OTÁVIO JOSÉ BERNARDES BRUSTOLINI

# Differential Gene Expression (DGE) by RNA sequencing analysis and development of software for integrating different DGE methods

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Agrícola, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2014

**Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa**

T

B912d
2014

Brustolini, Otávio José Bernardes, 1978-
Differential Gene Expression (DGE) by RNA sequencing analysis and development of software for integrating different DGEmethods. / Otávio José Bernardes Brustolini. - Viçosa, MG, 2014.
viii, 74f. : il. ; 29 cm.

Orientador : Elizabeth Pacheco Batista Fontes.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.5.

1. Tomate - Doenças e pragas. 2. Begomovirus. 3. Mosca Branca. 4. Expressão gênica. 5. RNA-seq transcriptoma. I. Universidade Federal de Viçosa. Departamento de Bioquímica e Biologia Molecular. Programa de Pós-graduação em Bioquímica Agrícola. II. Título.

CDD 22. ed 572.8

OTÁVIO JOSÉ BERNARDES BRUSTOLINI

# DIFFERENTIAL GENE EXPRESSION (DGE) BY RNA SEQUENCING ANALYSIS AND DEVELOPMENT OF SOFTWARE FOR INTEGRATING DIFFERENT DGE METHODS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Bioquímica Agrícola, para obtenção do título de *Doctor Scientiae*.

APROVADA: 28 de fevereiro de 2014

_____ _____
Fabyano Fonseca e Silva Benilton de Sa Carvalho

_____ _____
Humberto Josué de Oliveira Ramos (Coorientador) Juliana Lopes Rangel Fietto

_____
Elizabeth Pacheco Batista Fontes
(Orientadora)

## Acknowledgments

This thesis would not have been possible without the active support of many people known to me and many unknown. I am indebted to my family, friends and colleagues, who continuously guided and mentored me throughout my academic life. I am particularly grateful to my doctor supervisor, Prof. Dr. Elizabeth Pacheco Batista Fontes of University of Viçosa, who supported my interest in bioinformatics, otherwise, I would have gone to the tedious wet lab. I am still grateful to my adviser Prof. Dr. Fabyano Fonseca e Silva for his guides in statistics and useful discussions. Indeed, he was always there whenever I need him. My sincere thanks to Dr. Karen Lipkow and Dr. Sven from Cambridge System Biology Centre at University of Cambridge, UK, for the great opportunity to receive me at their laboratory, allowing and supporting my first international academic experience.

## Contents

# Resumo

BRUSTOLINI, Otávio José Bernardes, D.Sc., Universidade Federal de Viçosa, fevereiro de 2014. **Expressão Diferencial Gênica (EDG) por sequenciamento de RNA e o desenvolvimento de um sistema que integra esses métodos**.Orientadora: Elizabeth Pacheco Batista Fontes. Co-orientadores: Humberto Josué de Oliveira Ramos e Luciano Gomes Fietto.

Os begomovirus são geminivirus transmitidos pela mosca branca, e causam severos sintomas em cultivares com um grande impacto econômico na agricultura de regiões tropicais e subtropicais. Com as recentes mudanças climáticas é esperado uma forte alteração na distribuição da mosca branca ao redor do globo, tornando-a uma das mais sérias ameaças à agricultura. No tomateiro este fato ainda é mais preocupante, pois há uma complexa população de espécies emergentes de begomovirus que infectam estas plantas. Neste presente trabalho, nós propomos o estudo de um novo mecanismo regulatório presente em células vegetais que responde a infecção viral. Por meio da mutação no receptor imune NIK, o qual é alvo da proteína viral NSP, nós promovemos a ativação de um mecanismo de resposta antiviral que confere uma tolerância eficaz a diferentes espécies de begomovirus. Os nossos resultados também melhoraram o entendimento sobre o mecanismo de defesa intermediada pelo receptor NIK. Por meio de uma comparação usando quatro técnicas de agrupamentos hierárquico com quatro diferentes normalizadores presente no pacote edgeR, os perfis transcricionais do mutante T474D, que é o receptor NIK na sua forma ativa, com os das plantas infectas, mostraram que as plantas T474D mimetizaram o perfil das plantas infectadas, pois estes dois grupos se agruparam com um alto grau de confiabilidade, enquanto as plantas NIK induzidas e selvagens se diferenciaram em um grupo a parte. Além disso, a eliminação dos genes diferencialmente expressos (DE) do mutante T474D em todos os dados brutos dos tratamentos reforçaram que a resposta do mutante T474D mimetiza a planta com a infecção viral, pois os genótipos das plantas infectadas se agruparam com as selvagens também com um alto grau de confiabilidade. Portanto, estes resultados indicam que a infecção viral induziu a resposta antiviral mediada por NIK. Também foi empregado quatro diferentes métodos de expressão diferencial e um método

de enriquecimento de grupos gênicos (GSEA) nos dados de RNA-seq. Estes revelaram que a expressão ectópica do mutante T474D causa uma massiva “down” regulação de genes relacionados a tradução e uma “up” regulação de genes associados ao sistema imune. A “down” regulação mediada por T474D dos genes relacionados a tradução foi associado a uma supressão global na produção de proteínas, diminuindo assim a tradução dos polissomos do mRNA viral, aumentando, então a tolerância a begomovirus. Coletivamente nossos dados indicam que a sinalização antiviral mediada por NIK promove a resposta de defesa pela (i) supressão global da tradução e (ii) “up” regulação dos genes relacionado ao sistema imune da planta. O grande volume de dados provenientes do sequenciamento de RNA (RNA-seq) por meio de técnicas que geram uma grande quantidade de dados tal como os vindos de tecnologias de sequenciadores de nova geração, já estão disponíveis para a maioria dos laboratórios de pesquisa, e portanto está rapidamente se tornando uma ferramenta chave nos experimentos de expressão gênica. Os dados de RNA-seq são trabalhados da seguinte forma: os fragmentos (pequenas sequencias geradas pela tecnologia atual de RNA-seq) são mapeadas (alinhadas) com sequencias de referencia que podem ser o genoma ou transcriptoma, então uma tabela de contagemcontendo o número de fragmentos por gene é gerada e posteriormente analisada com os métodos de expressão diferencial. Na verdade este protocolo pode ser um trabalho árduo para pessoas que não têm muita experiência no ambiente R. Para encontrar genes cuja a expressão estão estatisticamente diferentes entre os tratamentos, além de avaliar o significado biológico através da anotação, muitos “scripts” do R precisam ser criados e as análises serem rodadas de forma correta. A variedade de opções contidas nas metodologias para a expressão gênica diferencial do ambiente R/Bioconductor fazem a tarefa de analisar esses tipos de dados ainda mais complicada. Portanto, nós desenvolvemos uma plataforma que facilita esse tipo de análise, fazendo com que as interações entre o usuário e o ambiente seja rápida e amigável. Ainda este sistema permite a possibilidade de combinar diferentes p-valores usando técnicas inspiradas na meta análise. Os métodos de expressão diferencial disponíveis são: edgeR, DESeq2, baySeq e NBPSeq. De fato, por meio de poucos passos uma análise poderá ser completada. Um diretório contendo o projeto que são informações das análises e todos os arquivos gerados incluindo os scripts serão armazenados juntamente com um banco de

dados em SQLite contendo os genes diferencialmente expressos com seu valor de expressão e anotação. Este programa é livre e de código aberto, permitindo assim quaisquer contribuições. Está plataforma é completamente livre.

# Abstract

BRUSTOLINI, Otávio José Bernardes, D.Sc., Universidade Federal de Viçosa, February, 2014. **Differential Gene Expression (DGE) by RNA sequencing analysis and development of software for integrating different DGE methods**. Adviser: Elizabeth Pacheco Batista Fontes. Co-Advisers: Humberto Josué de Oliveira Ramos and Luciano Gomes Fietto.

Begomoviruses (whitefly-transmitted geminiviruses) cause severe diseases of high economic impact on a variety of agriculturally relevant crops in tropical and subtropical areas. Current climate changes are expected to alter more the whitefly distribution along the globe posing a major threat to agriculture worldwide. This is particularly true for the case of tomato plants which are inflicted by a complex population of emergent species of tomato-infecting begomoviruses. Here we uncovered a novel regulatory mechanism of plant cells to fight plant DNA virus infection. By mutating the immune receptor NIK, which is a target of the begomovirus protein NSP, we promoted the activation of an antiviral defense response in tomato, which was effective to confer tolerance to different species of begomoviruses. Our results also shed light on the mechanism underlying the NIK-mediated defense. A comparison of the gain-of-function mutant (T474D)-induced transcriptome with infected WT transcriptome using a combination of four different clustering methods and four different normalization factors provided by the edgeR package revealed that the T474D-induced transcriptome mimicked the infected transcriptome as they clustered together with high confidence and they differ from the normal NIK-induced expression profile. Furthermore, the elimination of the mock T474D DE genes from the raw of all treatments further indicates that the expression profile induced by the T474D mutant mimics greatly the response to the viral infection, as the mock- and infected-induced transcriptomes from each genotype clustered together with high significance. These results indicate that the viral infection was the trigger of the NIK-mediated antiviral response. Furthermore, we employed four different methods for DGE analysis and the enrichment GSEA method to statistically analyze the RNA-seq data, which revealed that ectopic expression of T474D causes a massive down-regulation of the translation-related genes and up-regulatiom of immune system-associated genes. The T474D-mediated down-regulation of translation-related genes was associated with suppression of global protein, decreased viral mRNA loading in

actively translating polaysomes and enhanced tolerance against begomoviruses. Collectively our data indicate that the NIK-mediated antiviral signaling promotes a defense response by (i) suppressing global translation and (ii) up-regulating immune defense-related genes. The high-volume of RNA sequencing data provided by many high-throughput techniques like the next generation RNA sequencing technology (RNA-seq) is now within reach of any research laboratory and is quickly becoming established as a key research tool in any global gene expression experiments. In a RNA-seq workflow, the reads (short sequence generated by RNA-seq technology) are mapped (aligned) to a reference sequences data sets (transcriptome or genome), a counting table (number of reads per gene) can be set up and, then, a further downstream analysis can be executed to recover the biological meaning of the experiment. Actually, this protocol can be an arduous work for a person who is not an R experienced user. To find out which genes are statistically different in the expression profile among treatments and evaluate the biological meaning though annotation, many R scripts must be created and properly run. The variety of options for differential gene expression (DEG) methodology available on the R/Bioconductor makes this task even more troublesome. Our platform was designed to fill these gaps and make these iterations faster and easier. Yet, it reaches further expectations by allowing the combination of the p-values generated by the DGE methods: edgeR, DESeq2, baySeq and NBPSeq. Inspired on the meta analysis we used a combined p-value calculated by the Fisher's method, weighted Z-test, truncated product method, binomial test or a simple intersection (average or median) for helping the decision of the statistically significative DE genes based on these multiple DEG methods. To accomplish this goal, the friendly interface interacts with a low level R scripts to perform an RNA-Seq analysis without using directly a bunch of scripts. Indeed, by a few steps the analysis will be completely performed. A directory with the project name will be used to save all the generated files and store a SQLite database containing the DE genes with their expression values and annotation. As the program has the goal to be completely free and open to contributions, all the programs methods were designed to be a transparent system to the end user. This platform is completely free.

# CHAPTER I

# Global Gene Expression in Tomato Plants Tolerant to Begomovirus Infection

## Abstract

Begomoviruses (whitefly-transmitted geminiviruses) cause severe diseases of high economic impact on a variety of agriculturally relevant crops in tropical and subtropical areas. Current climate changes are expected to alter more the whitefly distribution along the globe posing a major threat to agriculture worldwide. This is particularly true for the case of tomato plants which are inflicted by a complex population of emergent species of tomato-infecting begomoviruses. Here we uncovered a novel regulatory mechanism of plant cells to fight plant DNA virus infection. By mutating the immune receptor NIK, which is a target of the begomovirus protein NSP, we promoted the activation of an antiviral defense response in tomato, which was effective to confer tolerance to different species of begomoviruses. Our results also shed light on the mechanism underlying the NIK-mediated defense. A comparison of the gain-of-function mutant (T474D)-induced transcriptome with infected WT transcriptome using a combination of four different clustering methods and four different normalization factors provided by the edgeR package revealed that the T474D-induced transcriptome mimicked the infected transcriptome as they clustered together with high confidence and they differ from the normal NIK-induced expression profile. Furthermore, the elimination of the mock T474D DE genes from the raw of all treatments further indicates that the expression profile induced by the T474D mutant mimics greatly the response to the viral infection, as the mock- and infected-induced transcriptomes from each genotype clustered together with high significance. These results indicate that the viral infection was the trigger of the NIK-mediated antiviral response. Furthermore, we employed four different methods for DGE analysis and the enrichment GSEA method to statistically analyze the RNA-seq data, which revealed that ectopic expression of T474D causes a massive down-regulation of the translation-related genes and up-regulatiom of immune system-associated

genes. The T474D-mediated down-regulation of translation-related genes was associated with suppression of global protein, decreased viral mRNA loading in actively translating polaysomes and enhanced tolerance against begomoviruses. Collectively our data indicate that the NIK-mediated antiviral signaling promotes a defense response by (i) suppressing global translation and (ii) up-regulating immune defense-related genes.

# 1 - Introduction

Tomato (*Solanum lycopersicum*) is one of the most consumed and cultivated crop worldwide. Among many threats to this culture, the begomoviruses represent a serious constraint to the tomato plantation. The begomoviruses are whitefly-transmitted single-stranded DNA viruses, which belong to the *Geminiviridae* family (Rojas et al., 2005). The *Begomoviridae* genus encompasses more than 200 species, which collectively cause severe diseases in major crops worldwide, inflicting significant economic losses in many dicotyledonous crops (Fauquet et al., 2008). Current climate changes are expected to alter more the whitefly distribution along the globe posing a major threat to agriculture worldwide. Despite unsuccessful attempts to develop plants resistant to begomovirus infections (Day et al., 1991; Hashmi et al., 2011; Lin et al., 2012), a variety of studies have succeeded in enhancing plant tolerance to begomovirus infection (Edelbaum et al., 2009; Santos et al., 2009; Vu et al, 2013). The ectopic expression of the host protein NSP-interacting kinase (NIK), which was identified as a virulence target of the begomovirus nuclear shuttle protein (NSP), has been employed as a molecular strategy to enhance tolerance to tomato-infecting begomoviruses (Fontes et al., 2004; Carvalho et al., 2008). The resulting tomato lines displayed delayed infection that nevertheless could not be controlled at later stages.

LeNIK (*Solanum lycopersicum* NIK), AtNIK1 (*Arabidopsis thaliana*) and GmNIK (*Glycine max* NIK) have been demonstrated to be involved in plant antiviral immunity (Mariano et al., 2004, Fontes et al., 2004). Based on sequence conservation and structural features, NIK belongs to the leucine-rich repeat (LRR) receptor-like kinase (RLK) subfamily II, LRR-RLKII. The members of LRR-RLKII subfamily are basically clustered into three distinct branches: antiviral defense proteins represented by the NIK subgroup; developmental and defense proteins represented by the SERK group; and functionally unassigned proteins (Santos et al., 2010). As a member of the LRR-RLKII subfamily, NIK is structurally organized into characteristic domains, including a serine/threonine kinase domain with a nucleotide binding site at the C-terminal region, an internal transmembrane segment and leucine-rich repeats (LRR) at the N-terminal portion (Mariano et al., 2004). Like for mammalian receptors, the

oligomerization of single-pass transmembrane receptor kinases has been proposed to be either induced or stabilized by ligands as the critical early event that triggers signaling and transduction from the receptor (Wang et al., 2005; Hubbard and Miller, 2007). As a single-pass transmembrane receptor kinase, NIK is expected to dimerize or multimerize in order to activate a defense response, although the molecular bases for the ligand are totally unknown. Probably, NIK interacts with itself and/or co-receptors to promote transphosphorylation and subsequent activation of the kinase (Santos et al., 2010). The three homologs of NIK in *Arabidopsis thaliana* (AtNIK1, AtNIK2, AtNIK3) have been found to interact with NSP through their kinase domain (Fontes et al., 2004). Thus, this interaction prevents the signal transduction that evokes the defense response cascade. It has also been shown that the NSP-NIK interaction is also conserved among begomovirus NSPs and NIK homologs from different hosts (Mariano et al., 2004). The tomato orthologs of NIK, SlNIKs, have also been shown to interact with the NSP of the tomato-infecting begomovirus ToYSV (*Tomato yellow spot virus*; Sakamoto et al., 2012). They have also found that SlNIK1 and SlNIK3 have analogous functions and structural conservation with those counterparts in Arabidopsis; same protein-protein interactions, similar expression profiles and predominate in tissues that support high efficiency of begomovirus infection.

