905	1856	47
AT1G01030.2	AT1G01030	1836	1787	0
AT1G01040.1	AT1G01040	6276	6227	897.39

Length: length of the reconstructed transcript.
Effective_length: transcript_length – mean_fragment_length + 1 .
Expected_count: number of expected RNA-seq fragments assigned to the transcript given maximum – likelihood transcript abundance estimates.

Table 7: **The content of the created data matrix for some genes at Gene-level**

Gene Id	1h_ Rep1	1h_ Rep2	3h_ Rep1	3h_ Rep2	6h_ Rep1	6h_ Rep2	12h_ Rep1	12h_ Rep2	24h_ Rep1	24h_ Rep2
AT1G01010	25	0	60	59.01	46	48	79	106	91	62
AT1G01020	56	9	61	163	150.05	216	198	328	200	152
AT1G01030	49	3	21	94	42	20	49	47	25	12
AT1G01040	139	13	289	334	240.28	639.6	872.11	1233.49	967	659.28
AT1G01046	0.5	0.5	0.5	0.5	1.5	11	12	18	8.5	4.5
AT1G01050	305	39	432.75	443	575	675.98	453.9	734	346.91	269.72
AT1G01060	2607.95	199	4620.58	401	2789.41	72.01	670.95	37	11	13
AT1G01070	10	0	10	12	16	56	51	87	51	53
AT1G01080	365	44	377	318	506.02	189	236	305.99	47	33

Differential Expression analysis

The alignment and quantification processes produce a count matrix that is the most commonly used format in DE analysis. There are many methods for DE analysis; DESeq and edgeR seem to be the most popular of those methods (Robinson, M. D. et al., 2010).

After completion of the phase 'Generation of Expression Matrix', analysis of the resulting matrix aims to detect the highly expressed genes which obviously affected by heat-stress.

In this work, two different methods are utilized in identification of differentially expressed genes and transcripts; FC, and edgeR methods. Each method tries to find the set of highly expressed genes to upgrade and define them as heat-stress genes.

Fisher criterion (FC)

Many researchers suggested using the Fisher's exact test, a likelihood ratio test, or t-statistics as an approximation to test whether a gene is differentially expressed between their two samples (Bullard, J. et al., 2010). FC is a popular statistical approach that measures separation between estimates of different classes. It is going to be used here, as a method for measuring the DE genes and transcripts within the resulting matrix. Detecting the overly expressed genes that are affected by the heat-stress is the main goal.

The following set of equations (1), (2), (3), and (4) describes the calculation of FC for genes resulting from the expression matrix *(Expected_count)*. FC is determined for each two consecutive time points, FC_{i-j} (2), taking into consideration that each time point has set of replicates that should be involved to correctly measure the variance between samples. The complete FC of each gene, FC_{total} (1), is the cumulative of all sub FC for given time points for that gene.

$$FC_{\text{total}} = FC_{1\_2} + FC_{2\_3} + \cdots + FC_{(n-1)\_n} \quad (1)$$

$$FC_{i-j} = (M_i - M_j)^2 / (SD_i^2 + SD_j^2) \quad (2)$$

Where,

- FC_{i-j} : Fisher criterion between time points $(i, \ and \ j)$
- M_i: mean of sample replicates at time point (i)
- SD_i: standard deviation at time point (i) (SD^2: variance of sample replicates)
- i, j: time points

$$M_i = \sum (Rep_{ik}) \ / \ m \quad (3)$$

$$SD_i^2 = \sum (Rep_{ik} - M_i)^2 \ / \ (m-1) \quad (4)$$

Where,

- m: number of replicates per sample at time point(i)
- k: replicate number $k = 1 \ldots m$

edgeR Analysis

This is the second method used to determine the overly expressed genes that are affected by the heat-stress. The following script is created by the assistance of edgeR software package. *Empirical Analysis of Digital Gene Expression Data in R (edgeR):* its main purpose is the DE analysis of RNA-seq expression profiles with biological replication. The edgeR implements a range of statistical methodology based on the negative binomial distributions, including empirical Bayes estimation, exact tests, generalized linear models and quasi-likelihood tests (McCarthy, D. J. et al., 2012).

