INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
18F–137/2018/20–11–2590–2598
DOI: 10.17957/IJAB/15.0834
http://www.fspublishers.org

Full Length Article

Analysis of *Cf-12* Tomato Transcriptome Profile in Response to *Cladosporium fulvum* Infection with Hisat, StringTie and Ballgown

Wenbo Xu[1,2], Long Chen[1], Ping He[1], Jun Yang[1], Chengdong Xu[2], Bo Wang[2], Zhenji Wang[2], Haiyan Yang[2], Meihua Xie[2], Shenming Yang[2], Lu Qiu[2] and Yunyue Wang[1,3*]

[1]*College of Plant Protection, Yunnan Agricultural University. Kunming City, China*

[2]*School of Chemistry and Life Sciences, Chuxiong Normal University, Chuxiong City, China*

[3]*National Engineering Research Center of Agricultural Biodiversity Applied Technology, Yunnan Agricultural University. Kunming City, China*

[*]For correspondence: ux1240@163.com

Abstract

Nature Protocols introduced a new method of RNA-seq data processing with HISAT, StringTie and Ballgown, all of which are open-source software. HISAT aligns RNA reads to the genome; StringTie assembles the transcripts and computes their abundance, while Ballgown identifies differentially expressed genes and transcripts between samples. The software contained in these methods are flexible, accurate and efficient, making them convenient for researchers to use PC to analyze transcriptome data. The method provides a new way for RNA-seq from raw data reading to different expressions. The *Cf* gene is effective in resisting tomato leaf mold disease caused by *Cladosporium fulvum*. The *C. fulvum*-resistant tomato cultivars introduce *Cf* resistant genes identified from wild *Solanum* species. The *C. fulvum* pathogen has many physiological races, and newer physiological races have continued to evolve. Different from *Cf* genes such as *Cf-2*, *Cf-4*, *Cf-5*, and *Cf-9*, no *C. fulvum* physiological races virulent to *Cf-12* carrying tomato lines have been identified. In order to better understand the molecular mechanisms involved in the *Cf-12* gene resistance to *C. fulvum*, Hisat, StringTie and Ballgown were used to analyze the transcriptome changes at three different time stages of *C. fulvum* infection: 0 dpi (Cf12_A), 4 dpi (Cf12_B), and 8 dpi (Cf12_C). A total of 6446 differentially expressed genes (DEGs) between 4 and 0 dpi, and 7322 DEGs between 8 and 0 dpi were identified. Through GO and KEGG analysis, these DEGs were significantly highly enriched in plant-pathogen interaction pathway and plant disease resistance-related pathway, indicating that these DEGs play important roles in *C. fulvum* defense response. GO analysis for the up-regulated genes showed that 25 terms were significantly regulated. Biological process categories such as salicylic acid metabolic pathway, and biosynthetic pathways for jasmonic acid and salicylic acid were highly enriched. Through KEGG pathway enrich analysis of the up-regulation, many up-regulated and DEGs were enriched in plant-pathogen interaction pathways and disease resistance-related pathways. These results indicate the resistance of *Cf-12* tomato. We found 15 pathogenesis-related genes and 18 resistance protein genes in the DEGs. This study used novel methods and large data analysis theory to explore the mechanism of disease resistance of *Cf-12* tomato, and it provides new insight on the molecular mechanism of Cf resistance to *C. fulvum.* © 2018 Friends Science Publishers

Keywords: HISAT; StringTie; Ballgown; *Cf* resistant genes; Leaf mold disease; Transcriptome

Introduction

Tomato (*Lycopersicon esculentum* Mill.) is an important vegetable and fruit in people’s daily life. It is also an economically important crop. Leaf mold disease usually affects the cultivation and production of tomato. The disease easily breaks out in a humid environment, which explains the greater damage seen in tomato greenhouse cultivation (Hand, 1988). Leaf mold disease is caused by the biotrophic pathogen *Cladosporium fulvum* (syn. *Passalora fulva*). It is transmitted through the air, and it invades the stomata at abaxial surface of the leaf. This fungus primarily infects tomato leaves, but sometimes, the petioles, flowers and stems are also infected. The symptoms are pale yellow spots on leaves. The infection results in reductions in fruit yield and fruit quality, or even death of the entire plant. The pathogen of tomato leaf mold disease has many physiological races, and new physiological races have continued to evolve (Jones and Dangl, 2006; Chisholm *et al*., 2006).

When pathogens infect the tomato plant, the plant has two levels of innate immune system. The first level is pathogen-associated, molecular pattern-triggered immunity (PTI). It is the immune response activated by the pattern recognition receptors (PRRs) of plant cell surface to identify the pathogen associated molecular patterns (PAMPs) of the pathogenic microbes. Sometimes, when the pathogenic microbes infect the specific host plants, they can release

To cite this paper: Xu, W., L. Chen, P. He, J. Yang, C. Xu, B. Wang, Z. Wang, H. Yang, M. Xie, S. Yang, L. Qiu and Y. Wang, 2018. analysis of *Cf-12* tomato transcriptome profile in response to *Cladosporium fulvum* infection with Hisat, StringTie and Ballgown. *Int. J. Agric. Biol.*, 20: 2590–2598

some virulence factors secreted by the cilia to inhibit the PTI pathway. Consequently, the host plants have evolved a second surveillance level to resist the infection of pathogenic microbes. This level of immunity is called effector triggered immunity (ETI). Through the resistance proteins (R-proteins), ETI directly or indirectly identifies the effectors produced by the pathogens. When the R-proteins are activated, they usually cause the death of cells around the pathogenic microbe infection site. This phenomenon is called hypersensitive response (HR). The HR reaction of plants can effectively prevent pathogenic microbe growth in the HR reaction region, and the HR reaction usually does not exceed the area of pathogenic microbe infection. The ETI pathway-induced HR reaction is more intense than the PTI pathway, but not all ETI pathways can induce HR reaction, and plant disease resistance does not all exhibit HR reactions (Rivas and Thomas, 2005).