NIK1 undergoes a stepwise pattern of phosphorylation within its activation-loop domain (A-loop) with distinct roles for different threonine residues (Santos et al., 2009). The impairment of the autophosphorylation by a mutation at the residues Thr-474 or Thr-468 has been shown to lead to defective kinase activation. In contrast, mutation at Thr-469 does not affect the autophosphorylation activity; but increases the substrate phosphorylation activity, suggesting an inhibitory role for the kinase function. Therefore, a model for NIK activation was proposed upon begomovirus infection (Santos et al., 2010). This model was based on the NIK oligomerization and transphosphorylation of the kinase domain at the key threonine residue in the 474 position. This phosphorylation-dependent activation of NIK leads to the phosphorylation of a downstream component, the ribosomal protein L10A (rpL10), which in turn translocates to the nucleus, where it may mount a defense against begomovirus infection (Rocha et al., 2008; Santos et al., 2009). To overcome this activation defense mechanism, the viral NSP binds to the

kinase domain of NIK and prevents phosphorylation of Thr-474, leading to the suppression of the kinase activity and the establishment of an environment that is more favorable to begomovirus infection. Consistent with this mechanism, overexpression of Arabidopsis (At) NIK1 in begomovirus-infected tobacco leaves titrates the virally produced NSP inhibitor and overcomes NSP-mediated inhibition (Carvalho et al., 2008). Similarly, the overexpression of AtNIK1 in tomato plants attenuates begomovirus infection. However, the effectiveness of the NIK-mediated signaling pathway against begomovirus infection is limited because the viral NSP functions as a NIK suppressor. Furthermore, activation of the antiviral pathway seems to be dependent on the onset of infection.

In order to enhance the effectiveness of the NIK-mediated defense pathway against virus infection, an AtNIK1 mutant was constructed in which the Thr-474 residue was replaced with a phosphomimic aspartate residue, leading to the hyperactivation of the kinase activity. The resulting mutant receptor, designated as T474D, displayed a 1.5-fold increase in substrate phosphorylation activity and an enhanced capacity to relocate rpL10 to the nucleus (Santos et al., 2009). The NSPs from the Arabidopsis-infecting begomovirus CaLCuV (Cabbage leaf curly virus) was shown to bind stably to the kinase domain of NIK (Fontes et al., 2004). Replacing T474 with aspartate does not prevent NSP binding to NIK but decreases the NSP-mediated inactivation of the kinase activity. These in vitro results suggest that the hyperactive NIK T474D mutant may be a more effective target for engineering resistance against begomovirus.

To understand the molecular nature of the defense response triggered by NIK1 in tomato plants, we have designed an experiment with tomato plants expressing the gain-of-function receptor T474D. To examine whether the virus infection alone triggered NIK-mediated defense signaling, the transgenic lines T474D-6 and NIK1-4, the latter of which overexpresses wild-type AtNIK1, were challenged with the begomovirus ToYSV. A global comparison of the expressed sequences among the mock-treated and infected wild-type (WT), NIK and T474D lines was performed using the high-throughput RNA sequencing (RNA-seq) Illumina protocol.

RNA-seq is a sequencing platform which addresses a multitude of applications, including relative expression analyses, alternative splicing, discovery of novel transcripts and isoforms, RNA editing, allele-specific

expression and the exploration of non-model-organism transcriptomes (Anders et al., 2013). Recent studies have also reported that RNA-seq has become more accurate over a larger dynamic range of gene expression technics including microarrays (Trapnell et al., 2013; Marioni et al., 2008; Fu et al., 2009). The RNA-seq differential analysis methods focus on tackling one of two major challenges. The first one consists on accurately deriving gene and isoform expression values from raw sequencing reads, which requires statistical computations at isoform-level resolution. The second one is accounting for variability in measurements across biological replicates of an experiment (Trapnell et al., 2013). The most used strategy is to count the number of reads that fall into annotated genes and perform statistical analysis on the table of counts to discover quantitative changes in expression levels between experimental groups (Anders et al., 2013).

The ongoing improvements in RNA-seq data generation have continuously raised the necessity for more accurate statistical and computational tools. Therefore, the analysis of our RNA-seq data was performed focusing on the computational methods needed to derive more reliable conclusions from the huge amount of generated data from the experiments designed to address the NIK-mediated defense signaling pathway. We also discuss how these different methodologies can impact the interpretation of the data and, in some cases, biological data are presented to validate the interpretation from the RNA-seq data.

# 2 – Materials and Methods

## 2.1 - Tomato plant transformation

The clone pK7F-NIK1T474D has been previously described (Santos et al., 2009). It harbors a GFP gene fused in-frame after the last codon of the respective mutant cDNA under the control of the CaMV 35S promoter. In the mutant cDNA T474D, the threonine residue at position 474 within the activation loop of NIK1 was mutated to an aspartate residue. Leaf discs from *in vitro*-grown tomato plants (*Solanum lycopersicum*, cultivar Moneymaker) were transformed with pK7F-NIK1T474D via *Agrobacterium*-mediated plant transformation (strain LBA4404). The transformed shoots were selected on MS medium supplemented with 6-benzylaminopurine (500 $mg.L^{-1}$), cefotaxime (300 $mg.L^{-1}$), and kanamycin sulfate (50 $mg.L^{-1}$). The regenerated shoots were rooted, transferred into soil and grown under standardized greenhouse conditions to generate seeds. The transgenic lines were confirmed using PCR. The analysis of transgene expression was performed by RT-PCR and real-time RT-PCR using transgene-specific primers, and actin was used as an endogenous control to normalize all values. The transgenic line 35S:NIK1-4 (NIK1-OX), which overexpresses the NIK1 receptor, has been previously described (Carvalho et al., 2008).

## 2.1 – RNA Sequencing

The transgenic and wild-type lines were infected at the six-leaf stage with ToYSV-[MG-Bi2], by biolistic delivery using tandemly repeated viral DNA-A and DNA-B and a microprojectile bombardment model PDS-1000/He accelerator (BIORAD) at 900 psi. In each experiment, 20 plants of each line were inoculated with 2 µg of tandemly repeated DNA-A plus DNA-B per plant and grown in a greenhouse under natural conditions of light, 70% relative humidity and approximately equal day and night lengths. Total nucleic acid was extracted from the systemically infected leaves (young leaves), and viral DNA was detected by PCR using DNA-A and DNA-B begomovirus-specific primers. After

10 days post-inoculation, total RNA from systemically infected leaves, as diagnosed by PCR, and mock-inoculated leaves from wild-type, 35S::NIK1-4 and 35S::T474D lines was isolated using TRIzol (Invitrogen). For the RNA sequencing experiments, we used two biological replicates of a pool of 10 plants at 10 days after inoculation when we detected high levels of viral DNA in systemic leaves but symptoms were not visible as yet.

Total RNA from wild-type mock-inoculated and ToYSV-infected plants, 35S::NIK1-4 (NIK1-OX) mock-inoculated and infected plants and 35S::T474D (T474D) mock-inoculated and infected-plants was isolated using TRIzol (Invitrogen).

The RNA sequencing was obtained using single end approach by Illumina Genome Analyzer IIx in the Fasteris facilities. The GEX-NIaIII protocol was used with the following quality filter parameters: maximum of 1 base below a quality of 5 in the first 30 bases, a minimum average quality of 10, no “N” calls allowed and not more than 35 identical bases (low information reads). The data were stored in a comma-separated values (csv) spreadsheet file.

**2.2 – Data sets**

The reads were pre-processed (filtering) by the Illumina Genome Analyzer IIx software in the Fasteris facilities and were organized into a counting table containing 60,000 different reads at length of 21 bp (Table 1). The experimental design was 6 treatments as follow: mock-inoculated (mock WT, mock NIK1-OX, mock T474D) lines and infected (inf WT , inf NIK1-OX, inf T474D) lines with 2 repetitions. The reads expression is counting values in each sample column.

Table1. Sequence depth of the libraries and treatments

| Treatments | Library | Total Reads |
|---|---|---|
| Mock WT | GHY.1 | 1 304 801 |
| | GHY.2 | 9 153 036 |
| Infected WT | GHY.3 | 834 934 |
| | GHY.4 | 776 138 |
| Mock NIK1-OX | GHY.5 | 4 187 108 |
| | GHY.6 | 3 065 855 |
| Infected NIK1-OX | GHY.7 | 4 192 870 |
| | GHY.8 | 4 306 404 |
| Mock T474D | GHY.9 | 4 367 471 |
| | GHY.10 | 2 935 660 |
| Infected T474D | GHY.11 | 4 507 533 |
| | GHY.12 | 3 483 973 |

The annotation files were obtained at the Phytozome database web site (http://www.phytozome.net) provided by the International Tomato Annotation Group (ITAG) at the release 2.3 (Tomato Genome Consortium 2012). One cDNA file in FASTA format containing 34727 sequences, and one genome file in FASTA format with the 12 chromosomes sequences and 1 mitochondria DNA sequence were the annotation source.

**2.2 – Mapping the reads to a reference**

Although, alignment of reads is a classic problem in bioinformatics with several solutions specifically for a short-read mapping (Kent, 2002; Wu and Watanabe, 2005), RNA-seq reads pose particular challenge. They are short (~21 – 125 bases), error rates are considerably high and many reads span an exon-exon junction. Additionally, the number of reads per experiment is increasingly large, currently as many as hundreds of millions. To overcome these problems, there are two major algorithmic approaches to map the RNA-seq reads to a reference genome or transcriptome (Langmead et al., 2009). The first is referred to as 'unspliced read aligners', which aligns the reads to a reference without allowing any large gaps. The unspliced read aligners fall into

two main categories, 'seed methods' and 'Burrows-Wheeler transform methods' (Garber et al., 2011a). Seed methods, such as mapping and assembly with quality MAQ (Ruan, and Durbin, 2008) and Stampy programs (Lunter and Goodson, 2011), find matches for short subsequences, termed 'seeds', assuming that at least one seed in a read will perfectly match the reference. Each seed is used to narrow candidate regions in which more sensitive methods (such as Smith-Walterman algorithmic) can be applied to extend seeds to full alignments. In contrast, the second approach includes Burrows-Wheeler (BWA) and Bowtie alignments (Langmead et al., 2009), which compact the reference into a data structure that is very efficient when searching for perfect matches. If allowing mismatches, the performance of Burrows-Wheeler methods decreases exponentially with the number of mismatches as they iteratively perform perfect searches (Garber et al., 2011b). Another type of aligners is considered the spliced events. Those algorithmics predict exons and junctions and attempt to recognize the splicing events involving new exons. Reads can be aligned to the entire genome, including intron-spanning reads that require large gaps for proper placement. Several methods exist, collectively referred to as 'spliced aligners', which fall into two main categories: 'exon first' and 'seed and extend'. Exon-first methods such as MapSplice (Wang et al., 2010), SpliceMap (Au et al., 2010) and TopHat (Trapnell et a;., 2009) use a two-step process. First, they map reads continuously to the genome using the unspliced read aligners. Second, unmapped reads are split into shorter segments and aligned independently. The genomic regions surrounding the mapped read segments are searched for possible spliced connections. Exon-first aligners are very efficient because only a small portion of the reads requires an intensive computation. Alternatively, seed-extend methods, such as 'genomic short-read nucleotide alignment program' (GSNAP) and 'computing accurate spliced alignments', PALMapper/QPALMA (Jean et al., 2010), break reads into short seeds, which are placed onto the genome to localize the alignment. Candidate regions are then examined with more sensitive methods, such as the initial seeds to determine the exact spliced alignment for the read. Many of these alignments methods also support paired-end read mapping, which increases alignment specificity (Garber et al., 2011b). Although, the 'spliced aligners' have an advantage towards 'unspliced aligners' because they would be able to detect new isoform of a gene or even a new gene, many splicing

junction algorithms are always optimized to one specific group of organism, such as mammalian; thus, the bias of this detection could strain the downstream analysis and the biological interpretation. In addition, the small length of the reads probably would increase the false positive of the mapped reads.

Therefore, to understand how the mapping algorithms influence our analysis, we have applied three 'unspliced aligners': Bowtie (Burrows-Wheeler transform), Stampy and MAQ (seed method), and two 'spliced aligners' Tophat (exon first) and PALMapper (seed-extend). We also compared those results with the classic Smith-Walterman (SW) algorithmic (implemented in this work) direct to the ITAG 2.3 cDNA data set.

### 2.3 – Normalization

The term normalization has been widely used in many biology fields as synonym of any data corrections or transformation. Although in many earlier statistical reports, it has been directly related to a data transformation or correction in order to fit a normal distribution, or even a more sophisticated adjustment in which the data are brought to a symmetrical distribution (Dutka and Hanson, 1989; Grigelionis, 1990), this term remains in the gene expression literature as a method of correction regardless of the data dispersion. Therefore, normalization has been reported as an important step in gene expression quantification, and an understanding how these methods work could provide meaningful biological insights (Dillies et al., 2012).

In order to estimate the correct gene expression on RNA-seq, some studies have reported the necessity of a proper normalization (Kadota et al., 2012; Dillies et al., 2012; Robinson and Oshlack, 2010; Bullard et al., 2010). As these data are expressed by reads count per gene, a meaningful expression estimative can be extracted by an appropriate normalization choice (Garber et al., 2011b). There are two main sources of systematic variability that require normalization. First, RNA fragmentation during library construction causes longer transcripts to generate more reads compared to shorter transcripts present at the same abundance in the sample. Second, the variability in the number of reads produced for each run causes fluctuations in the number of

fragments mapped across samples (Marioni et al., 2008; Mortazavi et al., 2008). Therefore,, these sources of variability differ both in the type of bias adjustment and in the statistical strategy that must be adopted for normalization (Kadota et al., 2012). However, as data accumulate, there is still no clear indication of how the choice of normalization method impacts the downstream analysis (Dillies et al., 2012). In addition, although effective and relevant methods have been derived and implemented to normalize RNA-seq data, they are not always properly used in practice. A small number of publications have compared normalization methods (Bullard et al., 2010), providing useful yet preliminary results that must be confirmed with additional data to yield clear and robust guidelines to the community (Dillies et al., 2012). Comparisons among the available normalization methods for gene expression analysis have either made use of simulation studies or real calibration data (Shedden et al., 2005; McCall and Irizarry, 2008; Qin et al., 2006; Jeanmougin et al., 2010). These studies rely on both the qualitative characteristics of normalized data and the impact of the normalization method on the results from a differential expression (DE) analysis. In addition, there are few investigations about the impact of the normalization method on the false-positive rate and on the detection power of DE analysis (Kadota et al., 2012).

Additional effects on differential expression analysis have also been reported, such as GC-content and gene length (Risso et al., 2011). Although the GC-content and gene length of each gene do not change from sample to sample, so they can be expected to have little effect on differential expression analyses, precedents in the literature have demonstrated that sample-specific effects for GC-content can still be detected (Risso et al., 2011; Hansen et al., 2012). Unlike the GC content, there is no strong evidence that gene length could have any effect on differential expression analysis, although a minor influence of gene length has been detected by Hansen et al., (2012).

A systematic comparison in our tomato data set of five representative normalization methods with and without the correction factors for CG-content and gene length was performed using the Bioconductor packages edgeR (Robinson et al., 2010), DESeq (Yang et al., 2013), and EDASeq (Risso et al., 2011). We have also adopted as a comparative experimental design a pairwise comparison and a false discover rate (FDR) p-value adjustment with the cutoff < 0.05. Four of those normalization methods were implemented on the edgeR:

Total Count (TC), Upper Quartile (UQ) (Bullard et al., 2010), Relative log Expression (RLE) (Anders and Huber, 2010) and Trimmed Mean of M values (TMM) (Robinson and Oshlack, 2010), and the last one was the normalization method implemented direct in DESeq package. The GC-content and gene length correction factors were performed by the Bioconductor package EDASeq (Risso et al., 2011). This package mainly focuses on biases related to GC-content and the existence of strong sample-specific GC-content effects on RNA-Seq read counts, which could substantially bias differential expression analysis. These methods are compared to state-of-the-art normalization procedures in terms of bias and mean squared error for expression fold-change estimation and in terms of type I error and p-value distributions for tests of differential expression.

**2.4 – Deferential Expression Detection**

After the normalizations and before the analysis of the Differential Gene Expression (DGE) methods, Robinson and Oshlack (2010) suggest an interpretation of the biological coefficient of variation (BCV) of the RNA-seq samples. They stated that this is the coefficient of variation (CV) with which the (unknown) true abundance of the gene varies between replicate samples. It represents the CV that would remain between biological replicates if sequencing depth could be increased indefinitely. The technical CV, another component of the total variation, decreases as the size of the counts increases. BCV on the other hand does not. BCV is therefore likely to be the dominant source of uncertainty for high-count genes, so reliable estimation of BCV is crucial for realistic assessment of differential expression in RNA-Seq experiments (McCarthyet al., 2012). If the abundance of each gene varies between replicate RNA samples in such a way that the genewise standard deviations are proportional to the genewise means, a commonly occurring property of measurements on physical quantities, then it is reasonable to suppose that BCV is approximately constant across genes. Therefore, the magnitude of BCV is more important than the exact probabilistic law followed by the true gene abundances (McCarthy et al., 2012). For mathematical convenience, they assume that the true gene abundances follow a gamma

distributional law between replicate RNA samples. This implies that the read counts follow a negative binomial probability law.