Filtering out low levels expressed genes prior to DE analysis reduces the need for correction and also improves the detection power. Some methods, such as the well-known edgeR, take as input raw read counts and present possible preference sources into the statistical model to perform an integrated normalization along with DE analysis (Conesa, A. et al., 2016).

The book chapter in (Chen, Y. et al., 2011) explains the *'estimateDisp'* function and the weighted likelihood empirical Bayes method. Different genes show different levels of variability, but the number of replicate samples from which variability is estimated can be very small indeed.

The following steps summarize the performed edgeR script:

Load gene expression data, gene-level and transcription-level annotations
Merge (Gene and Isoforms) with (Annotation)
Filter lowly expressed genes/transcripts and recompute the library sizes
Calculate normalization factors using TMM normalization to scale the raw library sizes
Estimating dispersion
Differential expression: quasi-likelihood F-test

Output files:

DE_analysis.gene.csv: are the set of overly expressed genes in the given quantified samples.
DE_analysis.transcript.csv: are the set of overly expressed transcripts in the given quantified samples.

RESULTS

After applying all steps of a workflow run, the output file *(genes.rsem.txt)* contained (34,218) differentially expressed genes. The next step is to discover the highest differentially expressed genes from this set based on the two methods described above. Additionally, a reasonable set of comparisons between results of the two analysis methods is held to provide meaningful justification of these results.

Heat stress Genes References

In this study, two different references are used to validate the results from the differentially expression matrix. Known heat-stress related genes for *A. thaliana* can be collected from two important references (Ma, C. et al., 2014):

1- TAIR10 (The Arabidopsis Information Resource) (TAIR team, 2019)

This reference maintains a database of genetic and molecular biology data for the higher model *A. thaliana* plant. Heat-stress genes from TAIR10 were retrieved based on **"***heat*" keyword search and returned (188) different genomic loci.

2- DRASTIC (Database Resource for the Analysis of Signal Transduction in Cells) (Gary Lyon, 2018)

A manually derived database of plant expressed sequence tags and genes up- or down-regulated in response to various pathogens (biotic stress), chemical treatments, and abiotic stress such as drought, salt, heat and cold. After searching, (43) non redundant genes are found for *A. thaliana* that were mostly experimentally validated to be heat-stress.

The intersection between TAIR10 and DRASTIC databases are only (6) genes. This affects the explanation of results gained and it is discussed later.

Differential Expression Analysis Using FC

FC method was used to detect the most differentially expressed genes in the (34,218) differentially expressed genes. Calculate FC for those genes by applying the equation (1) using the following formula:

$$FC_{total} = FC_{1-3} + FC_{3-6} + FC_{6-12} + FC_{12-24}$$

Given that the data samples have 5 time points, the FC score was calculated for each 2 sequential time points. Then, by accumulation, the total FC can be determined to estimate the actual reflection of DE for all genes. The resulting DE genes were ordered according to their total FC in descending order to get the highest differentially expressed ones. All genes that have FC_{total}= 0 were filtered out as the least expressed genes. After applying this filtering step, the total number of differentially expressed genes has become (30,959) genes. In Figure 2, the distribution of some genes that have the highest Total FC is shown.

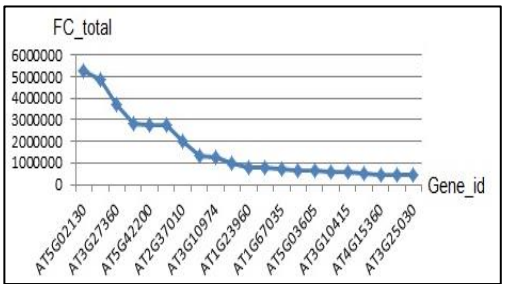

Figure 2: Distribution of some highest FC_{Total} genes predicted by FC

After arranging genes, they are matched to the two references of TAIR10 and DRASTIC databases. Table 8 shows the predicted differentially expressed genes as by FC and their intersection with DRASTIC and TAIR10 ordered by Top (%)

Differential Expression Analysis using edgeR

The edgeR software contains many features, options, and opens up flexible possibilities for RNA-seq data analysis. When applying edgeR to the experimental data, DE analysis is performed

by applying these functions in order: *DGEList, merge, calcNormFactors, estimateDisp, glmQLFit, glmQLFTest,* and *decideTestsDGE.*

The significance level (p-value) of DE was calculated with t-test, and Limma using R/Bioconductor package. The used p-value cutoff was p.value=0.01.