In the current perspective, natural plant immunity system is divided into the following four stages: the first stage is the PTI pathway induced by PAMPs; the second stage is the pathogenic microbe secretion of effector to inhibit the PTI pathway that makes the plant sensitive to the pathogenic bacteria; the third stage is the NB-LRR protein specific recognition effector so as to trigger ETI immune response; and the fourth stage is the pathogen-induced production of another mechanism to inhibit ETI under the natural selection pressure. The R-genes also have a new corresponding resistance that enables ETI to be triggered repeatedly (Bolton *et al*., 2007). *C. fulvum*-resistant tomato cultivars introduce *Cf* resistant genes identified from wild *Solanum* species. Since the discovery of first *Cf* gene (*Cf-1*) in the 1930s, more than 20 new *Cf* genes have been discovered and introduced into cultivated tomatoes. The *Cf-12*-carrying tomato lines is efficiently resistant to *C. fulvum*. Unlike other *Cf* gene-carrying lines, none of the *C. fulvum* physiological races is pathogenic to the *Cf-12* carrying lines. The molecular mechanisms triggered in *Cf-12* tomato in response to infection with *C. fulvum* are poorly understood. The use of big data theory to analyze the transcriptional data of *cf-12* tomato infected with *C. fulvum* has also not been reported before now.

High-throughput sequencing of mRNA (RNA-seq) technology is developing rapidly. It has been widely used for measuring and comparing gene expression levels (Rivas and Thomas, 2005). The RNA-seq experiment is usually comprehensive, and it produces large amount of data. Thus, a flexible and accurate software is needed to process the transcriptome data. At present, most of the data analysis and processing of transcriptome are carried out by companies, leading to high economic costs. In addition, the accuracy of data analysis and processing cannot be guaranteed, coupled with the limitations of data processing initiative. Recently, an article in *Nature Protocols* introduced the new method of RNA-seq data processing involving the use of HISAT, StringTie and Ballgown, all of which are open-source software. HISAT aligns RNA reads to the genome (Kim *et al*., 2015). StringTie assembles the transcripts and computes their abundance (Pertea *et al*., 2015), while Ballgown identifies differentially expressed genes and transcripts between samples (Frazee *et al*., 2014). The software contained in the methods are flexible, accurate and efficient which make it convenient for researchers to use PC to analyze transcriptome data. These methods provide a new way for RNA-seq from raw data reading to different expressions.

In the present study, we used HISAT StringTie and Ballgown to process the transcriptome data of *Cf-12* tomato infected with *C. fulvum*. Using the big data analysis theory at the transcriptome level, we studied the immune response process of the *Cf-12* tomato interaction with *C. fulvum* and analyzed the resistance metabolic pathways and molecular regulation network of *Cf-12* tomato. This was aimed at unravelling the disease-resistant mechanisms and identifying the disease-resistant genes involved.

Materials and Methods

Data Acquisition

In order to identify the DEGs, we selected three time-course RNA-seq data of *cf-12* tomato leaves infected by *C. fulvum* (0, 4 and 8 dpi). Every time stage sample had three replicates. The RNA-seq data was downloaded from the NCBI SRA database, with sequence numbers SRR4041970, SRR4041973, SRR4041974, SRR4041975, SRR4042017, SRR4042029, SRR4042030, SRR4042031, and SRR4042332. The reference genome and annotation files (*Solanum_lycopersicum*.SL2.5. dna. toplevel.fa and *Solanum_lycopersicum*.SL2.50.38.chr.gff3) were downloaded from Ensembl Genomes Databases (url: http://ensembl.gramene.org/Zea_mays/Info/Index?db=core

Hardware and Software Computer (Model Name: Lenovo-B40, CPU: AMD E1-6010 APU with AMD Radeon R2 Graphics x 2, RAM: 4Gb, Hard Disk: 500Gb)
Linux OS (Version: ubuntu 16.04 LTS).

HISAT (Version: hisat2-2.1.0); StringTie (Version: stringtie-1.3.3b. Linux_x86_64); SAMtools (Version: samtools-1.5); Rstudio (Version: rstudio-xenial-1.1.383-amd64); R package Ballgown (for estimating differential expression transcripts and genes); alyssafrazee/ RskittleBrewer (for setting up colors); genefilter (for fast calculation of means and variances); dplyr (for sorting and arranging results); devtools (for reproducibility and installing packages); Convert the sra data file into fastq data file

```
$ fastq-dump –split-3 SRR4041970
$ fastq-dump –split-3 SRR4041973
$ fastq-dump –split-3 SRR4041974
$ fastq-dump –split-3 SRR4041975
$ fastq-dump –split-3 SRR4042017
$ fastq-dump –split-3 SRR4042029
$ fastq-dump –split-3 SRR4042030
$ fastq-dump –split-3 SRR4042031
$ fastq-dump –split-3 SRR4042032
```

Quality Control and Preprocessing

The raw data of RNA seq was stored in fastq format, in order to protect its quality. It was necessary to pre-process the raw data. Here we used fastp to clean the raw data. The fastp software filters the sequences with low quality and more N, and it can automatically locate the adapter sequence and cut the adapter pollution. Sequence data quality score above Q20 were used for further analysis. The command was as follows:

```
$ fastp -i SRR4041970_1.fastq –I SRR4041970_2.fastq -o 4041970_1.fastq -O 4041970_1.fastq
$ fastp -i SRR4041973_1.fastq –I SRR4041973_2.fastq -o 4041973_1.fastq -O 4041973_1.fastq
$ fastp -i SRR4041974_1.fastq –I SRR4041974_2.fastq -o 4041974_1.fastq -O 4041974_1.fastq
$ fastp -i SRR4041975_1.fastq –I SRR4041975_2.fastq -o 4041975_1.fastq -O 4041975_1.fastq
$ fastp -i SRR4042017_1.fastq –I SRR4042017_2.fastq -o 4042017_1.fastq -O 4041970_1.fastq
$ fastp -i SRR4042029_1.fastq –I SRR4042029_2.fastq -o 4042029_1.fastq -O 4042029_1.fastq
$ fastp -i SRR4042030_1.fastq –I SRR4042030_2.fastq -o 4042030_1.fastq -O 4042030_1.fastq
$ fastp -i SRR4042031_1.fastq –I SRR4042031_2.fastq -o 4042031_1.fastq -O 4042031_1.fastq
$ fastp -i SRR4042032_1.fastq –I SRR4042032_2.fastq -o 4042032_1.fastq -O 4042032_1.fastq
```

The in.fq is the raw data to be filtered and quality-controlled, and the out.fq is the clean data. When the software was executed, the fastp.html was generated. Fastp.html is a visual quality control report.