Having quantified the BCV and normalized expression values, an important question is to understand how these expression levels differ across conditions. An extensive methodology for the statistical analysis of differential expression using microarrays has been developed in the last decade (Garber et al., 2011b). Although in principle these approaches are directly applicable to RNA-seq data as well, using read coverage to quantify transcript abundance provides additional information such as genetic variation and epigenetic state to transcriptional and post-transcriptional regulation (Trapnell et al., 2013). Moreover, as it has been shown in some studies of normalization analysis (Kadota et al., 2012; Dillies et al., 2012), the power to detect differential expression depends on many factors throughout the experiments. To accommodate the count-based nature of RNA-seq data, the original methods have modeled the observed reads using count-based distributions such as the Poisson distribution (Marioni et al., 2008; Trapnell et al., 2010; Jiang and Wong, 2009). However, several studies have reported that these distributions do not account for biological variability across samples (Li et al., 2010; Robinson and Smyth, 2007). Ideally, if one had enough replicates the variability across replicates could be estimated empirically using a permutation-derived approach (Grant et al., 2007; Grant et al, 2005), similar to that used by the Myrna method (Langmead et al., 2010). However, to date few RNA-seq expression studies have generated a sufficient number of replicates to achieve this goal (Garber et al., 2011b). To overcome this limitation, many methods attempt to model biological variability and provide a measure of significance in the absence of a large number of biological replicates. These methods, such as EdgeR (Robinson et al., 2010), differential expression analysis of count data (DESeq) (Anders and Huber, 2010), DEGSeq (Wang et al., 2010) and baySeq (Hardcastle and Kelly, 2010), model the count variance across replicates as a non-linear function of the mean counts using various different parametric approaches (such as the normal and negative binomial distributions) (Marioni et al., 2008; Trapnell et al., 2010; Robinson and Smyth, 2007; Anders and Huber, 2010; Wang et al., 2010). Although these approaches can assign significance to differential expression, the biological conclusions must be interpreted with caution (Garber et al., 2011b). For example, although the variability of the

sequencing process is low compared to microarray hybridization (Marioni et al., 2008), measurements can vary substantially because of differences in library construction protocols and most importantly because of intrinsic variability in biological samples (Levin et al., 2010). As with any biological measurement, biological replicates provide the only measurement of intrinsic, nontechnical transcript expression variability and, thus, are critical for differential expression analysis (Garber et al., 2011b).

In order to deal with this intrinsic variability on RNA-seq experiments, one of the most important issues among DE analysis is the distribution by which the counts mapping to a gene are fitting. As the non-negative integers follow a discrete distribution, in the methods, explicitly developed for differential expression analysis of this type of count data, the Poisson distribution and the negative binomial (NB) distribution are the two most commonly used models (Marioni et al., 2008; Robinson and Oshlack, 2010; Anders and Huber, 2010). Other distribution, such as the beta-binomial (Hardcastle and Kelly, 2010), has also been proposed. The Poisson distribution has the advantage of simplicity and has only one parameter, but it constrains the variance of the modelled variable to be equal to the mean. The negative binomial distribution has two parameters, encoding the mean ($\mu$) and the dispersion ($\varphi$), and hence allows modeling of more general mean-variance relationships. For RNA-seq, it has been suggested that the Poisson distribution is well suited for analysis of technical replicates, whereas the higher variability between biological replicates necessitates a distribution incorporating overdispersion, such as the negative binomial (Robinson and Smyth, 2007; Anders and Huber, 2010). Because of the Poisson model is a special case of the negative binomial, namely the case when dispersion ($\varphi$) $\rightarrow 0$, then, the negative binomial can actually be applied for both biological and technical replicates (Robinson and Smyth, 2008). Our analyses in the tomato RNA-seq experiment employed the normalized data provided by the counting table to the most common negative binomial methods presented in R/Bioconductor software: edgeR (Robinson et al., 2010), DESeq (Anders and Huber, 2010) and baySeq (Hardcastle and Kelly, 2010).

## 2.4 – Downstream analysis

Only a small percentage of the proteins encoded in plant genomes are sufficiently characterized with regard to their cellular functions (Monaco et al., 2013). The majority of plant proteins remain either completely unknown or only partially understood. In *Solanum lycopersicum* (tomato plant), in which the complete genome sequencing was released in 2012 (Tomato Genome Consortium 2012), the number of predicted protein-coding genes is 34,727 and 31,741 show high similarity to *Arabidopsis* genes. Among those predicted genes, at least 6,874 (19.79%) have no information about their sequences and around 17,786 (47,14%) have been poorly characterized. Despite this significant deficit in the annotation provided by the consortium database, our understanding of the molecular functions of the tomato plant transcriptome could be fundamentally revealed by analyzing the known genes through the computational methods provided by a wide spectrum of statistical and machine learning techniques (Monaco et al., 2013; Kawaji and Hayashizaki, 2008; Punta et al., 2011; Ashburner et al., 2000). A few major methods are generally in use for predicting protein functions in almost all organisms. The widely used approaches consider the similarity of proteins. It shows the detectable sequence or structural similarities to functionally characterized proteins in reference databases (Horan et al., 2008). In contrast, the more conservative empirical approach as the ortholog clusters could be applied to ensure with high confidence the evolutionary and functional relationship among the clustered proteins as well as with the supported information of the direct experimental evidence (Suzek et al., 2007). For instance, when a group of related sequences contains one or more members of known function, then the similarity approach tends to assign all of them as analogs, whereas the empirical approach distinguishes between functionally characterized and uncharacterized candidates within groups of related sequences. As all these approaches have already been applied by the International Tomato Annotation Group (ITAG) for the genome sequencing published by Tomato Genome Consortium 2012, this annotation was considered as our main reference for all analysis. It contains proteins names and functional description with their gene ontology (GO -

Ashburner et al., 2000) and KEGG (Kanehisa, 2013) information.

The entire annotation data set from ITAG/Phytozome (http://www.phytozome.net) was stored in the relational database PostgreSQL 9.3 (http://www.postgresql.org). The scripts operations in the database using the SQL (Structured Query Language) facilitate the data mining and the structural organization of the retrieved information. Yet, a deeper understanding of the biological meaning of the gene expression data requires a more powerful approach (Subramanian et al., 2005; Horan et al., 2008; Cumbie et al., 2011). The Gene Set Enrichment Analysis (GSEA) has become the most widely used annotation technique since the emerging of the microarray technology (Jansen and Gerstein, 2000; Subramanian et al., 2005). GSEA has also been applied for RNA-seq annotation analysis similarly as for microarray. The basic goal of GSEA is the detection of biological processes in gene sets instead of focused on identifying individual genes. GSEA features a number of advantages as compared with single-gene methods. First, it eases the interpretation of a large-scale experiment by identifying pathways and processes. Rather than focus on high scoring genes, which can be poorly annotated and may not be reproducible, the efforts can focus on gene sets, which tend to be more reproducible and more interpretable. Second, when the members of a gene set exhibit strong cross-correlation, GSEA can boost the signal-to-noise ratio and make it possible to detect modest changes in individual genes. Third, the leading-edge analysis can help define gene subsets to elucidate the biological phenomenon implied by the results (Subramanian et al., 2005). To detect gene set enrichment from our RNA-seq tomato infection DE data, we used the GSEA method provided by the R/Bioconductor GSEABase package based on the Gene Ontology (GO) database (Ashburner et al., 2000). In order to increase the pattern detection among the enrichment genes, we applied the cumulative hypergeometric function using KEGG groups (Kanehisa, 2013).

### 2.5 - *In vivo* labeling of leaf proteins

Tomato seedlings (300 mg) were incubated with 1 mL of nutrient solution containing 50 µg/ml chloramphenicol and 20 µCi of [35S]methionine (EasyTag Protein Labeling Mix, [$^{35}$S]-, 2mCi (74MBq), Perkin Elmer) for 3 h at room

temperature. To quantitate incorporation of [$^{35}$S]methionine into protein, aliquots of protein extracts were placed in 10% (w/v) TCA and incubated on ice for 30 min. The samples were filtered onto glass microfiber filters and the filters were washed three times with 5 ml of cold 5% (w/v) TCA and two times with 5 ml of 95% ethanol. After drying, the filters were counted with a scintillation counter.

**2.6 - Polysome fractionation**

Polysomes were fractionated over sucrose gradients. Briefly, 500 mg of 15-day-old tomato seedlings were ground in liquid nitrogen and 1 mL of extraction buffer (0.2 M Tris-HCl, pH 8.0, 50 mM KCl, 25 mM MgCl2, 1% Triton X-100, 400 units/mL of RNasin and 50 mg/mL of cycloheximide). After centrifuging for 10 min, the supernatant was loaded onto a 10-mL 15% to 50% sucrose gradient and spun in a Beckman SW41Ti rotor at 135,000 g for 3.5 h. Fractions were collected manually from the bottom, and total RNA was extracted with phenol/chloroform/isoamyl alcohol, precipitated with isopropanol, and treated with DNase I. The specific transcripts were amplified using semiquantitative RT-PCR. For cDNAs from the T474D line, the RT-PCR involved 25 cycles for rbcS, actin and the resistance-like gene targets, and each cycle comprised 95 ℃ for 30 s, 60 ℃ for 30 s and 72 ℃ for 1 min. The same conditions were used for the amplification of control genes from cDNAs from the wild-type line. Amplification of the resistance-like cDNA fragments from wild-type samples was performed with 45 cycles and an annealing temperature of 52℃.

**2.7 - Infectivity assays**

For the infectivity assays, we used T2 transgenic plants harboring the T474D mutant gene construct, which were derived from four independently regenerated kanamycin-resistant plants (35S::T474D-2, 35S::T474D-5, 35S::T474D-6 and 35S::T474D-9). We also used the previously described transgenic lines expressing AtNIK1 under the control of the CaMV 35S promoter, 35S::NIK1-4 and 35S::NIK1-6 (Carvalho et al., 2008). The transgenic

and wild-type lines were infected at the six-leaf stage with either ToYSV-[MG-Bi2] or ToSRV by biolistic delivery using tandemly repeated viral DNA-A and DNA-B and a microprojectile bombardment model PDS-1000/He accelerator (BIORAD) at 900 psi. In each experiment, 20 plants of each line were inoculated with 2 µg of tandemly repeated DNA-A plus DNA-B per plant and grown in a greenhouse under natural conditions of light, 70% relative humidity and approximately equal day and night lengths. Total nucleic acid was extracted from the systemically infected leaves (young leaves), and viral DNA was detected by PCR using DNA-A and DNA-B begomovirus-specific primers (PBL1v 2040, GCCTCTGCAGCARTGRTCKATCTTCATACA, and PCRC1, CTAGCTGCAGCATATTTACRARWATGCCA, or PAL1v1978, GCATCTGCAGGCCCACATYGTCTTYCCNGT, and PAR1c496, AATACTGCAGGGCTTYCTRTACATRGG) at 10 days post-infection.

**2.8 - Quantitation of viral DNA in infected plants**

Viral DNA accumulation was measured by quantitative PCR (qPCR). The reactions were prepared in a final volume of 10 µl using the Fast SYBR Green Master Mix (Applied Biosystems) according to the manufacturer's instructions and analyzed on a 7500 Real Time PCR System (Applied Biosystems). Virus-specific primers were designed using Primer Express 3.0 (Applied Biosystems) and tested by conventional PCR using plasmids containing the complete DNA-A of each virus ($10^6$ copies per reaction). The following primer sequences were used: ToSRVFwd, CACGTGCCCACATCGTCTT, and ToSRVRev, GGCCGGAACGACCTATTA-3', or ToYSVFwd, CCACGATTTTAAAGCTGCATTCT, and ToYSVRev, CAATCCTGGTGAGGGAGTCAGT. For viral DNA quantitation, standard curves were prepared using serial dilutions of these clones ($10^0$ to $10^6$ copies of viral genome per reaction). Standard curves were obtained by regression analysis of the Ct values of each of the three replicates of a given dilution in relation to the log of the amount of DNA in each dilution. For the absolute quantitation of the number of viral DNA molecules in the different treatments, 100 ng of total DNA from the infected plants was used in the qPCR reactions containing virus-specific primers. Each sample was analysed in triplicate from at least two biological replicates.

# 3 – Results

## 3.1 - Best statistical and computational program adjustments for the analysis of our RNA-seq data

The preprocessed reads were mapped to the tomato plant genome and transcriptome (ITAG 2.3) files and no differences were observed between the mapped reads from the genome and the transcriptome. Thus, we chose the transcript file as our main reference. The Figure 1 shows the quantity of the reads mapped according to the aligners software. The SW algorithm and Bowtie detected the same amount of mapped reads (Figure 1A) and the outcome from the others aligners programs were exactly the subsets of the Bowtie/SW (Figure 1B). Therefore, we used the result provided by the Bowtie for our analysis, as. the SW/Bowtie had a better performance on our mapping analysis.

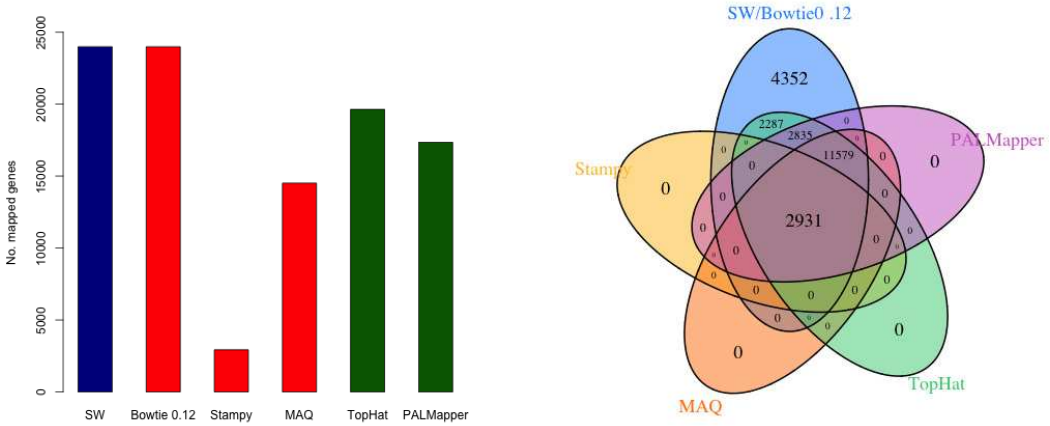



Figure 1: Number of mapped reads from genes. (A) Number of mapped genes by software: Red is 'unspliced aligners' and green is 'spliced aligners'. (B) Venn diagram showing the numbers of mapped gene detected by the aligners.

The counting tables provided by the mapping aligners were the input of the normalization methods. They deal with the bias from the sequencing process and attempt to strip off any technical variation of the data. Thus, the normalization factors were employed to the counting table with and without the EDASeq corrections (GC% content and gene length). When the boxplots graphics (Figure S2) were analyzed, they do not strongly suggest that the

overall data distribution had any modification. Only the scale has varied. The effects of the five normalization factors seem to have little influence on the data dispersion, but as shown in Figure 2, when the differential gene expression (DGE) was calculated, the amount of up and down-regulated genes showed a great variability among the conditions and normalization factors.

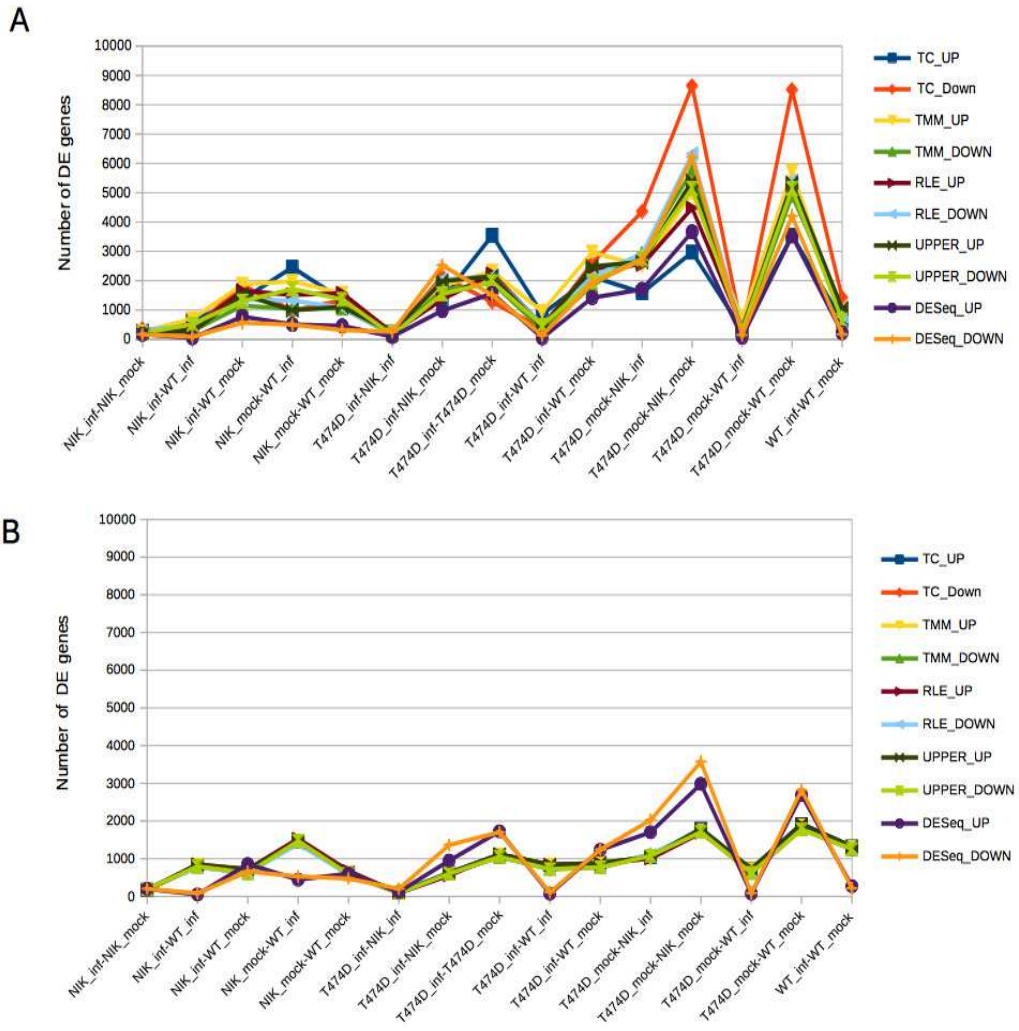



Figure 2: Number of differential expressed reads (A) and mapped genes with EDASeq corrections (B) by pairwise conditions. Each line represents a normalization method separated by up and down regulated.