The edgeR script described above is used to detect the most differentially expressed genes in the (34,218). The resulting files, 'DE_analysis.gene' and 'DE_analysis.transcript', contain only (18,541) differentially expressed genes after edgeR script execution. Moreover, F_{value}, for each gene is generated for each sample or replicate in such files. Correspondingly, genes that have F_{value}= 0 have been excluded to filter out the least expressed genes. After applying this filtration, the total number of differentially expressed genes has become (18,517) genes and some of highest F_{value} genes are shown in Figure 3.

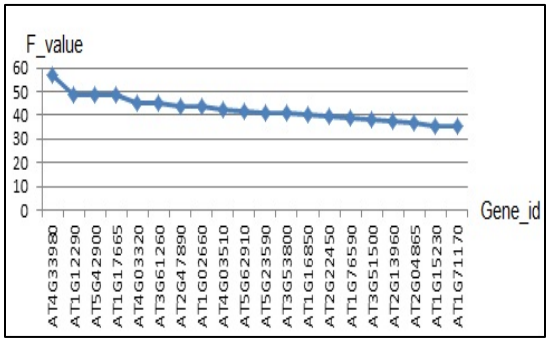

Figure 3: Distribution of some highest F_{value} genes predicted by edgeR

Similarly, after ordering genes by (F_{value}) in descending, they are matched to the two references of TAIR10 and DRASTIC databases. Table 9 shows the predicted differentially expressed by edgeR and their intersection with DRASTIC and TAIR10 ordered by Top (%).

Comparisons

The predicted heat-stress genes using both FC and edgeR show the effectiveness of both DE methods. An additional verification step by other methods is performed to cover the evaluation of the presented results and their relation to other trusted sources, and to each other.

Table 10 handles the intersection between equal slices (same percentage) genes generated by both edgeR, and FC methods gradually from 1% up to 100%. For example, the first row in the table, the highest (1%) genes in FC and edgeR are declared. As shown, for edgeR (1% of 18,517 = 185 genes) and for FC (1% of 30,959 = 309 genes). The intersected gene between these two sets is only one gene [AT5G23240]. Similarly, the second row compares top 5% of edgeR (925 genes) and FC (1547 genes) has 26 [AT5G42200, AT5G23240, AT5G05440, AT5G13220, AT1G02640, AT2G17840, AT2G47410, AT1G56220, AT4G37990, AT3G11590, AT2G33050, AT5G62200, AT1G67310, AT5G19140, AT2G43540, AT4G30780, AT1G03070, AT1G20620, AT1G03220, AT1G14970, AT3G13784, AT3G53990, AT5G16260, AT4G38580, AT3G04910, AT3G18800] intersected genes.

Figure 4 combines all intersections between the heat-stress genes generated by FC, and edgeR DE methods and their relations to the reference databases DRASTIC, and TAIR10. In Figure 4.(A), all the (43) DRASTIC genes exist in the FC predicted heat-stress genes. However, Figure 4.(B) shows that only (171) TAIR10 genes exist in the FC predicted heat-stress genes. In the same way, Figure 4.(C) indicates that (41) DRASTIC genes exist in the edgeR predicted heat-stress genes. However, Figure 4.(D) shows that (152) TAIR10 genes exist in the edgeR predicted heat-stress genes. Lastly, Figure 4.(E) decides that only (6) genes are common between DRASTIC and TAIR10 DBs. On the contrary, in Figure 4.(F) most of (18,517) edgeR predicted genes are included into the (30,959) FC predicted genes. Only (6) genes from edgeR are not included into FC predicted genes.