Use of HISAT to Map the RNA-seq Reads to the Reference

Aligning the reads to the genome is the first step of the RNA-seq analysis. HISAT is the fastest read mapping software currently available. It is based on the Burrows-Wheeler transform algorithm. HISAT builds two types index for the alignment: one is global Ferragina-Mantzini (FM) index, and the other is local FM index. The global index represents the whole genome, and the large number of local indexes represent the small regions of the genome with overlaps; they cover the whole genome. HISAT uses less memory than the other aligning software. Thus, the work can be transferred from the specific server to the ordinary personal computer (Kim *et al*., 2015). Before aligning, the user must extract the splice-site information and exon information form the genome annotation file:

```
$extract_splice_sites.pySolanum_lycopersicum.SL2.50.38.chr.gff3 > Solanum_lycopersicum.ss
$ extract_exons.py Solanum_lycopersicum.SL2.50.38.chr.gff3 > Solanum_lycopersicum.exon.
```

Then, index files were generated:

```
$ hisat2-build --ss Solanum_lycopersicum.ss --exon Solanum_lycopersicum.exon Solanum_lycopersicum._tran
```

Third, the reads were aligned to the genome reference:

```
$ hisat2 -p 8 –dta -x home/xwb/Downloads/Solanum_lycopersicum._tran -1 4041970_1.fastq -2 4041970_2.fastq –S 4041970.sam
$ hisat2 -p 8 –dta -x home/xwb//Download/Solanum_lycopersicum._tran -1 4041973_1.fastq -2 4041973_2.fastq –S 4041973.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4041974_1.fastq -2 4041974_2.fastq –S 4041974.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4041975_1.fastq -2 4041975_2.fastq –S 4041975.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4042017_1.fastq -2 4042017_2.fastq –S 4042017.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4042029_1.fastq -2 4042029_2.fastq –S 4042029.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4042030_1.fastq -2 4042030_2.fastq –S 4042030.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4042031_1.fastq -2 4042031_2.fastq –S 4042031.sam
$ hisat2 -p 8 –dta -x home/xwb/Download/Solanum_lycopersicum._tran -1 4042032_1.fastq -2 4042032_2.fastq –S 4042032.sam
```

Fourth, the SAM files were sorted and converted to BAM

```
$ samtools sort -@ 8 -o 4041970.bam 4041970.sam
$ samtools sort -@ 8 -o 4041973.bam 4041973.sam
$ samtools sort -@ 8 -o 4041974.bam 4041974.sam
$ samtools sort -@ 8 -o 4041975.bam 4041975.sam
$ samtools sort -@ 8 -o 4042017.bam 4042017.sam
$ samtools sort -@ 8 -o 4042029.bam 4042029.sam
$ samtools sort -@ 8 -o 4042030.bam 4042030.sam
$ samtools sort -@ 8 -o 4042031.bam 4042031.sam
$ samtools sort -@ 8 -o 4042032.bam 4042032.sam
```

Use of StringTie to Assemble and Quantitate Full-length Transcripts

StringTie is a fast and highly efficient assembler of read alignments into transcripts. Compared with other main stream transcript assemblers, StringTie can assemble the reads into transcripts more completely, and estimate the gene expression levels accurately at the same time. The assembly may be reference-based or non-reference based; the latter is less accurate than the former. In the present study, we used the reference-based assembly. In the first step, StringTie assembles transcripts from RNA-seq reads that had been aligned to the genome and clusters the reads. Then, it creates a splice graph for each cluster through which it identifies transcripts of different isoforms and reconstructs them. Using the maximum flow algorithm, it creates a separate flow network for each transcript to estimate the expression level beginning from the highest expressed transcript. Then, it removes all the reads belonging to that transcript and repeats the process. In order to make for consistency in the transcripts in different samples, StringTie provides the merge function which can improve the accuracy of assemblies. The output files generated by the StringTie can be directly used as the input files of Ballgown for downstream analysis. The file contains exon expression levels, intron expression levels, transcript expression levels, the corresponding relationship between exon and transcripts, and the corresponding relationship between intron and transcripts (Pertea *et al*., 2015). The Stringtie procedure is as follows:

First, it assembles transcripts for each sample:

```
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa -o 4041970.gtf -l 4041970 4041970.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa -o 4041973.gtf -l 4041973 4041973.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa -o 4041974.gtf -l 4041974 4041974.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa-o 4041975.gtf -l 4041975 4041975.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa-o 4042017.gtf -l 4042017 4042017.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa-o 4042029.gtf -l 4042029 4042029.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa-o 4042030.gtf -l 4042030 4042030.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa-o 4042031.gtf -l 4042031 4042031.bam
$ Stringtie -p 8 -G home/xwb/Downloads/ Solanum_lycopersicum.SL2.5.dna.toplevel.fa-o 4042032.gtf -l 4042032 4042032.bam
```

Secondly, it merges transcripts from all samples:

```
$ stringtie –merge -p 8 –G xwb/home/Downloads/Solanum_lycopersicum.SL2.50.38.chr.gff3 –o stringtie_merge.gtf xwb/home/Downloads/ mergelist.txt
```

Thirdly, it examines how the transcripts compare with the reference annotation

```
$ gffcompare –r xwb/home/Downloads/Solanum_lycopersicum.SL2.50.38.chr.gff3 –G –o merged stringtie_merged.gtf
```