In Figure 2A, the TC normalization shows a massive down-regulation trend in the differential expressed genes in the groups T474D_mock-WT_mock and T474D_mock-NIK_mock. In contrast, the Figure 2B exhibits a more balanced up and down-regulated genes, i.e., the application of the two corrections factors seemed to shrink the number of DE genes. It probably

occurs because the normalization factors were designed to correct the bias by applying a calculated factor according to the experimental design in the raw data (Dillies et al., 2012). When the data were previously treated by the EDASeq, the application of normalization factors may have imprinted another layer of adjustment in the analysis; thereby, diminishing the difference among those normalization factors as well as the variability within pairwise conditions. In order to exploit this issue, we have performed a short simulation study using the R/Bioconductor package TCC (Kadota, Nishiyama, and Shimizu 2012). This simulation study was set up for a direct application of the five normalization factors in the raw and EDASeq corrections data sets. The sequence depth and the proportion of up/down-regulated DE genes found by TC normalization on our real data were used as the real simulation parameters (Table 2). The Figures 3B and 3C revealed that both data sets, the raw and EDAseq, differed only by the DE genes numbers instead of proportions of up/downregulated genes. None of those methods in the simulated data were able to find out the proportion of 70% downregulated and 30% of upregulated DE genes set up in the simulations 1 and 3 (Figure 3A). In addition, this simulation suggests the variability among the normalization factors is not as huge as shown on the real data set (Figure 3A). Thus, the EDASeq corrections can definitely eliminate some real effects that could be present in a biological variation. Furthermore, many of the normalization factors such as TMM (Robinson and Oshlack 2010) are an empirical strategy that equates the overall expression levels of genes between samples under the assumption that the majority of them are not differentially expressed (DE). However, it would not be possible to use only the sequence depth correction provided by the TC without any further corrections, because the down-regulated genes could be over represented upon the up-regulated. Therefore, we decided to continue only with the raw data set without the EDASeq corrections, as for our data, the direct application of the normalization factors has already been able to correct enough bias.

Table 2. Proportions of up/down-regulated DE genes found by TC normalization

| Simulation number | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| Contrast | WTinf-WTmock | NIKmock-WTmock | T474Dmock-WTmock | NIKinf-NIKmock |
| DGE: | 3% | 4% | 20% | 1% |
| UP: | 0.31 | 0.5 | 0.29 | 0.4 |
| Down: | 0.69 | 0.5 | 0.71 | 0.6 |
| | | | | |
| Simulation number | 5 | 6 | 7 | 8 |
| Contrast | NIKmock-WTinf | NIKinf-WTinf | NIKinf-WTmock | T474Dinf-NIKinf |
| DGE: | 6% | 2% | 5% | 1% |
| UP: | 0.71 | 0.66 | 0.42 | 0.62 |
| Down: | 0.29 | 0.34 | 0.58 | 0.38 |
| | | | | |
| Simulation number | 9 | 10 | 11 | 12 |
| Contrast | T474Dinf-NIKmock | T474Dinf-T474Dmock | T474Dinf-WTinf | T474Dinf-WTmock |
| DGE: | 6% | 8% | 2% | 8% |
| UP: | 0.39 | 0.74 | 0.70 | 0.45 |
| Down: | 0.61 | 0.26 | 0.30 | 0.55 |
| | | | | |
| Simulation number | 13 | 14 | 15 | |
| Contrast | T474Dmock-NIKinf | T474Dmock-NIKmock | T474Dmock-WTinf | |
| DGE: | 10% | 19% | 1% | |
| UP: | 0.27 | 0.26 | 0.5 | |
| Down: | 0.73 | 0.74 | 0.5 | |

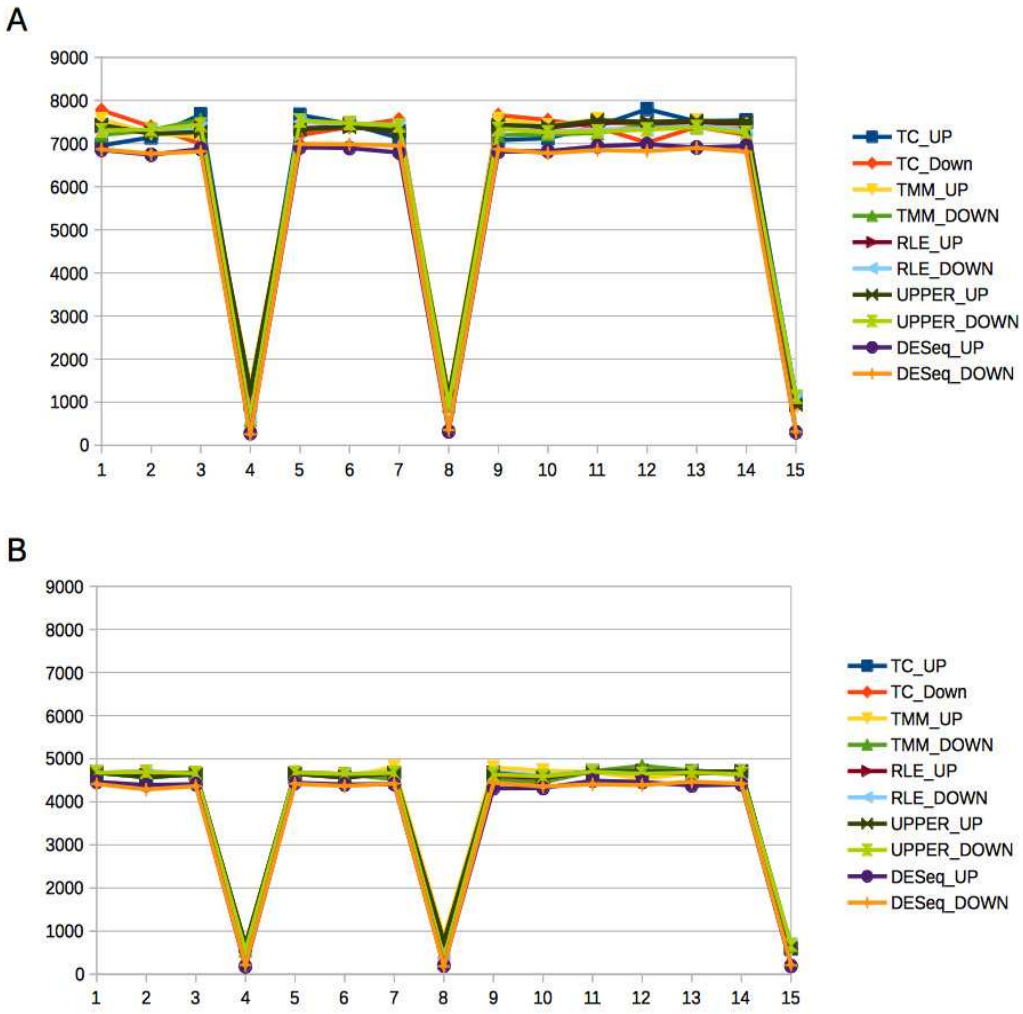



Figure 3. RNA-seq simulations performed by TCC package between the uncorrected and corrected data set by EDASeq (CG content and gene length). (A) DE genes proportions of TC normalization data sets. (B) number of DE genes without EDA (C) EDASeq corrected data sets (C) comparing the normalizations factors.

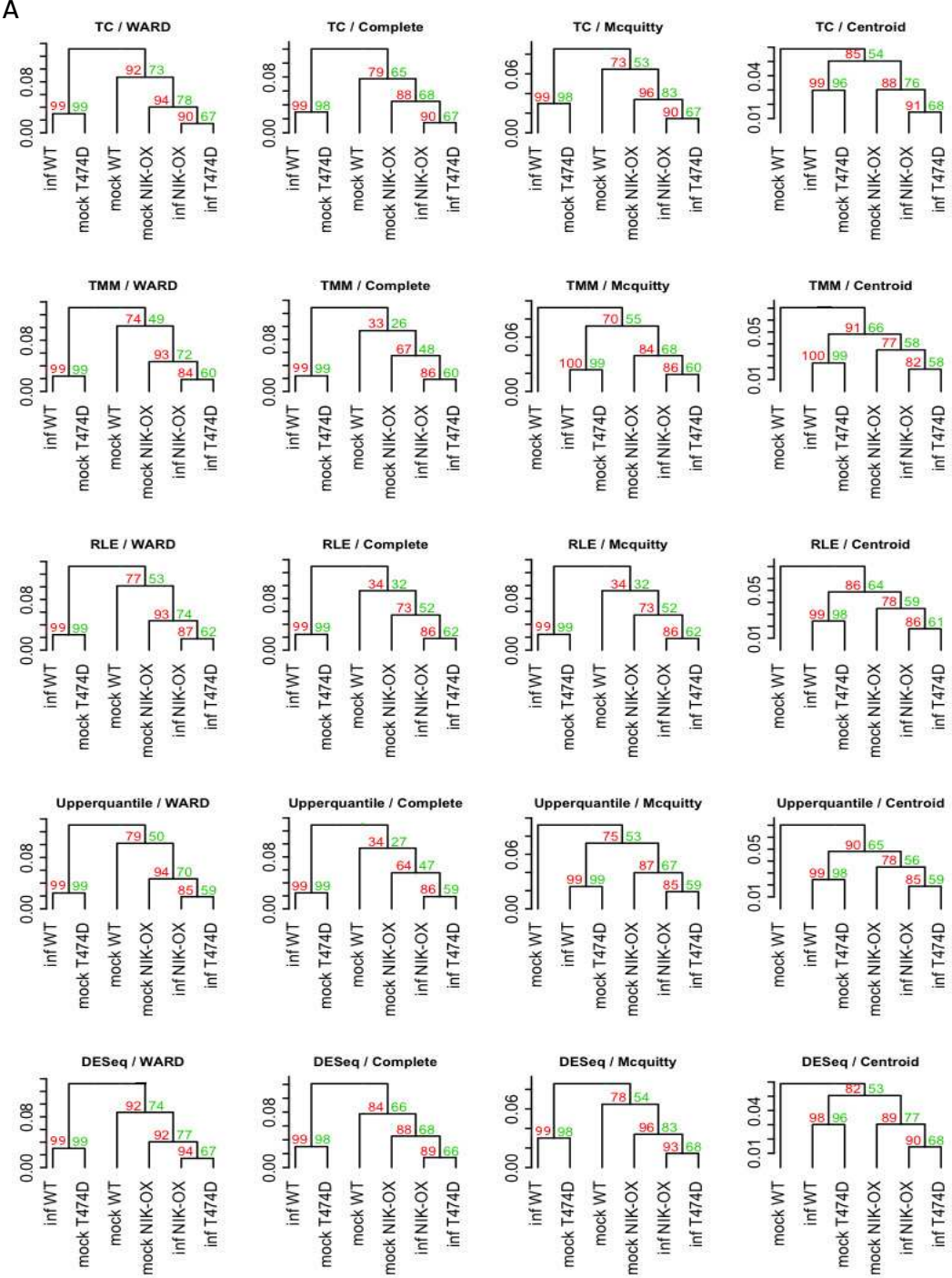
### 3.2 - The virus infection is the trigger of the NIK-mediated antiviral signaling

We next examined the overall expression profile among all the treatments. The hierarchical clustering via multiscale bootstrap resampling method was employed to obtain the clusters from the normalized treatments data sets. We have applied four different agglomerative cluster methods: ward (Ward 1963), complete (Hartigan 1975), mcquitty (McQuitty 1966) and centroid

(Hartigan 1975). All of these methods are implemented in the R package pvclust (Shimodaira 2002).

The transcriptomes of mock T474D and infected WT clustered together regardless of the clustering and normalization methods used (Figure 4A). The values of the approximately unbiased p-value (au) and the bootstrap probability (bp) were significative using the threshold of 0.05 for all combined clustering and normalization methods. Therefore, this analysis clearly shows up a strong similarity between the general gene expression profile of the mock T474D and infected WT plants. Consistent with this finding, by trimming the mock T474D – mock WT DE genes off from all treatments, the effect of viral infection seems to be titrated off and each T474D, NIK-OX and WT mock-inoculated treatment grouped together with its infected counterpart. This shift was independently shown by 75% of the combined clustering and normalization methods (Figure 4B).

A


TC / WARD
TC / Complete
TC / Mcquitty
TC / Centroid
TMM / WARD
TMM / Complete
TMM / Mcquitty
TMM / Centroid
RLE / WARD
RLE / Complete
RLE / Mcquitty
RLE / Centroid
Upperquantile / WARD
Upperquantile / Complete
Upperquantile / Mcquitty
Upperquantile / Centroid
DESeq / WARD
DESeq / Complete
DESeq / Mcquitty
DESeq / Centroid
inf WT
mock T474D
mock WT
mock NIK-OX
inf NIK-OX
inf T474D

B

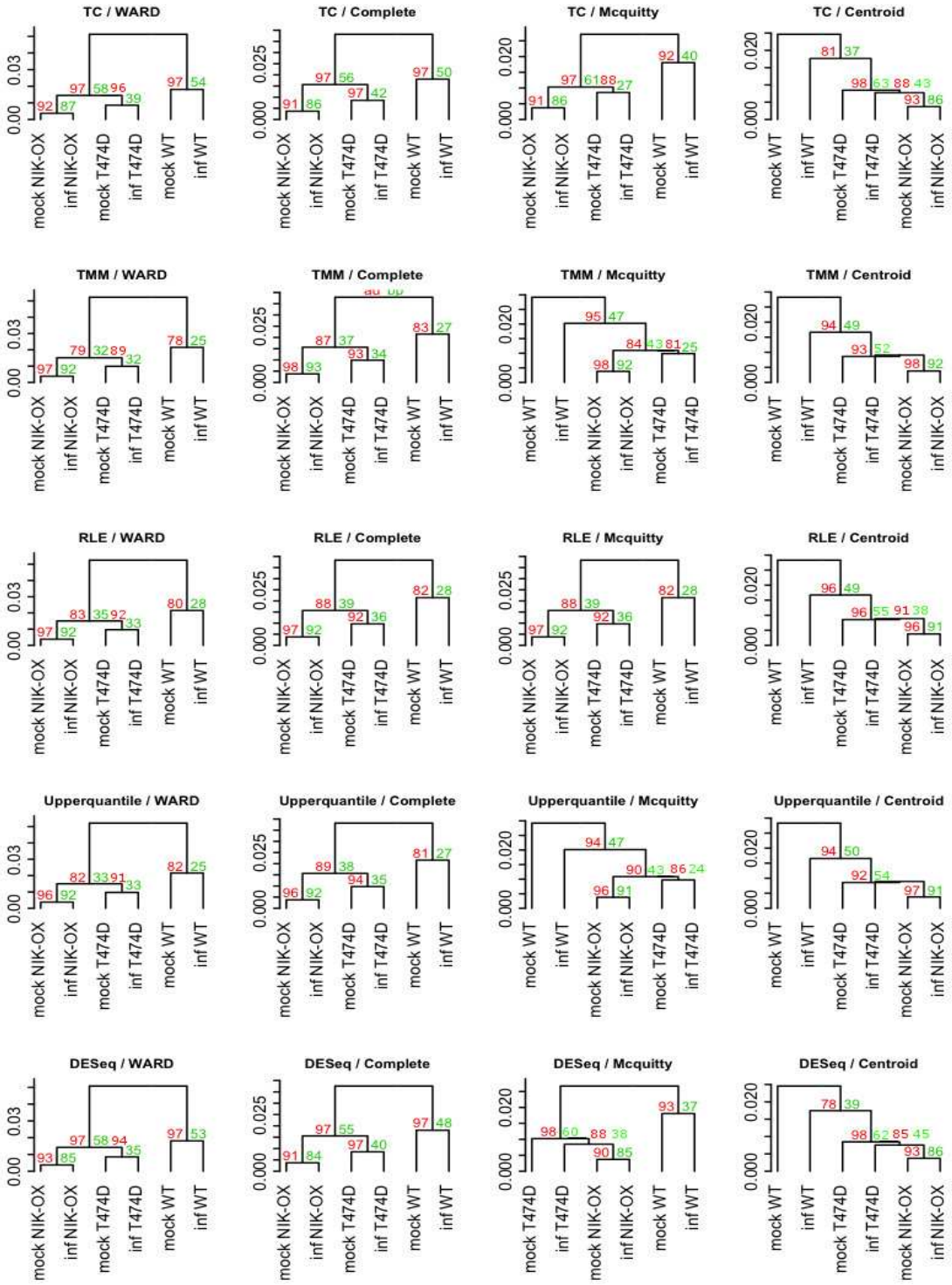



Figure 4: Clustering comparison between normalization factors and the agglomerative clustering methods. (A) The raw data without any previews correction before the normalization. (B) The raw data trimmed off the DE genes of mock T474D – mock WT. The red numbers means unbiased p-value (au) and green ones bootstrap probability.