Literature Based verification of 5% recommendations

The set of genes that were detected in Table 10 can be a potential addition to the body of knowledge in systems biology. Top 5% detected heat-stress genes shared between FC and edgeR based results are (26) genes in total. Exactly, (23) genes out of them do not exist in both DRASTIC and TAIR10 DBs. It is hypothesized that these (23) genes can be a newly discovered heat-stress genes. Here, a brief explanation as to why these can be response related genes is given, while due to space limitation of the article further verification is planned for a future study. After reviewing recent research published in the National Center for Biotechnology Information advances science and and (NCBI) as a trusted source for genes information, the following notes are collected about some of the selected genes (NCBI gene database, 2019).

Gene [AT5G42200] has neither reported phenotype nor known in vivo function.

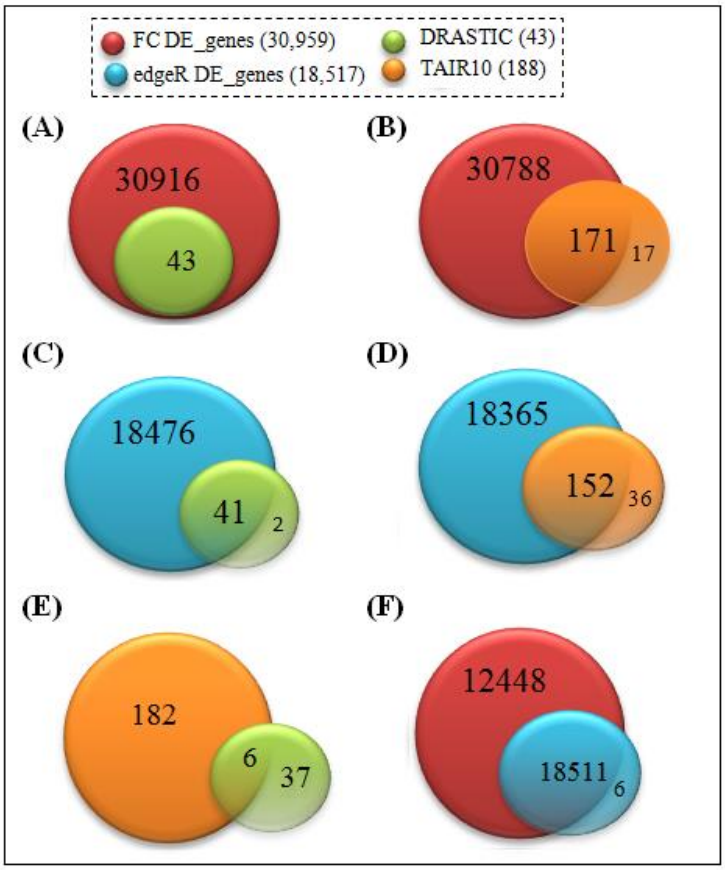

Figure 4: Venn diagram showing the common heat-stress genes between FC, edgeR, DRASTIC and TAIR10
[(A) FC vs. DRASTIC, (B) FC vs. TAIR10, (C) edgeR vs. DRASTIC, (D) edgeR vs. TAIR10, (E) DRASTIC vs. TAIR10, and (F) FC vs. edgeR].

Table 8: Top percentage intersected genes between DE FC genes and Reference Databases, each cell contains number of intersected genes and their gene Ids