Thirdly, the transcript abundances were estimated and table counts for Ballgown were created.

```
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4041970/4041970.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4041973/4041973.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4041974/4041974.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4041975/4041975.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4042017/4042017.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4042029/4042029.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4042030/4042030.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4042031/4042031.gtf
$ stringtie –e –B 8 –G stringtie_merged.gtf –o ballgown/4042032/4042032.gtf
```

Use of Ballgown to Estimate Differential Expression Transcripts and Genes

Ballgown is an R package designed for differential expression analysis of RNA-Seq data. It functions to organize, visualize, and analyze the expression measurements for transcriptome assembly. The differential expression analysis is based on the flexible linear model framework. In order to improve the accuracy of the FPKM value and the stability of the variance, Ballgown performs a log transformation to the FPKM, and simultaneously filters out the confounders which may affect the results. The results include fold value between time courses, *p* value and *q* value of the differential expression (Frazee *et al*., 2014, 2015). The work flow is as follows:

1. Loading of R packages

```
$ R
> library (ballgown)
> library (RSkittleBrewer)
> library (genefilter)
> library (dplyr)
> library (devtools)
```

2. Load of the phenotype data.

```
> pheno_data = read.csv(“geuvadis_phenodata.csv”)
```

3. Loading of the expression data calculated by Stringtie.

```
> bg_ Solanum_lycopersicum = (dataDir = “ballgown”, samplePattern = “404”, pData = pheno_data)
```

4. Filtering and removing low-abundance genes

```
>bg_Solanum_lycopersicum_filt=subset (bg_Solanum_lycopersicum, “rowVars (texpr (bg_Solanum_lycopersicum))>1”, genomesubset=TURE)
```

5. Identifying differential expression transcripts.

```
>results_transcripts = stattest (bg_Solanum_lycopersicum_ filt, feature = “transcript”, covariate= “ids”, getFC = TURE, meas = “FPKM”)
```

6. Identifying differential expression genes.

```
> results_genes = stattest (bg_Solanum_lycopersicum_filt, feature = “gene”, covariate = “ids”, getFC = TRUE, meas = “FPKM”)
```

7. Adding gene names and gene IDs to the results_ transcripts data frame.

```
> results_transcripts = data.frame (geneNames = ballgown::geneNames (bg_Solanum_lycopersicum_filt), results_transcripts)
```

8. Sorting the results from *p* value

```
> results_transcript = arrange (results_transcripts,pval)
> results_genes = arrange (results_genes,pval)
```

9. Writing the results to a csv file that can be shared and distributed

```
> write.csv (results_transcripts, "Solanum_lycopersicum_transcript_results.csv", row.names =FALSE)
> write.csv (results_genes, "Solanum_lycopersicum_gene_results.csv", row.names = FALSE)
```

10. Identifying transcripts and genes with *q* value <0.05

```
> subset (results_transcripts, results_transcripts $ q value <0.05)
> subset (results_genes, results_genes $ q value <0.05)
```

Function Analysis and KEGG Enrichment Analysis of DEGs

Go is a commonly used classification system of gene function which is based on biological pathway, molecular function and cellular components. The *p* value Go term and the FDR value of *p* value were calculated through statistical analysis of Go terms enrichment degree of differentially expressed genes. GO analysis, which is most likely to be related to differential genes, was helpful in the experimental results. Through Go analysis of the differential genes, we found the Go classification items that enrich the differentially expressed genes, identified genes which may be related to the functional changes of the genes in different samples (Camon *et al*., 2004).

We used agriGO to perform the Go annotation. The URL of agriGO is http://bioinfo.cau.edu.cn/agriGO/index.php. Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database for systematic analysis of gene function and genomic information which can be used for metabolism analysis and metabolic network research (Kanehisa and Goto, 2000). We used KOBAS3.0 to perform the KEGG enrichment analysis. The KOBAS3.0 URL is http://kobas.cbi.pku.edu.cn/. Before KEGG enrichment, we used g: Profiler to covert the tomato genes id to enterz gene id, so the KOBAS could identify the tomato gene list.

Results

Raw reads Pre-processing, Aligning and Assembly

RNA seq data of *cf-12* tomato leaves infected by *C. fulvum* (0, 4 and 8 dpi) was downloaded from the NCBI SRA database. After cutting the adapters, filtering the low quality and more N reads, the average clean read counts of Cf12_A, Cf12_B and Cf12_C were 57.936, 54.818 and 69.249489 M, respectively. The average Q20 base counts were 7.119, 6.741 and 8.525 G, respectively, while the average Q30 base counts were 6.954, 6.590 and 8.354 G, respectively; and the average GC contents were 42.502, 42.573 and 42.717%, respectively. The alignment result indicated that 91.08% of all the pair reads were aligned concordantly to the tomato reference genome exactly once or more than once. The mean overall alignment levels of Cf12_A, Cf12_B and Cf12_C were 95.70, 94.69 and 92.50%, respectively. StringTie assembled the transcripts and merged all the transcripts of the 9 samples. Gffcompare compared the merged transcripts to the reference annotation. The Gffcomopare results showed that 25656 transcripts matched the intron chains, 35145 transcripts matched the reference transcripts, while 34741 transcripts matched the loci. The levels of novel exons, missed introns and novel loci were 6.8, 0.1 and 6.1%, respectively. After assembly, there were 21262 transcripts and 14676 unigenes in each sample. Table 1 shows relative gene distribution measured as FPKM value across samples the results are showed in Fig. 1.