The similarity between the mock T474D and infected WT genes profiles suggested that the T474D NIK-mediated and the infection response can share a similar portrait. In addition, it seems that a sustained NIK-mediated response led to a “priming” state that further enhanced the response to begomovirus infection in T474D- and NIK-overexpressing leaves; the virus-induced transcriptomes of these samples cluster together and differ from the T474D mock-inoculated transcriptome (Figure 4A). Furthermore, the elimination of the mock T474D DE genes from the raw of all treatments further indicates that the expression profile induced by the T474D mutant mimics greatly the response to the viral infection, as the mock- and infected-induced transcriptomes from each genotype clustered together with high significance (Figure 4B). Taken together, these results suggest that the gain-of-function mutant T474D can sustain an activated NIK-mediated antiviral response in the absence of the virus and that virus infection is the trigger of the NIK-mediated signalling pathway.

### 3.3 - Ectopic expression of the gain-of-function T474D mutant causes a general down-regulation of translation-related genes

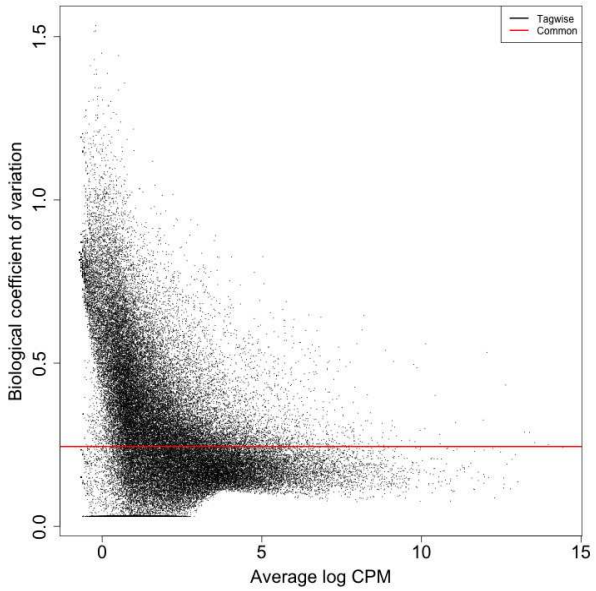



Figure 5: Biological Coefficients variations (BCV) dispersion against log2-CPM (count per milion) of RNA-seq tomato samples.

To estimate the differential expression of a gene, we estimated first the overall Biological Coefficient Variation (BCV), which was 17% for the total

variation of the experiment, by using thee edgeR package. . This BCV is not considered atypical for technical replicates in observational studies (Oberg et al., 2012). The edgeR package also allows us to estimate the BCV per gene using the tagwise method (Robinson et al., 2010). The Figure 5 shows the estimated BCV for each gene as the black dots and the red line, as the overall BCV. This analysis also demonstrates that there is a group of genes responsible for a great variation among the entire experiment.

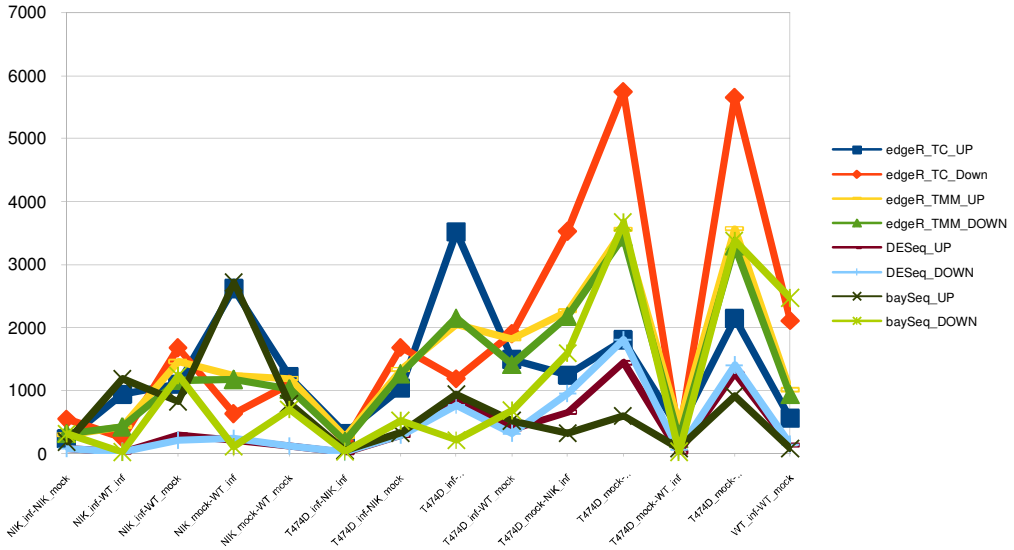



Figure 6: Comparisons among DE methods based on negative binomial distribution.

The differential gene expression (DGE) methods have been performed using pairwise design and edgeR, DESeq and baySeq. First, we have compared the numbers of accepted up and down-regulated genes using false discover rate (FDR) adjusted p-value of 0.05 (Figure 6). The difference among all the methods does not seem to follow a pattern. However, the data have shown a similar variation in some contrasts and methods. One strong trend is the massive down-regulation presented in the contrasts T474D_mock-WT_mock and T474D_mock-NIK_mock detected by edgeR TC normalization and by baySeq. Despite the DGE methods, which are based on the same distribution (negative binomial), large numerical variations in up and down-regulation of DE genes were observed. This unpredicted variation could be due to differences in the estimative of model parameters and initial assumption among the methods used .

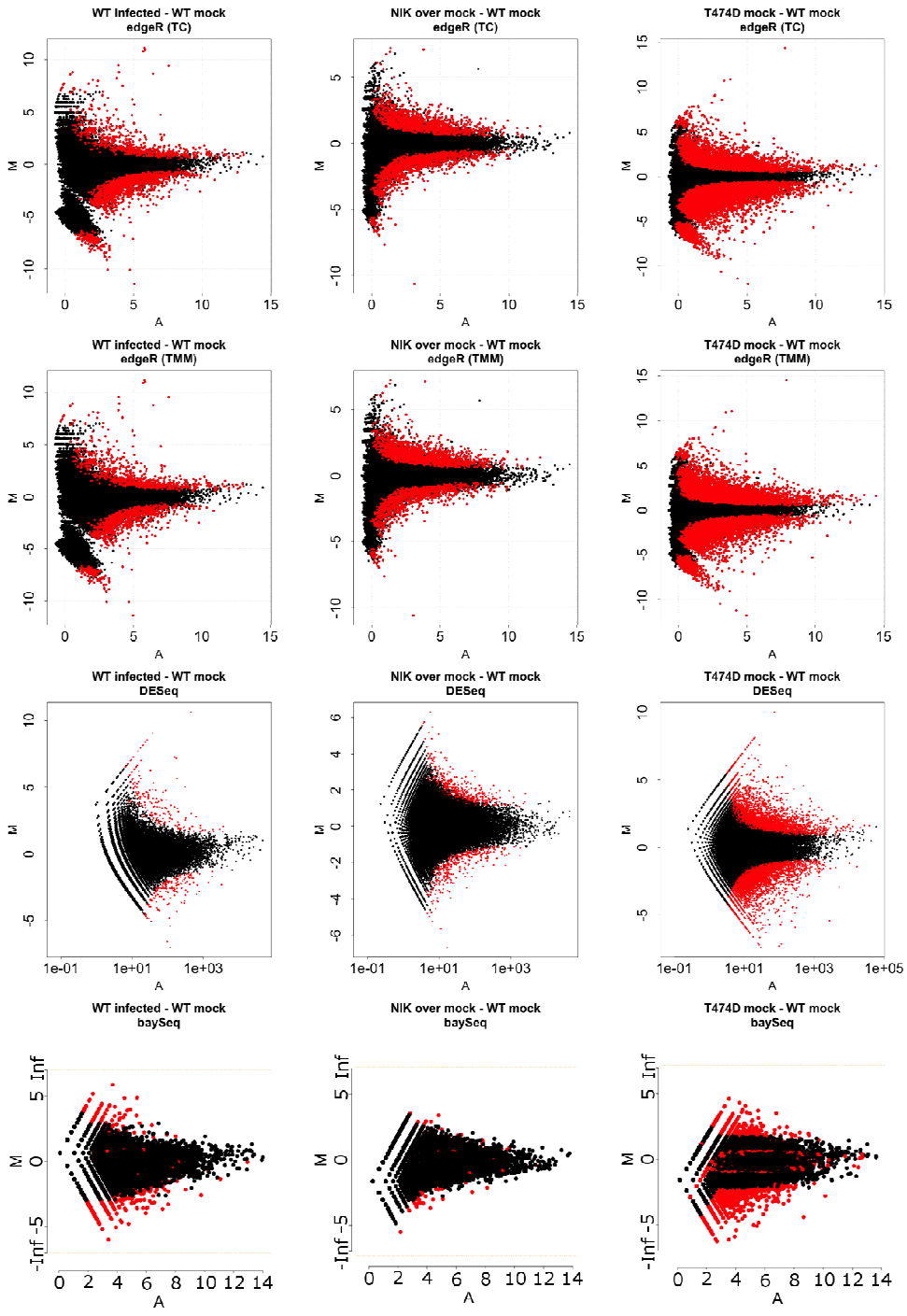



Figure 7: Each DGE data can be visualized as 'MA' plots (M = log ratio versus A = abundance) in which each dot represents a gene. This plot shows RNA-seq gene expression for WT infected versus WT mock, NIK over-expressed mock versus WT mock and T474D mock versus WT mock. Red dots represent the DE genes accepted as statistically significant.

To understand about the gene expression variation of the infected WT and the mock-inoculated overexpressed NIK and T474D tomato plants, we

have analyzed the following contrasts: infected WT – mock WT, mock NIK-OX – mock WT and mock T474D – mock WT. The MA plots show the genes dispersion. The DE genes are in red (Figure 7). The edgeR/TMM and edgeR/TC plots were similar in the overall data dispersion between infected WT – mock WT and mock T474D – mock WT, mainly in the down-regulation region. Although the DGE detection was numerically dissimilar (Figure 6), the similarity in the DGE profile of the infected WT – mock WT and mock T474D – mock WT MA plots reinforces that some infection response mechanisms linked to the down-regulated genes were already active in mock T474D plants.

The huge difference observed among the DGE methods made difficult to choose a unique normalization and DGE method for further analysis of the data, which would ensure that all real biological effects were uncovered. Thereby, it was reasonable to assume that the application of all gene expression methods would be the right choice to perform the further downstream analysis.

The DE genes from all normalizations and DEG analysis were stored using SQL tables at the PostgreSQL relational database (http://bioinfo-1.bioagro.ufv.br/fonteslab/tomatodb), which listed the corresponding $\log_2$(Fold Change), uncorrected p-value and p-value corrected by FDR (q-value) for all DE genes.

With the data from DE genes, there would be necessary to merge them with the annotations provided by the ITAG/Phytozome. Using the same SQL table structure presented at PostgreSQL relational database, the DE genes were easily merged with the gene annotations by the in house SQL scripts developed for this study (Supplementary information). In order to have a better understand of the biological phenomena, we performed a gene enrichment analysis using the GSEA methods based on biological process from the GO data. The same contrasts, infected WT – mock WT, mock NIK-OX – mock WT and mock T474D – mock WT, presented in DGE analysis were used here. A great number of enriched categories were found using p-value cutoff < 0.05 (Table S1, http://bioinfo-1.bioagro.ufv.br/fonteslab/tomatodb) . Because of the poor annotation of the tomato genome regarding some GO categories, we have decided to change the p-value cutoff to <0.01 and only accepted the GO categories, which had been labeled by at least four DGE methods or normalization factors (Figure 8, Table 1).

Unfortunately, a strong bias has arisen from some poor annotated GO

categories in the enrichment analysis, which led to the recognition of some statistically significant groups in spite of their very low number of genes. In fact, we have observed that some accepted GO categories had few significative genes (p-value < 0.01, Table 1). Thus, we did not consider those groups with less than 3 genes.

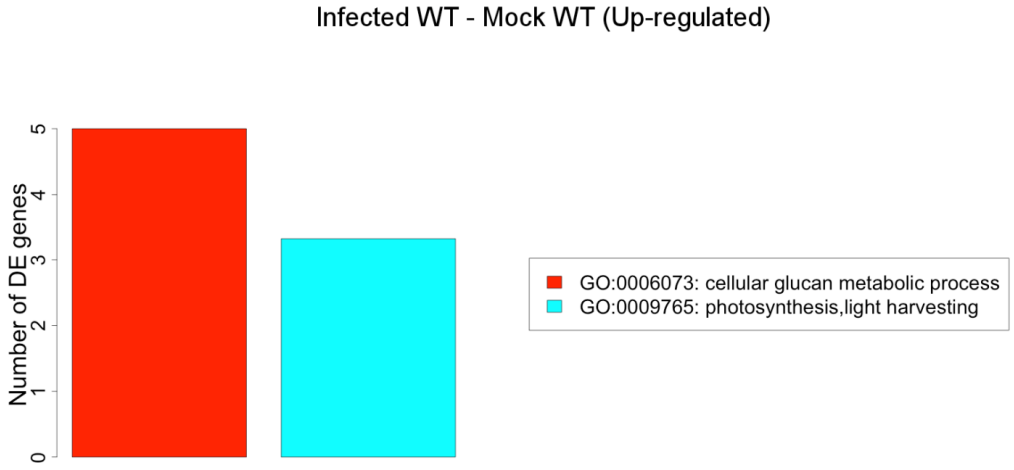
According to these criterion, the contrast mock NIK-OX – mock WT had a few GO enriched categories, one category for the up-regulated (GO:0046148: pigment biosynthetic process) and three for the down-regulated DE genes (GO:0009611: response to wounding, GO:0043648: dicarboxylic acid metabolic process and GO:0044262: cellular carbohydrate metabolic process). The enriched categories for the contrasts infected WT – mock WT and T474D mock – WT mock are shown in Figure 8.

The contrast mock T474D – mock WT down-regulated genes shows the highest number of enriched GO categories. The most significative p-value for one enriched category among all the analyzed contrasts was the GO:0006412 (translation) with the average p-value 4.69e-07. To determine whether a down-regulation of the translation machinery-related genes would be a direct result of the T474D expression, we plotted all the annotated genes attributed by the GO:0006412 (translation) in a smear-plot as the red dots (Fig. 9A). In the T474D mock-WT mock, the translation-related genes (red dots) tended to be down-regulated, because they were clearly concentrated at the bottom of the graphic. In contrast, this trend was not shared by the down-regulated GO:0006629: lipid metabolic process enriched category in the mock T474D – mock WT constrast which displayed a high dispersion without any tendency for up- and down-regulated profile (Figure 9B). These results indicate that merging of DGE data may have underestimated the number of translation-related genes in the T474Dmock-WTmock.

Another enriched group that could be highlighted is the GO:0009607 (response to biotic stimulus) from the up-regulated genes in the mock T474D – mock WT constrast. Despite the proximity of the DGE p-values to the cutoff, this group indicated that some pathogen-related genes were up-regulated. As GO does not have a category, which embraces the plant resistance genes, we have performed a search in the annotation database by the keywords: “resistance + disease” and “pathogenisis-related”. The outcome was 416 genes, from which 34 was found to be differentially expressed by at least three methods. Using the

hypergeometric distribution test, the enriched p-value was 0.0093287 (Table 2).

As a complement for the GO analysis, we used the KEGG (Kanehisa 2013) pathways enrichment to detect pathways not discriminated by GO analysis. We found that some KEGG-enriched categories confirmed the GO analysis, although the KEGG analysis resulted in fewer enriched categories. The GO categories of translation in down-regulated genes and biotic stress in the up-regulated genes from the contrast mock T474D – mock WT were also observed in the KEGG pathway enrichment by the 'ribosome' and 'Ribosome biogenesis in eukaryotes' (Translation) and 'Plant-pathogen interaction' (resistance) maps (Table 3). Collectively, these results indicate that overexpression of the gain-of-function mutant promotes a massive down-regualtion of the translational machinery genes and induces the immune system-related genes.