DE FC_{total}	DRASTIC(43 genes)	TAIR10(188 genes)
Top 1% (309 genes)	1 (AT1G08830)	2 (AT5G23240, AT5G41920)
Top 5% (1547 genes)	5 (AT1G20620, AT3G52880, AT1G08830, AT5G25220, AT1G15100)	8 (AT5G23240, AT2G25140, AT5G41920, AT4G21320, AT3G53990, AT1G06460, AT4G11660, AT1G51670)
Top 10% (3095 genes)	7 (AT1G20620, AT5G02500, AT3G52880, AT3G45310, AT1G08830, AT5G25220, AT1G15100)	12 (AT1G79930, AT5G23240, AT5G02500, AT2G25140, AT5G41920, AT4G37910, AT1G12180, AT4G21320, AT3G53990, AT1G06460, AT4G11660, AT1G51670)
Top 20% (6191 genes)	16 (AT5G07090, AT3G18780, AT3G08580, AT5G13490, AT5G08670, AT1G20620, AT3G23990, AT5G02500, AT3G52880, AT4G02940, AT4G16190, AT3G45310, AT1G08830, AT5G18100, AT5G25220, AT1G15100)	27 (AT4G36990, AT2G41690, AT1G79930, AT3G17210, AT5G18340, AT4G39150, AT5G23240, AT2G22360, AT5G27240, AT5G17020, AT5G02500, AT2G25140, AT2G03020, AT5G18730, AT5G41920, AT4G37910, AT2G46240, AT1G12180, AT5G43840, AT4G21320, AT3G53990, AT1G06460, AT1G05850, AT4G11660, AT1G51670, AT3G23990, AT1G75220)

Table 9: Top percentage intersected genes between DE edgeR genes and Reference Databases, each cell contains number of intersected genes and their gene Ids

edgeR genes (F_{value})	DRASTIC (43 genes)	TAIR10 (188 genes)
Top 1% (185 genes)	0	2 (AT5G23590, AT5G23240)
Top 5% (925 genes)	2 (AT1G20620, AT4G16190)	12 (AT4G36990, AT2G32120, AT5G05750, AT5G23590, AT5G23240, AT2G20560, AT3G51910, AT3G53990, AT4G18880, AT3G56740, AT1G75220, AT1G67970)
Top 10% (1851 genes)	3 (AT5G47120, AT1G20620, AT4G16190)	24 (AT4G36990, AT3G63350, AT2G32120, AT5G05750, AT3G09350, AT5G48850, AT5G23590, AT5G23240, AT2G20560, AT5G27660, AT2G25140, AT3G08970, AT5G03720, AT3G57340, AT5G21160, AT3G51910, AT3G24500, AT3G53990, AT1G65280, AT4G18880, AT3G16770, AT3G56740, AT1G75220, AT1G67970)
Top 20% (3703 genes)	4 (AT5G47120, AT1G20620, AT4G16190, AT5G18100)	38 (AT4G36990, AT3G63350, AT2G32120, AT5G02490, AT5G05750, AT3G09350, AT5G48850, AT5G23590, AT1G18700, AT5G23240, AT2G15970, AT2G20560, AT5G27660, AT5G53150, AT1G79920, AT2G25140 , AT5G15450, AT4G19020, AT3G08970, AT5G03720, AT3G07770, AT3G57340, AT5G21160, AT2G46240, AT3G51910, AT5G09590, AT3G24500, AT3G53990, AT1G51670, AT5G53400, AT5G58410, AT1G65280, AT4G18880, AT5G16820, AT3G16770, AT3G56740, AT1G75220, AT1G67970)

Table 10: The genes intersected between equal slices of the total edgeR, and FC genes

Slice size	Count of edgeR genes	Count of FC genes	Intersected genes
Top 1%	185	309	1
Top 5%	925	1547	26
Top 10%	1851	3095	127
Top 20%	3703	6191	561
100 %	18517	30959	18511

However, genes [AT5G05440], and [AT5G13220] have functions in protein binding. On the other hand, genes [AT1G56220], [AT1G02640], [AT4G37990], [AT3G11590], [AT5G62200], [AT5G19140], [AT2G43540], [AT4G30780], [AT1G03070], [AT1G03220], [AT1G14970], [AT3G13784] reported to perform as protein coding in different purposes.