Gene Differential Expression Analysis

After screening, 6446 differentially expressed genes were identified between Cf12_A and Cf12_B, including 3848 up-regulated genes, which accounted for 59.70% of the total DEGs, and 2598 down-regulated genes, which accounted for 40.30% of the total number of DEGs. Differentially expressed genes (7322) were identified between Cf12_A and Cf12_C, including 3816 up-regulated genes, which accounted for 59.70% of the total DEGs, and 3516 down-regulated genes, which formed 40.30% of the total number of DEGs. These results are shown in Fig. 2. In the up-regulated DEGs, we found 15 pathogenesis-related genes and 18 resistance protein genes which are shown in Table 2 and Table 3

GO Analysis of Differentially Expressed Genes

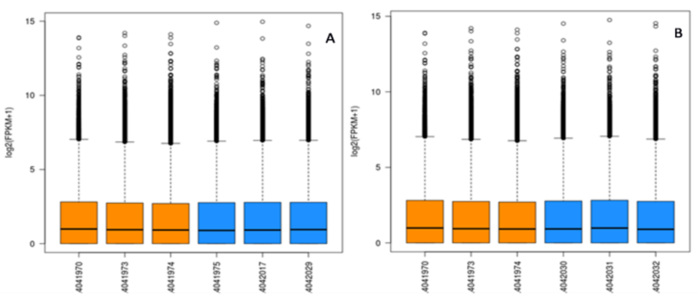
The differential expression genes (DEGs) between the treatment group (Cf12_A) and the control groups (Cf12_B and Cf12_C) was analyzed by Go function analysis. The results are shown in Fig. 3. Biological process categories such as salicylic acid metabolism, jasmonic acid and salicylic acid biosynthesis; phosphorus metabolism, protein phosphorylation, and cellular protein modification relative to disease resistance were significantly enriched. Cellular component category such as thylakoid, plastid thylakoid membranes, and photosynthetic membranes were also significantly enriched. Moreover, the catalytic activities of transferases, phosphotransferases and protein serine/threonine kinases were enriched, as well as anion binding, purine nucleoside binding, nucleoside binding, adenyl nucleotide binding, and ATP binding. These catalytic activity terms and binding terms were closely related to signal recognition and signal transduction.

KEGG Pathway Enrichment Analysis of Differentially Expressed Genes

The Unigene was compared with the KEGG database to identify the biological pathways of incompatible interactions. Statistical results showed that 1705 of the different expressed upregulated unigenes between *Cf*12_A

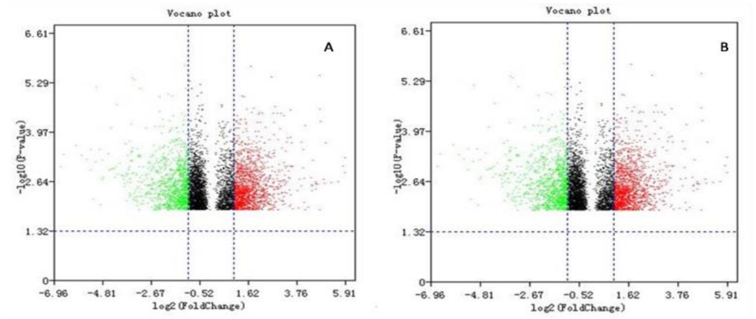
Table 1: Quality control of the RNA seq data of *cf-12* tomato leaves infected by *C. fulvum*

Summary statistics	Cf12_A1	Cf12_A2	Cf12_A3	Cf12_B1	Cf12_B2	Cf12_B3	Cf12_C1	Cf12_C2	Cf12_C3
Total reads (M)	61.26898	62.723930	53.88178	57.19780	63.06396	48.13255	58.28756	77.17739	77.00365
Total bases (G)	7.658623	7.840491G	6.735223	7.149725	7.882996	6.016570	7.285946	9.647175	9.625457
Average length (bp)	120	120	120	120bp	120bp	120bp	120bp	120	120
Reads with low quality	1.543610	1.421016	1.032810	1.299944	1.565468	1.009722	1.308816	1.689496	1.641354
Reads with too many N (K)	22.27200	23.87400	21.30000	21.94600	23.74600	18.98800	21.79400	29.93400	28.75800
Clean reads (M)	59.70310	61.27904	52.82767	55.87591	61.47475	47.10384	56.95695	75.45796	75.33354
Clean bases (G)	7.458317	7.653967	6.598022	6.979850	7.679373	5.883671	7.113867	9.424488	9.408955
Q20 bases (G)	7.314011	7.536348	6.506437	6.871191	7.554701	5.796675	7.009752	9.290506	9.273653
Q30 bases (G)	7.115630	7.371558	6.374020	6.717678	7.381588	5.673115	6.867619	9.106513	9.086950
GC content (%)	42.45268	42.65321	42.40128	42.63051	42.47985	42.60992	42.87908	42.58219	42.68961



F

Fig. 1: The distribution of gene abundance. A, The distribution of gene abundances Cf12_A and Cf12_B, The yellow color represents Cf12_A, the blue color represents Cf12_B; B, The distribution of gene abundances of Cf12_A and Cf12_C, The yellow color represents Cf12_A, the blue color represents Cf12_C



F

Fig. 2: Cf12_B *vs.* Cf12_A. A, Differential Expressed Genes (6446), up regulated: 3848, down regulated:2598; B, Differential Expressed Genes (7332), up regulated:3816, down regulated:3516

and *Cf*12_C were annotated in 113 KEGG Pathway, 16% participate in metabolic pathways, 10.79% in are involved in pathways of biosynthesis of secondary metabolites, while 2.46% are involved in plant-pathogen interaction pathways as shown in Fig. 4.

As shown in Fig. 5, the levels of different expressed genes enriched in plant disease resistance-related pathways were also significantly higher than their levels in the other pathways such as metabolism of amino acids (phenylalanine, tyrosine and tryptophan); carbohydrate metabolism (glycolysis/gluconeogenesis, pyruvate); and biosynthesis of secondary metabolites (phenylalanine metabolism, phenylpropanoid biosynthesis). In addition, it was found that plant hormone signal transduction, biosynthesis of unsaturated fatty acids, and fatty acid metabolism were involved in *Cf-12* tomato response to *C. fulvum* infection.