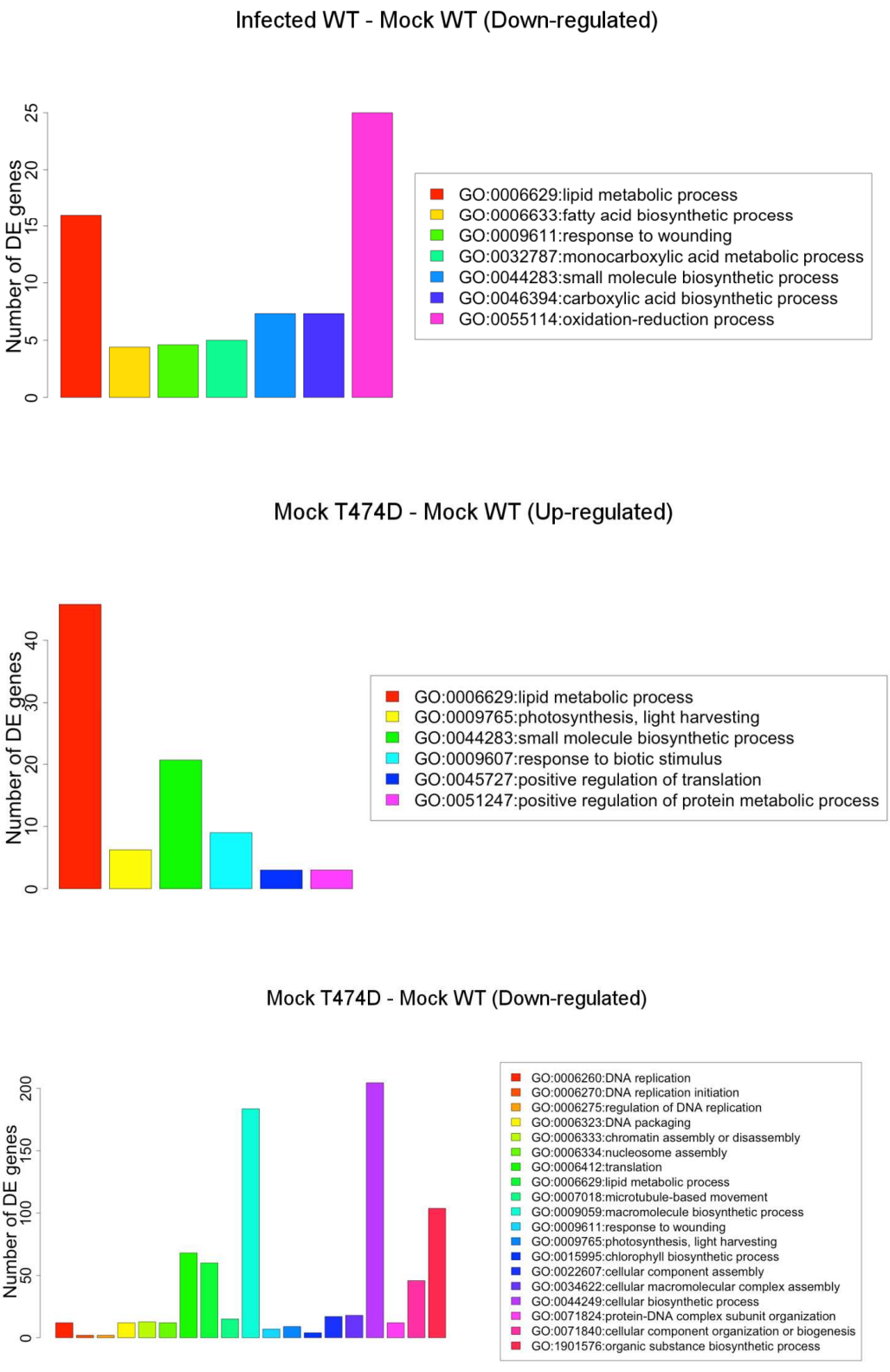



Figure 8: Bar charts of the enriched categories from Biological Process from gene ontology (GO) database using at least four DGE methods or normalization factors.

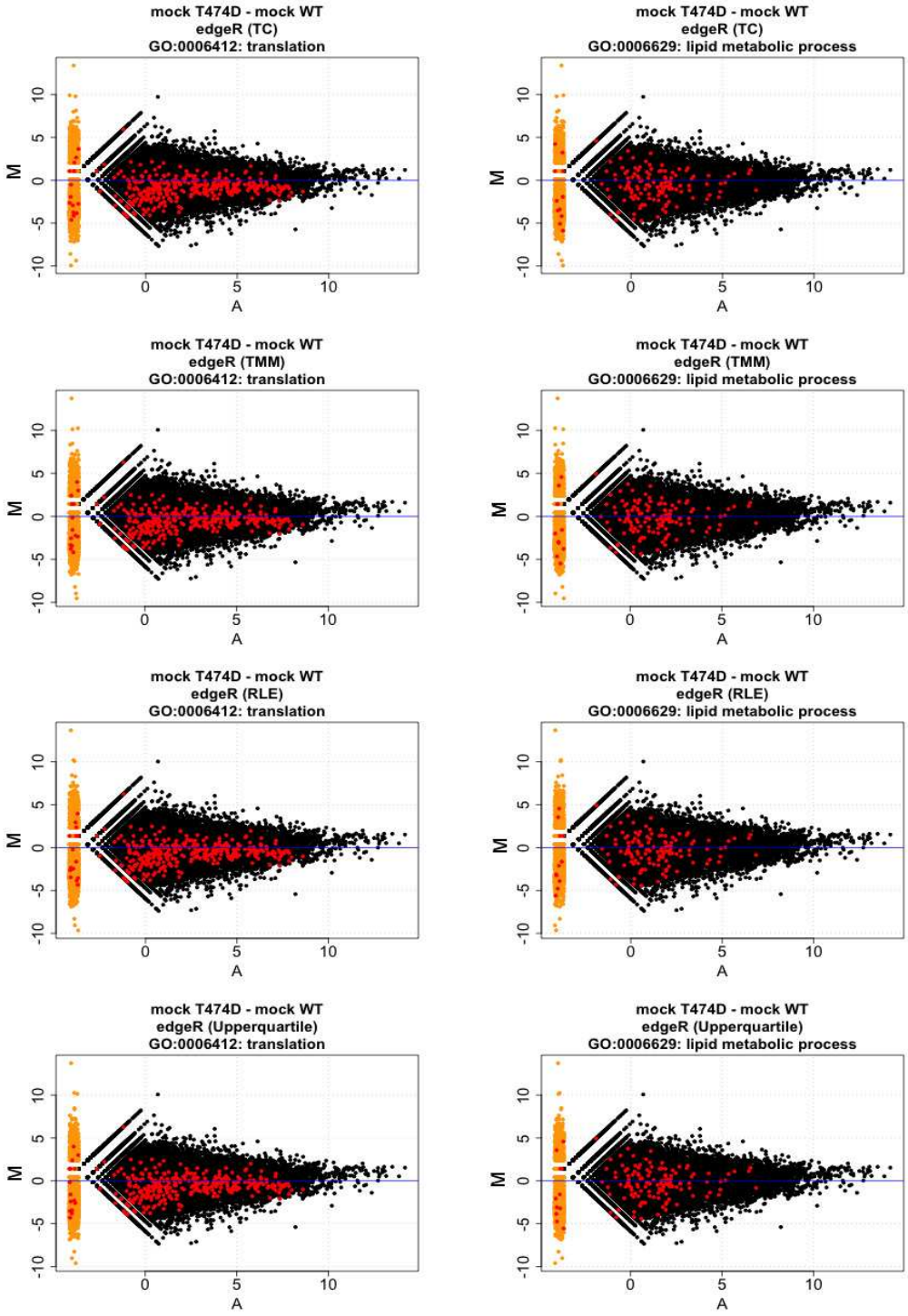



Figure 9: SmearPlots of the comparison of the dispersion between GO:0006412 (translation) and GO:0006629. (A) The density of the red dots are concentrated at the bottom of the graphics presented in the first column, while (B) in the second column the red dots are sparse to the overall dispersion. The smear of yellow points on the left side signifies that genes were observed in only one group of replicate samples.

Table 3: Enriched Biological process categories from GO database using GSEA method

| **Infected WT – mock WT (UP-regulated)** | | |
|---|---|---|
| ***Enriched Categories*** | ***DGE methods: nunber of genes / number of genes in the GO group (p-value)*** | ***AN*** |
| GO:0006073: cellular glucan metabolic process | edgeR/RLE: 4/93 (0.007615) edgeR/upper: 4/93 (0.008759) baySeq: 7/93 (0.006532) | 5 |
| GO:0009765: photosynthesis, light harvesting | edgeR/RLE: 3/33 (0.002581) edgeR/upper: 3/33 (0.002895) baySeq: 4/33 (0.007371) | 3 |
| GO:0030418: nicotianamine biosynthetic process | edgeR/TMM: 1/1 (0.006295) edgeR/TC: 1/1 (0.00526) edgeR/RLE: 1/1 (0.008365) edgeR/upper: 1/1 (0.00871) DESeq: 1/1 (0.001897) | 1 |
| | | |
| **Infected WT – mock WT (DOWN-regulated)** | | |
| ***Enriched Categories*** | ***DGE methods: nunber of genes / number of genes in the GO group (p-value)*** | ***AN*** |
| GO:0006629: lipid metabolic process | edgeR/TMM: 19/592 (1.79e-05) edgeR/TC: 22/592 (0.0004238) edgeR/RLE: 16/592 (8.847e-05) edgeR/upper: 7/343 (0.009761) | 16 |
| GO:0006633: fatty acid biosynthetic process | edgeR/TMM: 5/69 (0.00087) edgeR/TC: 5/69 (0.006144) edgeR/RLE: 5/69 (0.0004101) edgeR/upper: 5/69 (0.0001461) DESeq: 2/69 (0.006187) | 5 |
| GO:0009611: response to wounding | edgeR/TMM: 5/22 (3.054e-06) edgeR/TC: 5/22 (2.732e-05) edgeR/RLE: 5/22 (1.351e-06) edgeR/upper: 5/22 (4.486e-07) DESeq: 3/22 (6.617e-06) | 5 |
| GO:0032787: monocarboxylic acid metabolic process | edgeR/TMM: 5/96 (0.003758) edgeR/RLE: 5/96 (0.001836) edgeR/upper: 5/96 (0.000681) | 5 |
| GO:0044283: small molecule biosynthetic process | edgeR/TMM: 8/217 (0.002388) edgeR/RLE: 7/217 (0.003716) edgeR/upper: 7/217 (0.001046) | 7 |
| GO:0046394: carboxylic acid biosynthetic process | edgeR/TMM: 8/189 (0.0009954) edgeR/RLE: 7/189 (0.001713) edgeR/upper: 7/189 (0.000462) | 7 |
| GO:0055114: oxidation-reduction process | edgeR/TMM: 29/1645 (0.004426) edgeR/RLE: 25/1645 (0.00631) edgeR/upper: 21/1645 (0.006762) | 25 |
| | | |
| **Mock NIK-OX – mock WT (UP-regulated)** | | |
| ***Enriched Categories*** | ***DGE methods: nunber of genes / number of genes in the GO group (p-value)*** | ***AN*** |
| GO:0046148: pigment biosynthetic process | edgeR/TMM: 3/28 (0.006671) edgeR/TC: 3/28 (0.004615) edgeR/RLE: 3/28 (0.0075) edgeR/upper: 3/28 (0.003001) | 3 |
| | | |
| **Mock NIK-OX – mock WT (DOWN-regulated)** | | |
| ***Enriched Categories*** | ***DGE methods: nunber of genes / number of genes in the GO group (p-value)*** | ***AN*** |
| GO:0009611: response to wounding | edgeR/TMM: 3/22 (0.001621) edgeR/TC: 3/22 (0.003048) edgeR/RLE: 3/22 (0.001937) edgeR/upper: 3/22 (0.002832) | 3 |
| GO:0043648: dicarboxylic acid metabolic process | edgeR/TMM: 3/27 (0.002961) edgeR/TC: 3/27 (0.005512) edgeR/RLE: 3/27 (0.003528) edgeR/upper: 3/27 (0.005128) | 3 |
| GO:0044262: cellular carbohydrate metabolic process | edgeR/TMM: 7/214 (0.00841) edgeR/TC: 8/214 (0.00833) edgeR/upper: 8/214 (0.007165) | 8 |
| | | |
| **Mock T474D – mock WT (UP-regulated)** | | |
| ***Enriched Categories*** | ***DGE methods: nunber of genes / number of genes in the GO group (p-value)*** | ***AN*** |

| | | |
|---|---|---|
| GO:0006629: lipid metabolic process | edgeR/TMM: 47/592 (0.007967) edgeR/RLE: 46/592 (0.003633) edgeR/upper: 45/592 (0.007517) baySeq: 45/343 (5.446e-06) | 46 |
| GO:0009765: photosynthesis, light harvesting | edgeR/TMM: 6/33 (0.008713) edgeR/upper: 6/33 (0.006678) DESeq: 5/33 (0.005769) baySeq: 8/33 (0.001161) | 7 |
| GO:0009607: response to biotic stimulus | edgeR/TMM: 9/58 (0.009423) edgeR/upper: 8/58 (0.007652) baySeq: 10/58 (0.004553) | 9 |
| GO:0044283: small molecule biosynthetic process | edgeR/TMM: 21/217 (0.009161) edgeR/RLE: 20/217 (0.008819) edgeR/upper: 21/217 (0.004946) | 21 |
| GO:0051247: positive regulation of protein metabolic process | edgeR/TMM: 3/7 (0.00502) edgeR/RLE: 3/7 (0.004125) edgeR/upper: 3/7 (0.004283) | 3 |
| | | |
| **Mock T474D – mock WT (DOWN-regulated)** | | |
| ***Enriched Categories*** | ***DGE methods: nunber of genes / number of genes in the GO group (p-value)*** | ***AN*** |
| GO:0006260: DNA replication | edgeR/TMM: 12/87 (0.001556) edgeR/RLE: 12/87 (0.003899) edgeR/upper: 12/87 (0.00235) | 12 |
| GO:0006270: DNA replication initiation | edgeR/TMM: 2/2 (0.002646) edgeR/TC: 2/2 (0.007579) edgeR/RLE: 2/2 (0.003304) edgeR/upper: 2/2 (0.002919) | 2 |
| GO:0006275: regulation of DNA replication | edgeR/TMM: 2/2 (0.002646) edgeR/TC: 2/2 (0.007579) edgeR/RLE: 2/2 (0.003304) edgeR/upper: 2/2 (0.002919) DESeq: 2/2 (0.00171) | 2 |
| GO:0006323: DNA packaging | edgeR/TMM: 12/77 (0.0005285) edgeR/RLE: 12/77 (0.001394) edgeR/upper: 12/77 (0.0008159) | 12 |
| GO:0006333: chromatin assembly or disassembly | edgeR/TMM: 12/80 (0.0007529) edgeR/TC: 15/80 (0.003447) edgeR/RLE: 12/80 (0.001953) edgeR/upper: 12/80 (0.001154) | 13 |
| GO:0006334: nucleosome assembly | edgeR/TMM: 12/77 (0.0005285) edgeR/RLE: 12/77 (0.001394) edgeR/upper: 12/77 (0.0008159) | 12 |
| GO:0006412: translation | edgeR/TMM: 53/579 (2.923e-05) edgeR/TC: 110/579 (1.131e-15) edgeR/RLE: 67/579 (2.465e-08) edgeR/upper: 59/579 (1.62e-06) DESeq: 51/579 (2.342e-07) | 68 |
| GO:0006629: lipid metabolic process | edgeR/TMM: 56/592 (6.552e-06) edgeR/TC: 82/592 (1.355e-05) edgeR/RLE: 61/592 (5.595e-06) edgeR/upper: 58/592 (6.822e-06) DESeq: 43/592 (0.0002239) | 60 |
| GO:0007018: microtubule-based movement | edgeR/TMM: 14/60 (1.519e-06) edgeR/TC: 19/60 (3.965e-07) edgeR/RLE: 16/60 (1.667e-07) edgeR/upper: 15/60 (4.582e-07) DESeq: 11/60 (2.969e-05) | 15 |
| GO:0009059: macromolecule biosynthetic process | edgeR/TC: 250/2200 (1.178e-06) edgeR/RLE: 156/2200 (0.00198) edgeR/upper: 145/2200 (0.004376) | 184 |
| GO:0009611: response to wounding | edgeR/TMM: 7/22 (8.009e-05) edgeR/TC: 7/22 (0.001975) edgeR/RLE: 7/22 (0.0001609) edgeR/upper: 7/22 (0.0001092) DESeq: 7/22 (1.977e-05) | 7 |
| GO:0009765: photosynthesis, light harvesting | edgeR/TMM: 9/33 (3.03e-05) edgeR/TC: 9/33 (0.001575) edgeR/RLE: 9/33 (7.225e-05) edgeR/upper: 9/33 (4.459e-05) DESeq: 9/33 (5.25e-06) | 9 |
| GO:0015995: chlorophyll biosynthetic process | edgeR/TMM: 4/13 (0.003425) edgeR/RLE: 4/13 (0.005109) edgeR/upper: 4/13 (0.00409) | 4 |
| GO:0022607: cellular component assembly | edgeR/TMM: 17/153 (0.002249) edgeR/RLE: 17/153 (0.006862) edgeR/upper: 17/153 (0.003725) | 17 |
| GO:0034622: cellular macromolecular complex assembly | edgeR/TMM: 17/131 (0.0003895) edgeR/TC: 21/131 (0.004471) edgeR/RLE: 17/131 (0.001345) edgeR/upper: 17/131 (0.0006802) | 18 |
| GO:0044249: cellular biosynthetic process | edgeR/TMM: 163/2634 (0.004014) edgeR/TC: 296/2634 (2.09e-07) edgeR/RLE: 187/2634 (0.0005531) | 205 |

| | | |
|---|---|---|
| | edgeR/upper: 172/2634 (0.002566) | |
| GO:0071824: protein-DNA complex subunit organization | edgeR/TMM: 12/77 (0.0005285) edgeR/RLE: 12/77 (0.001394) edgeR/upper: 12/77 (0.0008159) | 12 |
| GO:0071840: cellular component organization or biogenesis | edgeR/TMM: 39/488 (0.004133) edgeR/TC: 62/488 (0.001511) edgeR/RLE: 42/488 (0.005659) edgeR/upper: 40/488 (0.005529) | 46 |
| GO:1901576: organic substance biosynthetic process | edgeR/TMM: 173/2717 (0.0007624) edgeR/TC: 60/517 (0.002539) edgeR/RLE: 43/517 (0.003325) DESeq: 139/2717 (0.002492) | 104 |