Moreover, gene [AT2G17840] is identified as drought-inducible gene and early-responsive to dehydration and through differential hybridization. It is described as up regulated by abiotic stress; high light, drought, cold and salt stress (Yin, M. et al., 2017). However, gene [AT2G47410] is declared as domain-containing protein which is a conserved part of a certain protein sequence and structure. It can function, and exist independently of the rest of the protein chain (Lee, J. H. et al., 2008). In addition, gene [AT2G33050] is found as a natural antisense transcript and has a gene encoding function as a receptor-like protein (Kondo, S. et al., 2016). Gene [AT1G67310] is considered Calmodulin-binding transcription activator protein (CAMTAs) that its functions are involved in developmental regulation and environmental stress response including abiotic and biotic stresses (Shen, C. et al., 2015). Also, gene [AT4G38580] is considered a farnesylated protein that can mediate protein-protein interactions and protein membrane interactions. As well, it is reported as heavy metal transport detoxification superfamily (Petzold, H. E. et al., 2017). Besides, gene [AT3G04910] is a serine/threonine protein kinase, whose transcription is regulated by circadian rhythm. This gene expression pattern of tissue specific and under various abiotic stresses reveals differential expression pattern (Manuka, R. et al., 2015). Lastly, gene [AT3G18800] is considered trans-membrane protein that functions as gateway to permit the transport of specific substances across the membrane (Sahoo, S. et al., 2019). Some of these genes can be ideal candidates to be tested in vivo to confirm their relations to abiotic stress phenomenon.

CONCLUSION

Understanding the impact of various kinds of biotic and abiotic stress continues to gain more attention in plant research community in order to grow better, stress tolerant plants. Information about genes expressed during the abiotic stress response will provide better understanding of the stress resistance phenomenon. The most common application of RNA-seq is to estimate gene and transcript expression. The presented research dedicated for studying the effect of heat-stress on *A. thaliana* plant via utilizing RNA-seq analysis. A proposed workflow for discovering the differentially expressed genes is presented and covers stages of mapping, quantification, and building of expression matrix. Two commonly used RNA-seq analysis methods were assessed; FC and edgeR. In order to evaluate their performance as DE analysis methods for RNA-seq data, the two methods were applied to ten RNA-seq *A. thaliana* samples under heat-stress. Next, DRASTIC and TAIR10 databases were used to provide a point of reference. FC predicted heat-stress genes contain all DRASTIC and (171) TAIR10 genes. Similarily, edgeR predicted heat-stress genes cover (41) DRASTIC and (152) TAIR10 genes. By comparing relative expression estimates, it was observed that results were more comprehensive and richer than provided by ad hoc methods. This evaluation showed improved and promising results for detected genes using FC and ensures the validity and applicability of the presented work. Additionally, workflow presented does not put any assumptions on *A. thaliana* plant nature so it can be applied for various organisms. Set of significant genes from the top 5% recommendation genes were collected to be verified to demonstrate the presented results in more useful manner. Furthermore, more analysis methods can be applied in the future to enrich this workflow and in vivo tests can be performed on the set of high recommended genes to maximize the benefit of the work results.

CONFLICT OF INTEREST

The authors declared that present study was performed in absence of any conflict of interest.

ACKNOWLEGEMENT

The authors are extremely grateful to all participants.

AUTHOR CONTRIBUTIONS

HZ and MN designed the study, followed up the cases and performed computational analysis and wrote the manuscript. AB contributed to the writing, amending and approving of manuscript. AF reviewed the manuscript. All authors read and approved the final version.

REFERENCES

BA-HPC group, Alexandria Library, https://www.bibalex.org/, Egypt, (accessed Oct 2018).

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics, 11(1), (pp. 94).

Chandramohan, R., Wu, P. Y., Phan, J. H., & Wang, M. D. (2013). Benchmarking RNA-Seq quantification tools. 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 647-650). IEEE.

Chen, Y., Lun, A. T., & Smyth, G. K. (2014). Differential expression analysis of complex RNA-seq experiments using edgeR. In Statistical analysis of next generation sequencing data, Somnath Data and Daniel S Nettleton (eds), (pp. 51-74). Springer, New York.

Chen, Y., Lun, A. T., & Smyth, G. K. (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research, 5, (pp. 1438).

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17(1), (pp. 13).

De la Fuente, A. (2010). From 'differential expression'to 'differential networking'–identification of dysfunctional regulatory networks in diseases. Trends in genetics, 26(7), (pp. 326-333).

Friedel, S., Usadel, B., Von Wirén, N., & Sreenivasulu, N. (2012). Reverse engineering: a key component of systems biology to unravel global abiotic stress cross-talk. Frontiers in plant science, 3, (pp. 294).