Discussion

In this study, the Hisat+ StringTie +Ballgown transcriptome analytical methods recently published in *Nature Protocol*》was used to analyze the three sets of

Table 2: Pathogenesis-related gene

Gene ID	Gene function
Solyc12g056590.1	Ethylene responsive transcription factor 2a
Solyc08g080660.1	Osmotin-like protein (Fragment)
Solyc06g068570.2	AP2-like ethylene-responsive transcription factor At1g16060
Solyc08g080640.1	Osmotin-like protein (Fragment)
Solyc05g051200.1	Ethylene-responsive transcription factor 1A
Solyc12g056980.1	Ethylene responsive transcription factor 2b
Solyc08g078180.1	Ethylene-responsive transcription factor 1A
Solyc09g089910.1	Ethylene responsive transcription factor 1a
Solyc04g081550.2	Thaumatin-like protein
Solyc01g065980.2	Ethylene responsive transcription factor 2b
Solyc02g064960.2	AP2-like ethylene-responsive transcription factor At1g16060
Solyc09g066360.1	Ethylene-responsive transcription factor 2
Solyc03g123500.2	Ethylene responsive transcription factor 2a
Solyc04g054910.2	Ethylene-responsive transcription factor 13
Solyc05g009450.1	Ethylene responsive transcription factor 2a

Table 3: Resistance protein gene

Gene ID	Gene function
Solyc05g007640.2	Cc-nbs-lrr, resistance protein with an R1 specific domain
Solyc05g008070.2	Cc-nbs-lrr, resistance protein
Solyc01g014840.2	Tir-nbs-lrr, resistance protein
Solyc12g094660.1	Cc-nbs-lrr, resistance protein
Solyc07g049700.1	Cc-nbs-lrr, resistance protein
Solyc04g007320.1	Tir-nbs-lrr, resistance protein
Solyc01g008800.1	Tir-nbs-lrr, resistance protein
Solyc12g096920.1	Cc-nbs-lrr, resistance protein
Solyc04g007060.2	Cc-nbs-lrr, resistance protein
Solyc01g102880.1	Tir-nbs-lrr, resistance protein
Solyc10g047320.1	Cc-nbs-lrr, resistance protein
Solyc11g020100.1	Cc-nbs-lrr, resistance protein
Solyc07g056200.2	NBS-LRR class disease resistance protein
Solyc04g079420.2	Nbs-lrr, resistance protein
Solyc04g007490.2	Cc-nbs-lrr, resistance protein with an R1 specific domain
Solyc03g005660.2	Cc-nbs-lrr, resistance protein
Solyc04g005550.1	Cc-nbs-lrr, resistance protein
Solyc09g098100.2	Cc-nbs-lrr, resistance protein

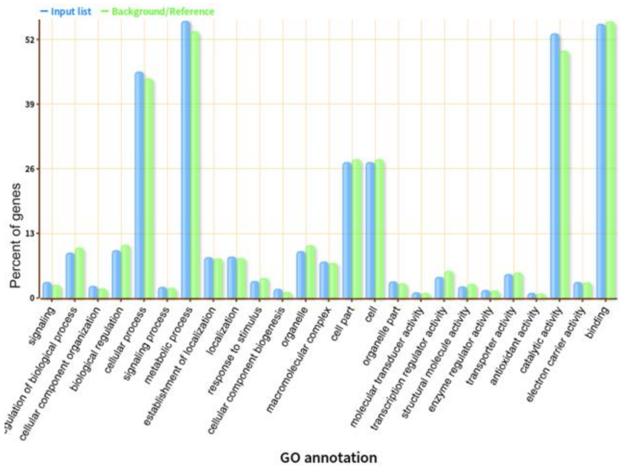


Fig. 3: Annotation of up regulated DEGs in GO

transcriptional data from *C. fulvum*-infected *Cf-12* tomato, *Cf12*_A (0dpi), *Cf12*_B (4dpi), and *Cf12*_C (8dpi). After assembling and quality control, an average of 21262 transcripts and 14676 unigenes were obtained in each sample. The *Cf*12_A *vs*. *Cf*12_B term had 6446 differentially expressed genes, 3848 of which were up-regulated genes, while 2598 were down-regulated genes. The *Cf*12_A *vs*. *Cf*12_C term had 7322 DEGs, 3816 of which were up-regulated genes, while 3516 were down-regulated genes. Enriched functional GO analysis for the up-regulated genes showed that 25 terms were significantly regulated. Biological process categories such as salicylic acid metabolic process, and biosynthesis of jasmonic acid and salicylic acid were highly enriched. Through KEGG pathway analysis of up-regulation, many up-regulated DEGs were enriched in plant-pathogen interaction pathway and disease resistance-related pathways (Ellis and Turner, 2001; Walters *et al*., 2002). which are evidence of the resistance of *Cf-12* tomato. The chitin elicitor receptor kinase 1 (*CERK1*) and Solyc07g049180.2, a pattern recognition protein, were expressed abundantly after tomato infection. These may play roles as pattern recognition receptors involved in the first layer of defense that recognizes *C. fulvum* (Cai *et al*., 2014). After the infection, *Cf-12* tomato established complex signal defense pathways. CDPK (Solyc03g113390.2, Solyc10g074570.1, Solyc02g083850.2, and Solyc10g076900.1) and MEKK1 (Solyc01g104530.2, and Solyc07g053170.2) were expressed highly after the infection, and subsequently stimulated the respiratory burst oxidase homolog (Rboh, Solyc01g099620.2, and Solyc03g117980.2). These findings are consistent with those obtained in previous studies (Malinovsky *et al*., 2014), and suggest that these genes play critical roles in *Cf-12* tomato response to *C. fulvum* infection.

A large number of studies have shown that signal molecules such as jasmonic acid bring about series of signal transductions of disease resistance and activate the expression of pathogenesis-related gene PR when the plant is infected by a biotrophic pathogen. Jasmonic acid-mediated disease resistance is the basis of disease resistance (Loake and Grant, 2007). In this study, the jasmonic-acid gene (JAZ, Solyc12g009220.1) which encodes a major protein in the jasmonic acid signaling pathway was upregulated following *C. fulvum* infection, suggesting that it may play a role in the tomato resistance against *C. fulvum*.