**AN – Average number*

Table 4: Gene enrichment by pathways from KEGG using hypergeometric test with threshold 0.01

| **Infected WT – mock WT (UP-regulated)** | | |
|---|---|---|
| ***Enriched pathway*** | ***DGE methods: number of DE genes/number in pathway (p-value)*** | ***AN*** |
| 00040: Pentose and glucuronate interconversions | baySeq: 5/59 (0.005185) | 5 |
| 00053: Ascorbate and aldarate metabolism | baySeq: 4/34 (0.003773) | 4 |
| 00071: Fatty acid metabolism | baySeq: 4/41 (0.007426) | 4 |
| 00196: Photosynthesis - antenna proteins | egdeR/RLE: 3/22 (0.0004045) egdeR/TC: 2/22 (0.003636) egdeR/TMM: 2/22 (0.004839) egdeR/RLE: 3/22 (0.0004045) egdeR/upper: 3/22 (0.000486) | 3 |
| 00310: Lysine degradation | baySeq: 3/15 (0.0026) | 3 |
| 00340: Histidine metabolism | baySeq: 4/19 (0.0003946) | 4 |
| 00520: Amino sugar and nucleotide sugar metabolism | egdeR/TC: 3/99 (0.007877) | 3 |
| 00561: Glycerolipid metabolism | baySeq: 4/35 (0.004197) | 4 |
| 00710: Carbon fixation in photosynthetic organisms | egdeR/TMM: 3/77 (0.005895) | 3 |
| 00941: Flavonoid biosynthesis | baySeq: 3/19 (0.005235) | 3 |
| 04075: Plant hormone signal transduction | baySeq: 11/232 (0.004992) | 11 |
| 04626: Plant-pathogen interaction | egdeR/TMM: 4/148 (0.005468) | 4 |
| | | |
| **Infected WT – mock WT (DOWN-regulated)** | | |
| ***Enriched pathway*** | ***DGE methods: number of genes (p-value)*** | ***AN*** |
| 00941: Flavonoid biosynthesis | egdeR/TC: 3/19 (0.001939) egdeR/TMM: 3/19 (0.000637) egdeR/RLE: 3/19 (0.0003642) egdeR/upper: 3/19 (0.000198) | 3 |
| 03010: Ribosome | egdeR/TC: 10/220 (0.0007847) | 10 |
| | | |
| **Mock NIK-OX – mock WT (UP-regulated)** | | |
| ***Enriched pathway*** | ***DGE methods: number of genes (p-value)*** | ***AN*** |
| 00010: Glycolysis / Gluconeogenesis | baySeq: 5/105 (0.009524) | 5 |
| 00030: Pentose phosphate pathway | egdeR/RLE: 4/53 (0.006231) | 4 |
| 00053: Ascorbate and | baySeq: 3/34 (0.00822) | 3 |

| | | |
|---|---|---|
| aldarate metabolism | | |
| 00100: Steroid biosynthesis | egdeR/TMM: 3/30 (0.007618) | 3 |
| 00520: Amino sugar and nucleotide sugar metabolism | egdeR/TC: 6/99 (0.001176) egdeR/TMM: 6/99 (0.002302) egdeR/RLE: 7/99 (0.000464) egdeR/upper: 6/99 (0.000406) | 6 |
| 01100: Metabolic pathways | egdeR/TC: 29/1489 (0.006205) egdeR/TMM: 32/1489 (0.006955) egdeR/RLE: 35/1489 (0.001701) egdeR/upper: 25/1489 (0.005301) DESeq: 11/1489 (0.005527) baySeq: 32/1489 (0.001521) | 27 |
| 01110: Biosynthesis of secondary metabolites | baySeq: 18/790 (0.009103) | 18 |
| 03010: Ribosome | egdeR/RLE: 9/220 (0.003758) | 9 |
| 03015: mRNA surveillance pathway | egdeR/TC: 5/82 (0.002935) egdeR/TMM: 6/82 (0.0008711) egdeR/RLE: 5/82 (0.005683) | 5 |
| 03040: Spliceosome | egdeR/TMM: 7/115 (0.0009872) egdeR/RLE: 6/115 (0.005408) egdeR/upper: 6/115 (0.0008969) DESeq: 3/115 (0.005296) | 6 |
| 04626: Plant-pathogen interaction | baySeq: 6/148 (0.009959) | 6 |
| 04712: Circadian rhythm - plant | baySeq: 3/29 (0.00525) | 3 |
| | | |
| **Mock NIK-OX – mock WT (DOWN-regulated)** | | |
| ***Enriched pathway*** | ***DGE methods: number of genes (p-value)*** | ***AN*** |
| 00260: Glycine, serine and threonine metabolism | baySeq: 1/47 (0.008409) | 1 |
| 00270: Cysteine and methionine metabolism | egdeR/RLE: 4/84 (0.008527) | 4 |
| 00310: Lysine degradation | egdeR/TMM: 2/15 (0.007741) egdeR/RLE: 2/15 (0.008726) | 2 |
| 00330: Arginine and proline metabolism | egdeR/TC: 4/68 (0.007011) egdeR/RLE: 4/68 (0.004049) egdeR/upper: 4/68 (0.007491) | 4 |
| 00400: Phenylalanine, tyrosine and tryptophan biosynthesis | baySeq: 1/43 (0.007696) | 1 |
| 00460: Cyanoamino acid metabolism | egdeR/upper: 3/28 (0.003841) | 3 |
| 00600: Sphingolipid metabolism | egdeR/TMM: 2/14 (0.006748) egdeR/RLE: 2/14 (0.00761) | 2 |
| 01040: Biosynthesis of unsaturated fatty acids | baySeq: 1/32 (0.005732) | 1 |
| 01100: Metabolic pathways | egdeR/TC: 30/1489 (0.001305) egdeR/TMM: 25/1489 (0.001923) egdeR/RLE: 26/1489 (0.002184) egdeR/upper: 31/1489 (0.0008773) | 28 |
| 01110: Biosynthesis of secondary metabolites | egdeR/TC: 20/790 (0.0006046) egdeR/TMM: 18/790 (0.0002926) egdeR/RLE: 17/790 (0.001598) egdeR/upper: 21/790 (0.0003018) | 19 |
| 04145: Phagosome | egdeR/TC: 4/62 (0.005058) egdeR/TMM: 4/62 (0.002317) egdeR/upper: 4/62 (0.00541) | 4 |
| | | |
| **Mock T474D – mock WT (UP-regulated)** | | |
| ***Enriched pathway*** | ***DGE methods: number of genes (p-value)*** | ***AN*** |
| 00010: Glycolysis / Gluconeogenesis | egdeR/TC: 12/105 (7.832e-05) egdeR/TMM: 15/105 (0.0001233) egdeR/RLE: 15/105 (5.155e-05) egdeR/upper: 13/105 (0.0007389) DESeq: 13/105 (1.035e-05) | 14 |
| 00030: Pentose phosphate pathway | egdeR/TMM: 8/53 (0.003243) egdeR/RLE: 8/53 (0.002) egdeR/upper: 8/53 (0.002149) | 8 |
| 00040: Pentose and glucuronate | baySeq: 10/59 (0.001445) | 10 |

| | | |
|---|---|---|
| interconversions | | |
| 00071: Fatty acid metabolism | egdeR/TMM: 7/41 (0.002838) egdeR/RLE: 7/41 (0.001826) DESeq: 5/41 (0.006213) baySeq: 8/41 (0.001677) | 7 |
| 00100: Steroid biosynthesis | baySeq: 6/30 (0.005547) | 6 |
| 00196: Photosynthesis - antenna proteins | egdeR/TMM: 6/22 (0.0004269) egdeR/RLE: 5/22 (0.002233) egdeR/upper: 6/22 (0.0002983) DESeq: 5/22 (0.0003424) baySeq: 6/22 (0.001021) | 6 |
| 00310: Lysine degradation | baySeq: 4/15 (0.007989) | 4 |
| 00350: Tyrosine metabolism | egdeR/upper: 7/26 (0.0001016) egdeR/TMM: 9/26 (1.719e-06) egdeR/RLE: 9/26 (8.977e-07) egdeR/upper: 7/26 (0.0001016) DESeq: 4/26 (0.006158) | 7 |
| 00480: Glutathione metabolism | egdeR/TMM: 8/62 (0.008529) egdeR/RLE: 8/62 (0.005408) egdeR/upper: 8/62 (0.005786) | 8 |
| 00620: Pyruvate metabolism | DESeq: 7/75 (0.005825) baySeq: 10/75 (0.008549) | 9 |
| 00640: Propanoate metabolism | baySeq: 6/31 (0.00656) | 6 |
| 00710: Carbon fixation in photosynthetic organisms | egdeR/TC: 9/77 (0.0005101) egdeR/TMM: 11/77 (0.0009563) egdeR/RLE: 11/77 (0.000502) egdeR/upper: 11/77 (0.0005522) DESeq: 11/77 (1.247e-05) | 11 |
| 00950: Isoquinoline alkaloid biosynthesis | egdeR/upper: 4/9 (0.0004001) egdeR/TMM: 4/9 (0.0005163) egdeR/RLE: 4/9 (0.000383) egdeR/upper: 4/9 (0.0004001) | 4 |
| 01100: Metabolic pathways | egdeR/TC: 64/1489 (0.00289) egdeR/TMM: 115/1489 (1.355e-07) egdeR/RLE: 106/1489 (4.984e-07) egdeR/upper: 105/1489 (1.485e-06) DESeq: 75/1489 (1.933e-06) baySeq: 118/1489 (6.857e-05) | 97 |
| 01110: Biosynthesis of secondary metabolites | egdeR/TC: 37/790 (0.005873) egdeR/TMM: 67/790 (2.975e-06) egdeR/RLE: 64/790 (1.741e-06) egdeR/upper: 62/790 (9.825e-06) DESeq: 42/790 (0.000117) baySeq: 63/790 (0.002806) | 56 |
| 03030: DNA replication | baySeq: 9/45 (0.0007247) | 9 |
| 03430: Mismatch repair | baySeq: 6/33 (0.008977) | 6 |
| | | |
| **Mock T474D – mock WT (DOWN-regulated)** | | |
| ***Enriched pathway*** | ***DGE methods: number of genes (p-value)*** | ***AN*** |
| 00040: Pentose and glucuronate interconversions | DESeq: 7/59 (0.00323) | 7 |
| 00061: Fatty acid biosynthesis | egdeR/TMM: 5/27 (0.005144) | 5 |
| 00072: Synthesis and degradation of ketone bodies | DESeq: 2/4 (0.006269) | 2 |
| 00100: Steroid biosynthesis | egdeR/RLE: 7/30 (0.000406) egdeR/TC: 9/30 (0.0002224) egdeR/TMM: 7/30 (0.0002156) egdeR/RLE: 7/30 (0.000406) egdeR/upper: 7/30 (0.0002802) DESeq: 6/30 (0.0003867) | 7 |
| 00196: Photosynthesis - antenna proteins | egdeR/TC: 6/22 (0.004306) egdeR/TMM: 6/22 (0.0002451) egdeR/RLE: 6/22 (0.0004286) egdeR/upper: 6/22 (0.0003089) DESeq: 6/22 (6.091e-05) | 6 |
| 00230: Purine metabolism | egdeR/TC: 20/133 (0.001844) | 20 |
| 00240: Pyrimidine metabolism | egdeR/TC: 17/97 (0.0007093) egdeR/TMM: 12/97 (0.0008578) egdeR/RLE: 13/97 (0.0006448) egdeR/upper: 12/97 (0.00124) DESeq: 11/97 (0.0003631) | 13 |

| | | |
|---|---|---|
| 00270: Cysteine and methionine metabolism | egdeR/TMM: 9/84 (0.009458) egdeR/RLE: 10/84 (0.006216) | 9.5 |
| 00330: Arginine and proline metabolism | egdeR/TMM: 9/68 (0.002311) egdeR/RLE: 10/68 (0.001289) egdeR/upper: 10/68 (0.000812) | 10 |
| 00640: Propanoate metabolism | egdeR/TMM: 5/31 (0.00941) DESeq: 5/31 (0.003251) | 5 |
| 00909: Sesquiterpenoid biosynthesis | baySeq: 1/1 (0.006498) | 1 |
| 00941: Flavonoid biosynthesis | egdeR/TC: 8/19 (3.171e-05) egdeR/TMM: 6/19 (9.945e-05) egdeR/RLE: 6/19 (0.0001761) egdeR/upper: 6/19 (0.000126) baySeq: 2/19 (0.006678) | 6 |
| 01040: Biosynthesis of unsaturated fatty acids | egdeR/TC: 7/32 (0.007786) | 7 |
| 01100: Metabolic pathways | egdeR/TC: 158/1489 (2.273e-06) egdeR/TMM: 107/1489 (8.155e-08) egdeR/RLE: 115/1489 (1.409e-07) egdeR/upper: 109/1489 (2.093e-07) DESeq: 80/1489 (1.399e-05) baySeq: 21/1489 (0.0007097) | 98 |
| 01110: Biosynthesis of secondary metabolites | egdeR/TC: 87/790 (0.0001255) egdeR/TMM: 63/790 (1.555e-06) egdeR/RLE: 67/790 (3.052e-06) egdeR/upper: 62/790 (1.197e-05) DESeq: 44/790 (0.0005858) baySeq: 14/790 (0.0006895) | 56 |
| 03008: Ribosome biogenesis in eukaryotes | egdeR/TC: 12/76 (0.009601) | 12 |
| 03010: Ribosome | egdeR/TC: 52/220 (4.015e-14) egdeR/TMM: 29/220 (7.252e-08) egdeR/RLE: 32/220 (1.608e-08) egdeR/upper: 31/220 (1.419e-08) DESeq: 28/220 (1.177e-09) baySeq: 6/220 (0.003243) | 30 |
| 03030: DNA replication | egdeR/TC: 12/45 (7.438e-05) egdeR/TMM: 12/45 (2.649e-07) egdeR/RLE: 12/45 (8.003e-07) egdeR/upper: 12/45 (4.188e-07) DESeq: 9/45 (1.381e-05) | 11 |
| 03420: Nucleotide excision repair | egdeR/TMM: 8/59 (0.00342) egdeR/RLE: 8/59 (0.006359) egdeR/upper: 8/59 (0.004429) | 8 |
| 03430: Mismatch repair | egdeR/TC: 11/33 (1.477e-05) egdeR/TMM: 10/33 (7.184e-07) egdeR/RLE: 10/33 (1.843e-06) egdeR/upper: 10/33 (1.061e-06) DESeq: 7/33 (8.504e-05) | 10 |
| 04075: Plant hormone signal transduction | egdeR/TC: 28/232 (0.007199) egdeR/TMM: 20/232 (0.00233) egdeR/RLE: 21/232 (0.003478) egdeR/upper: 20/232 (0.003775) DESeq: 17/232 (0.001944) | 21 |
| 04144: Endocytosis | egdeR/TMM: 8/67 (0.007471) egdeR/upper: 8/67 (0.009552) | 8 |
| 04146: Peroxisome | egdeR/TC: 11/61 (0.004708) | 11 |

**AN – Average number*

### 3.4 - Constitutive activation of NIK impairs translation and confers broad-spectrum tolerance against begomoviruses

To confirm that protein synthesis was impaired by the constitutive activation of NIK in the T474D lines, we labeled leaf proteins *in vivo* with [$^{35}$S]Met in the control plants and T474D overexpression lines (Figure 10A).