Gary Lyon, The DRASTIC gene expression database, http://www.drastic.org.uk, (accessed Nov 2018).

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols, 8(8), (pp. 1494).

Hu, Y. (2013). A Novel Computational Framework for Transcriptome Analysis with RNA-seq Data. D. Phil. Thesis, Kentucky University.

I-Hsuan. Lin. (2016). Hands-on Training in RNA-Seq Data Analysis. National Yang-Ming University, Taipei, Taiwan

Kondo, S., Ohto, C., Mitsukawa, N., & Ogawa, K. (2016). Gene capable of imparting environmental stress resistance to plants and method for utilizing the same. U.S. Patent No. 9,476,059. Washington, DC: U.S. Patent and Trademark Office.

Lee, J. H., Terzaghi, W., Gusmaroli, G., Charron, J. B. F., Yoon, H. J., Chen, H., & Deng, X. W. (2008). Characterization of Arabidopsis and rice DWD proteins and their roles as substrate receptors for CUL4-RING E3 ubiquitin ligases. The Plant Cell, 20(1), (pp. 152-167).

Less, H., Angelovici, R., Tzin, V., & Galili, G. (2011). Coordinated gene networks regulating Arabidopsis plant metabolism in response to various stresses and nutritional cues. The Plant Cell, 23(4), (pp. 1264-1271).

Ma, C., Xin, M., Feldmann, K. A., & Wang, X. (2014). Machine learning–based differential network analysis: A study of stress-responsive Transcriptomes in Arabidopsis. The Plant Cell, 26(2), (pp. 520-537), American Society of Plant Biologists.

Manuka, R., Saddhe, A. A., & Kumar, K. (2015). Genome-wide identification and expression analysis of WNK kinase gene family in rice. Computational biology and chemistry, 59, (pp. 56-66).

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research, 40(10), (pp. 4288-4297).

Moreno-Risueno, M. A., Busch, W., & Benfey, P. N. (2010). Omics meet networks - using systems approaches to infer regulatory networks in plants. Current opinion in plant biology, 13(2), (pp. 126-131).

Petzold, H. E., Rigoulot, S. B., Zhao, C., Chanda,

B., Sheng, X., Zhao, M., & Brunner, A. M. (2017). Identification of new protein–protein and protein–DNA interactions linked with wood formation in Populus trichocarpa. Tree physiology, 38(3), (pp. 362-377).

Quinn, T. P., Crowley, T. M., & Richardson, M. F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. BMC bioinformatics, 19(1), (pp. 274).

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), (pp. 139-140).

Sahoo, S., Das, S. S., & Rakshit, R. (2019). Codon usage pattern and predicted gene expression in Arabidopsis thaliana. Gene:X, Elsevier, 100012.

Shameer, K., Ambika, S., Varghese, S. M., Karaba, N., Udayakumar, M., & Sowdhamini, R. (2009). STIFDB - Arabidopsis stress responsive transcription factor dataBase. International journal of plant genomics.

Shen, C., Yang, Y., Du, L., & Wang, H. (2015). Calmodulin-binding transcription activators and perspectives for applications in biotechnology. Applied microbiology and biotechnology, Springer. 99(24), (pp. 10379-10385).

Sundar, A. S., Varghese, S. M., Shameer, K., Karaba, N., Udayakumar, M., & Sowdhamini, R. (2008). STIF: identification of stress-upregulated transcription factor binding sites in Arabidopsis thaliana. Bioinformation, 2(10), (pp. 431).

Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., & Radenbaugh, A. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic acids research, 36(suppl_1), (pp. D1009-D1014).

The Arabidopsis Information Resource (TAIR), http://www.Arabidopsis.org, (accessed Jan 2019).

Yin, M., Wang, Y., Zhang, L., Li, J., Quan, W., Yang, L., & Chan, Z. (2017). The Arabidopsis Cys2/His2 zinc finger transcription factor ZAT18 is a positive regulator of plant tolerance to drought stress. Journal of Experimental Botany, 68(11), (pp. 2991-3005).