In the tomato infected by *C. fulvum* transcriptome, there was significant upregulation of DEGs enriched in phenylalanine and tryptophan biosynthesis pathways. This suggests that these pathways may play important roles in the defense response of *Cf-12* tomato against *C. fulvum*. Amino acids also have critical roles for plant growth, development, reproduction, defense, and environmental responses (Maeda and Dudareva, 2012). Tryptophan is a precursor of alkaloids, phytoalexins, and indole glucosinolates, whereas phenylalanine is a common precursor of numerous phenolic compounds such as flavonoids, condensed tannins, lignans, lignin, and phenylpropanoid/benzenoid volatiles (Maeda and Dudareva, 2012; Vogt, 2010). In *Arabidopsis* mutants,

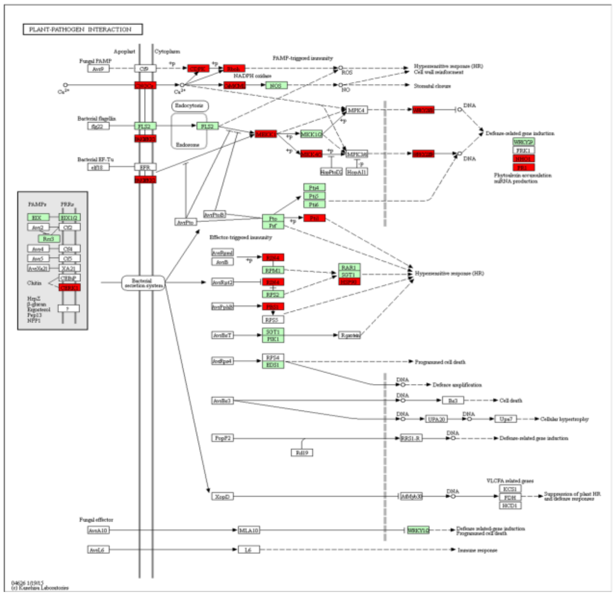


Fig. 4: Plant-pathogen interaction - *Solanum lycopersicum* (tomato)

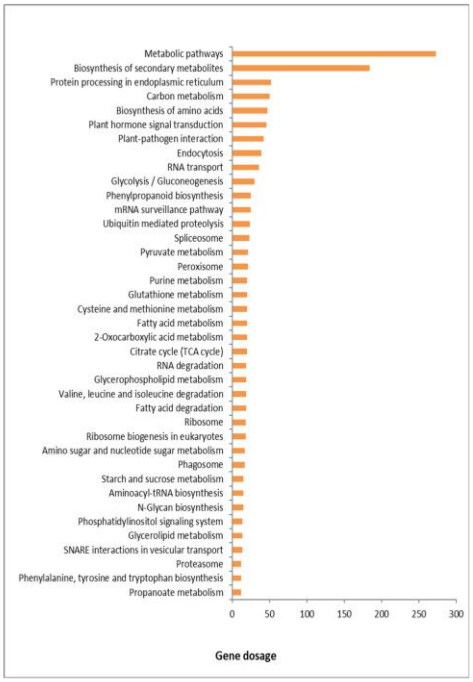


Fig. 5: KEGG Pathway analysis

glutathione and tryptophan metabolisms are required for immunity during the hypersensitive response to *Colletotrichum* genus of fungi (Hiruma *et al*., 2013).

Several pathways of carbohydrate metabolism such as glycolysis and gluconeogenesis, and pyruvate metabolism were upregulated in *Cf-12* tomato infected by *C. fulvum*, suggesting their possible role in the defense response. Presently, little is known about the role of carbohydrate metabolic pathways in the innate immunity of plants (Rojas *et al*., 2014).

Nonetheless, carbohydrate metabolism is not only critical for growth and development of the plant, but evidence suggest its involvement in the induction of a large number of defense responses to prevent or even avoid the proliferation of a potential pathogen (Berger *et al*., 2007; Rojas *et al*., 2014).

Secondary metabolites of plants form a group of diverse organic molecules that often promote growth and development of the plant. In many cases they are capable of inducing the synthesis of defense molecules (Patra *et al*., 2013). In this study, pathways of secondary metabolites such as phenylalanine metabolism and phenylpropanoid biosynthesis were significantly upregulated. This indicates that they may be involved in the synthesis of plant defense molecules in *Cf-12* tomato infected with *C. fulvum.* Furthermore, phenylalanine and phenylpropanoids are common precursors of numerous phenolic compounds, and have vital role in the resistance against pathogens (Dixon *et al*., 2002; Naoumkina *et al*., 2010). Flavonoids are important derivatives of phenylpropanoids, and they are important in plant responses to both biotic and abiotic stresses (Petrussa *et al*., 2013; Nakabayashi *et al*., 2014).

In the present study, the plant hormone signal transduction pathway was found to be enriched. Plant hormones are involved in plant disease resistance. The involvement of salicylic acid, jasmonic acid and ethylene in plant immune response has been reported (Ellis and Turner, 2001; Adie *et al*., 2007). In addition, the roles of auxin, cytokinin, abscisic acid and rapesinolide in plant disease resistance have been reported. Many hormones are involved in plant immune response, and the timing, species and quantity of hormones depend on the type of plant and the pathogen involved. In order to effectively resist different diseases, it is necessary for plants to regulate the complex network of hormonal signal transduction pathways. The genes enriched in the plant hormone signal transduction pathway may play important roles in the plant resistance (Santner and Estelle, 2009).

RNA-seq experiments generate very large, complex data sets that demand fast, accurate and flexible software to reduce the raw read data to comprehensible results. Recently we have developed three new software tools that accomplish the same tasks while running much faster, using substantially less memory, and providing more accurate overall results. Previously, most of the transcriptional data processing in China were handed over to companies. The present study used public databases to download transcriptome data, and used open source transcriptome group data analysis software. The bioinformatics analysis of tomato transcriptome data infected with *C. fulvum* by PC provides a reference for enhancing the ability of independent analysis and processing of transcriptional group data. At the same time, it provides a way to make better use of public database in carrying out scientific research on biological information resources in the future.

Conclusion

The development of biotechnology has greatly accelerated the advent of the bioinformation age. Bioinformatics data Network Center enables big data to be shared globally. Researchers can freely download the corresponding information data from the Internet according to their needs and mine the data through open source bioinformatics analysis software. Discover new scientific viewpoints and solve new scientific problems. In this study, three open source softwares named HISAT, StringTie and Ballgown proposed on Nature Protocols were used to process *Cf*-12 tomato transcriptional profile downloaded from public database, and to analyze the *Cf*-12 tomato disease resistance. mechanism.

References

Adie, B., J.M. Chico, I. Rubio-Somoza and R. Solano, 2007. Modulation of plant defenses by ethylene. *J. Plant Growth Regul*., 26: 160–177

Berger, S., A.K. Sinha and T. Roitsch, 2007. Plant physiology meets phytopathology: plant primary metabolism and plant-pathogen interactions. *J. Exp. Bot*., 58: 4019–4026

Bolton, M.D., J.H. Vossen, I.J.E. Stulemeijer, V. Den, B.P.H.J. Thomma, H.L. Dekker and C.D. Koster, 2007. Proteomic analysis of the apoplast of *Cladosporium fulvum*-infected tomato reveals novel virulence factors. *In: Book of Abstracts XIII International Congress Molicular Plant-Microbe Interaction,* p: 271. *Sorrento Italy*

Cai, G., G. Wang, L. Wang, J. Pan, Y. Liu and D. Li, 2014. *ZmMKK1*, a novel group A mitogen-activated protein kinase kinase gene in maize, conferred chilling stress tolerance and was involved in pathogen defense in transgenic tobacco. *Plant Sci*., 214: 57–73

Camon, E., M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez and R. Apweiler, 2004. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucl. Acids Res*., 32: 262–266

Chisholm, S.T., G. Coaker, B. Day and B.J. Staskawicz, 2006. Host-Microbe Interactions: Shaping the Evolution of the Plant Immune Response. *Cell,* 124: 803–814

Dixon, R.A., L. Achnine, P. Kota, C.J. Liu, M.S.S. Reddy and L. Wang, 2002. The phenylpropanoid pathway and plant defence—A genomics perspective. *Mol. Plant Pathol*., 3: 371–390

Ellis, C. and J.G. Turner, 2001. The *Arabidopsis* mutant cev1 has constitutively active jasmonate and ethylene signal pathways and enhanced resistance to pathogens. *Plant Cell*, 13: 1025–1033

Frazee, A.C., G. Pertea, A.E. Jaffe, B. Langmead, S.L. Salzberg and J.T. Leek, 2015. Ballgown bridges the gap between transcriptome assembly and expression analysis. *Nat. Biotechnol*., 33: 243-246

Frazee, A.C., G. Pertea, A.E. Jaffe, B. Langmead and S.L. Salzberg, 2014. Flexible isoform-level differential expression analysis with Ballgown. *BioRxiv*, 22: 1-13

Hand, D.W., 1988. Effects of atmospheric humidity on greenhouse crops. *Acta Hortic*., 229: 143–158

Hiruma, K., S. Fukunaga, P. Bednarek, M. Pislewska-Bednarek, S. Watanabe and Y. Narusaka, 2013. Glutathione and tryptophan metabolism are required for *Arabidopsis* immunity during the hypersensitive response to hemibiotrophs. *Proc. Natl. Acad. Sci. USA*, 110: 9589–9594

Jones, J.D. and J.L. Dangl, 2006. The plant immune system. *Nature*, 444: 323–329

Kanehisa, M. and S. Goto, 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucl. Acids Res*., 28: 27–30

Kim, D., B. Langmead and S.L. Salzberg, 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Meth.,* 12: 357–360

Loake, G. and M. Grant, 2007. Salicylic acid in plant defence-the players and protagonists. *Curr. Opin. Plant Biol*., 10: 466–472

Maeda, H. and N. Dudareva, 2012. The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. *Annu. Rev. Plant Biol*., 63: 73–105

Malinovsky, F.G., J.U. Fangel and W.G.T. Willats, 2014. The role of the cell wall in plant immunity. *Front. Plant Sci*., 5: 178

Nakabayashi, R., K. Yonekura-Sakakibara, K. Urano, M. Suzuki, Y. Yamada and T. Nishizawa, 2014. Enhancement of oxidative and drought tolerance in *Arabidopsis* by overaccumulation of antioxidant flavonoids. *Plant J*., 77: 367–379

Naoumkina, M., Q. Zhao, L. Gallego-Giraldo, X. Dai, P.X. Zhao and R. Dixon, 2010. Genome-wide analysis of phenylpropanoid defence pathways. *Mol. Plant Pathol*., 11: 829–846

Patra, B., C. Schluttenhofer, Y. Wu, S. Pattanaik and L. Yuan, 2013. Transcriptional regulation of secondary metabolite biosynthesis in plants. *Biochim. Biophys. Acta*, 1829: 1236–1247

Pertea, M., G.M. Pertea, C.M. Antonescu, T.C. Chang, J.T. Mendell and S.L. Salzberg, 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol*., 33: 290–295

Petrussa, E., E. Braidot, M. Zancani, C. Peresson, A. Bertolini and S. Patui, 2013. Plant flavonoids-biosynthesis, transport and involvement in stress responses. *Int. J. Mol. Sci*., 14: 14950–14973

Rivas, S. and C.M. Thomas, 2005. Molecular interactions between tomato and the leaf mold pathogen *Cladosporium fulvum*. *Annu. Rev. Phytopathol*., 43: 395–436

Rojas, C.M., M. Senthil-Kumar, V. Tzin and K.S. Mysore, 2014. Regulation of primary plant metabolism during plant-pathogen interactions and its contribution to plant defense. *Front. Plant Sci*., 5: 17

Santner, A. and M. Estelle, 2009. Recent advances and emerging trends in plant hormone signalling. *Nature*, 459: 1071–1078

Vogt, T., 2010. Phenylpropanoid biosynthesis. *Mol. Plant*, 3: 2–20

Walters, D., T. Cowley and A. Mitchell, 2002. Methyl jasmonate alters polyamine metabolism and induces systemic protection against powdery mildew infection in barley seedlings. *J. Exp. Bot*., 53: 747–756

(Received 03 July 2018; Accepted 09 July 2018)