There was a significant decrease (34% in T474D-5, 29.5% in T474D-6 and 27% in T474D-2; $P<0.05$) in the amount of newly synthesized protein in T474D-overexpressing leaves compared with the amounts found in wild-type and NIK-overexpressing leaves (Figure 10A). We observed a slight variation in the T474D-mediated inhibition of translation during development, as the level of translation suppression was about 7% less when incorporation of [$^{35}$S]Met into total proteins was measured in 28 days-old leaves (Figure 10B). This down-regulation of translation might underlie, at least in part, the molecular mechanisms involved in NIK-mediated antiviral defenses.

We next examined whether the constitutive activation of NIK was effective at controlling begomovirus infection. To this end, four independent T474D-overexpressing transgenic lines (T474D-9, T474D-6, T474D-5 and T47D-2), a wild-type (untransformed) line and the NIK-overexpressing lines NIK1-4 and NIK1-6 (Carvalho et al., 2008) were inoculated with tandemly repeated ToYSV DNA-A and DNA-B (Carvalho et al., 2008) using biolistic delivery, and the plants were assayed for symptoms of infection and the accumulation of viral DNA, as detected by PCR and qPCR. The wild-type plants displayed typical symptoms of ToYSV infection, such as leaf curling and yellow spots all over the leaves (>10 spots/cm$^2$; Figure 11A). Consistent with a previous observation (Caarvalho et al., 2008), the NIK-overexpressing line NIK1-4 displayed attenuated symptoms (less accentuated leaf distortion and a lower number of yellow spots per leaf area, <6 spots/cm$^2$). The symptoms in the T474D-overexpressing lines, however, were even more attenuated, with few spots per leaf area (varying among the lines) and no visible leaf curling (see T474D-2 and T474D-5 lines). The T474D-6 transgenic line displayed typical tolerance against begomoviruses, as it did not develop symptoms (Figure 11A and 11D), but we detected viral DNA accumulation in both inoculated and systemically infected leaves (Figure 11E). The symptomless ToYSV infections of the T474D-6 line were associated with a delayed course of infection (Figure 11F); a lower rate of infection (DPI50, days post inoculation to infect 50% of plants); and a lower accumulation of viral DNA in the systemically infected leaves, as shown by qPCR for the viral DNA (Figure 11H). We also observed a significant reduction in the polysome loading of viral mRNA (coat protein mRNA) in systemically infected leaves of the T474D-6-overexpresing line as compared to infected wild type and NIK1-overexpressing leaves (Figure 12A and 12B).

The accumulation of total virus transcripts in all infected lines was confirmed in our RNA sequencing data (Figure 12C).

Likewise, in the T474D-2 and T474D-5 lines, the progress and rate of infection were delayed compared with those of the wild-type control lines and the NIK1-overexpressing lines (Figure 10F and 10G) and the loading of coat protein mRNA in the actively translating polysomes was lower than that in wild-type and NIK1-overexpressing lines (Figure 12B). In the case of ToYSV, which showed high levels of accumulation in all samples analysed, both the T474D-2 and T474D-6 lines displayed lower viral DNA accumulation levels in the systemically infected leaves, although the high dispersion of the data among the samples prevented us from ascertaining the statistical significance of these findings (Figure 11H).

These transgenic lines were also challenged with the tomato-infecting begomovirus ToSRV (*Tomato severe rugose virus*), which caused severe leaf distortion in wild-type leaves, but not in the T474D-6 overexpressing line (Figure 13A). In these T474D-overexpressing lines, the viral DNA accumulation in the systemic leaves from ToSRV infections was significantly lower at 14 and 21 DPI ($P<0.05$; Figure 13B and 13C). The performance of the T474D-overexpressing lines upon begomovirus infection further confirmed that the T474D mutant protein could mount a sustained NIK-mediated antiviral defense in the absence of viral infection. The constitutive activation of T474D and its ability to bypass viral NSP inhibition likely account for the tolerance to begomovirus infection displayed by the T474D-6 line.

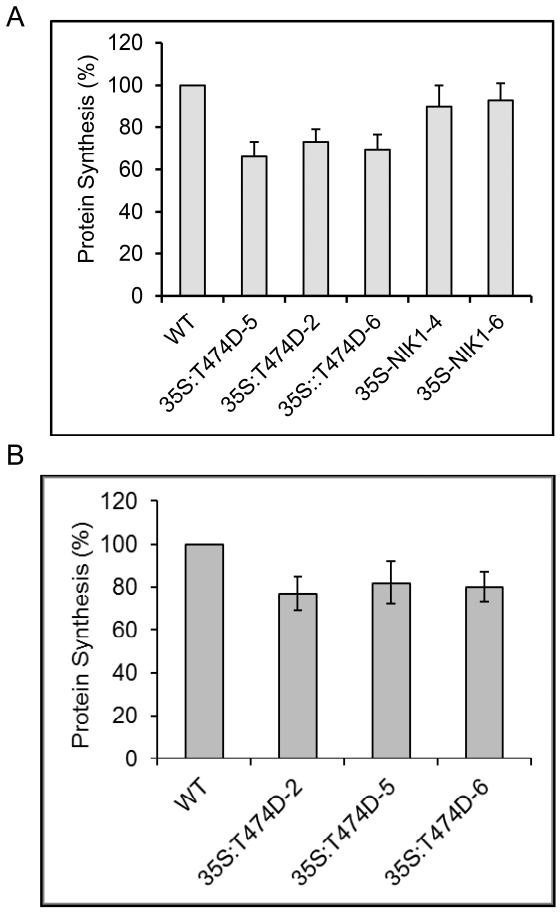



Figure 10. Ectopic expression of the T474 mutant receptor down-regulates global protein synthesis in leaves of 10 days (A)-old ad 28 days-old (B) tomato plants. Equal fresh weight of tomato leaves (300 mg) were incubated with 50 µg/ml chloramphenicol and 20 µCi of [35S]methionine for 3 h at room temperature. Incorporation of [$^{35}$S]Met into protein was measured in the TCA-precipitated total protein (mean ± SD, n= 3, $P<0.05$) from wild-type and T474D transgenic lines. Asterisks indicate means significantly different from the wild type control ($P<0.05$, Student's t test). In the transgenic 28 days-old leaves, the extent of translation inhibition by T474D overexpression was reduced (about 7%) as compared to the level of suppression in transgenic seedlings.

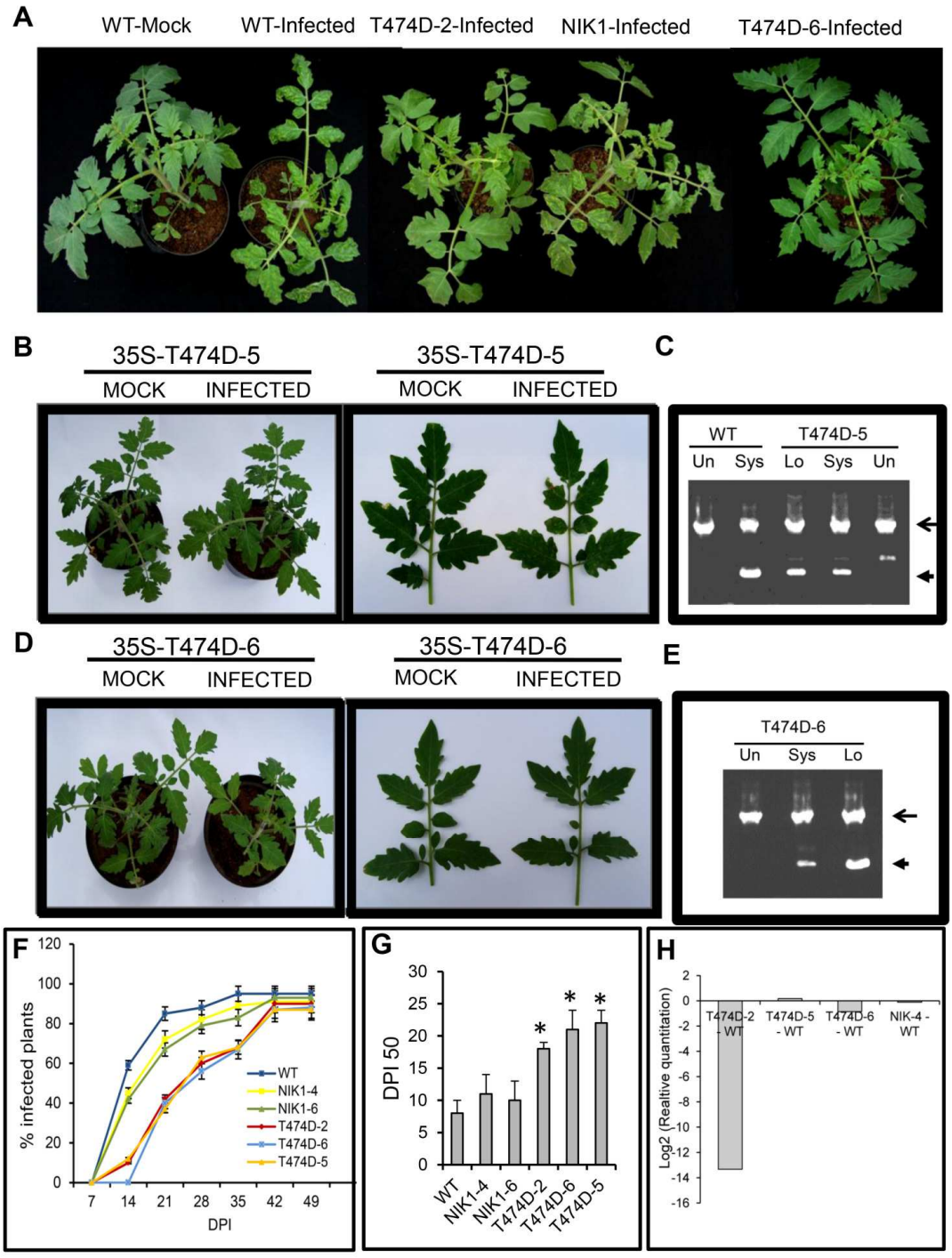
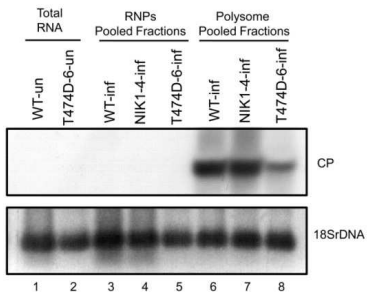


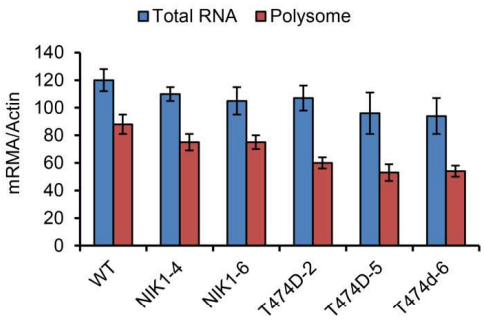

Figure 11. Ectopic expression of T474D in tomato confers tolerance to ToYSV infection. (A) Ectopic expression of T474D in tomato plants attenuates the development of symptoms upon ToYSV infection. Tandemly repeated ToYSV DNA-A and DNA-B sequences were introduced into the indicated lines by biolistic inoculation. Photographs were taken at 21 days post inoculation (DPI). (B) Symptoms associated with ToYSV infection in the 35S-T474D-5 line. Photographs were taken at 21 DPI. (C) Viral DNA accumulation in the infected leaves of the T474-5 line. Total DNA was isolated from infected plants at 21 DPI and PCR was performed with viral DNA-B-specific primers and actin-specific primers (as an internal control) in the same reaction. Lo indicates inoculated

and Sys denotes systemic leaves. “Un” indicates mock-inoculated leaves. The gel shows representative samples of Col-0 (WT) and 35S::T474D-5 transgenic line. The upper band (1.2 kb, arrow) is the amplified genomic fragment from actin and the lower band (0.5 kb, arrowhead) is the viral DNA fragment. (D) and (E) The OE line 35S-T474D-6 displayed tolerance to ToYSV infection. The T474D-6 line was infected with ToYSV, and photographs were taken at 21 DPI. Viral DNA accumulation was detected by PCR in inoculated (“Lo”) and systemic (“Sys”) leaves. (F) The course of infection was delayed in T474D lines. The indicated lines were infected with ToYSV by the biolistic method, and the course of infection was monitored by PCR amplification of viral DNA. Values represent the percentage of systemically infected plants at different DPI. (G) Infection efficiency in T474D-overexpressing lines. The infection efficiency is expressed as the DPI required to infect 50% of the plants (mean ± SD of three replicates). Asterisks indicate means significantly different ($P<0.05$, Student’s t test). (H) Viral DNA accumulation in T474D-overexpressing lines, as determined by quantitative PCR at 28 DPI. The fold variation (± SD, n = 3 biological replicates) is shown as log2-scaled copy units of the viral genome. Viral DNA accumulation was determined in systemic leaves at 21 DPI.

A




B




C

| WT infected – WT mock | | | |
|---|---|---|---|
| Tag | log2FC | Q-value | Viral Transcritps |
| CATGTGGTCTACAATCACCAG | 36.4214416953467 | 2.72242902865784e-21 | CP |
| CATGTTGTATTTAGACCGATT | 36.5706856869958 | 8.88820536256974e-23 | Rep/Trap/Ren |
| CATGCATCTAACCCTGTGTAT | 36.2836036544981 | 6.32696523733335e-21 | Coat protein |
| CATGTAACGGTGTTTATGCAA | 34.2901676257986 | 6.70065426546673e-08 | NSP |
| CATGTCAGAGGCCCAAATAAA | 34.9384840911855 | 1.0731864468325e-11 | MP |
| | | | |
| T474D infected – WT mock | | | |
| Tag | log2FC | Q-value | Viral Transcritps |
| CATGTGGTCTACAATCACCAG | 37.8173855440972 | 2.43394976695905e-39 | CP |
| CATGTTGTATTTAGACCGATT | 37.6871003260982 | 2.43394976695905e-39 | Rep/Trap/Ren |
| CATGCATCTAACCCTGTGTAT | 37.7116639425534 | 5.11589391144195e-40 | Coat protein |
| CATGTAACGGTGTTTATGCAA | 35.9742546966172 | 1.06764307374209e-24 | NSP |
| CATGTCAGAGGCCCAAATAAA | 35.724639031332 | 1.2249941122528e-22 | MP |
| | | | |
| NIK infected – WT mock | | | |
| Tag | log2FC | Q-value | Viral Transcritps |
| CATGTGGTCTACAATCACCAG | 35.729213273961 | 1.94295586580166e-22 | CP |
| CATGTTGTATTTAGACCGATT | 35.7464814641608 | 3.73443700707988e-23 | Rep/Trap/Ren |
| CATGCATCTAACCCTGTGTAT | 35.4402538329626 | 5.70874754573028e-21 | Coat protein |
| CATGTAACGGTGTTTATGCAA | 33.597222781461 | 1.80507146753756e-09 | NSP |
| CATGTCAGAGGCCCAAATAAA | 33.8551626703184 | 6.08744380555846e-11 | MP |

Figure 12. Polysome loading of viral mRNA is reduced in in systemically infected leaves of T47D-6-overexpressing lines. (A) Polysome loading of coat protein (CP) mRNA from ToYSV DNA-A in systemically infected leaves of WT, NIK-overexpressing and T47D-6-overexpressing lines. Polysomes from infected WT, NIK1-4-overexpressing and T474D-6-overexpressing lines were isolated from systemic leaves at 10 days post-inoculation with tandemly copy of DNA-A and DNA-B of ToYSV, as shown in Fig. S7. Polysome-bound RNA form pooled fractions was extracted with phenol/chloroform/isoamyl alcohol, precipitated with isopropanol, blotted and probed with the coat protein DNA (CP) and 18S rDNA. The identity of the polysome pooled fraction was confirmed by treatment with 25 mM EDTA prior to sucrose gradient (data not shown), which releases mRNA from polysomes. (B) Quantitation of polysome loading of coat protein viral transcripts in T474D overexpressing lines by qRT-PCR. Polysomes from infected WT, NIK1-4-overexpressing and T474D-6-overexpressing lines were isolated 10 days post-inoculation with infectious ToYSV clones. Polysome-bound RNA form pooled fractions was extracted with phenol/chloroform/isoamyl alcohol and precipitated with isopropanol and quantified by qRT-PCR. Values were normalized to the expression of Actin. Error bars represent SD from three measurements. (C) Expression profile of viral gene transcripts in T474D overexpressing lines. RNA-sequencing data of viral gene transcripts in systemic leaves of WT, NIK1-overexpressing and T474D-overexpressing plants, 10 days post-inoculation with tandemly repeated copies of ToYSV DNA-A and DNA-B. CP is coat protein, Rep/Trap/Ren, corresponds to the transcript encoding replication protein (Rep), transactivator protein (Trap) and replication enhancer protein (Ren), MP is movement protein and NSP, nuclear shuttle protein.

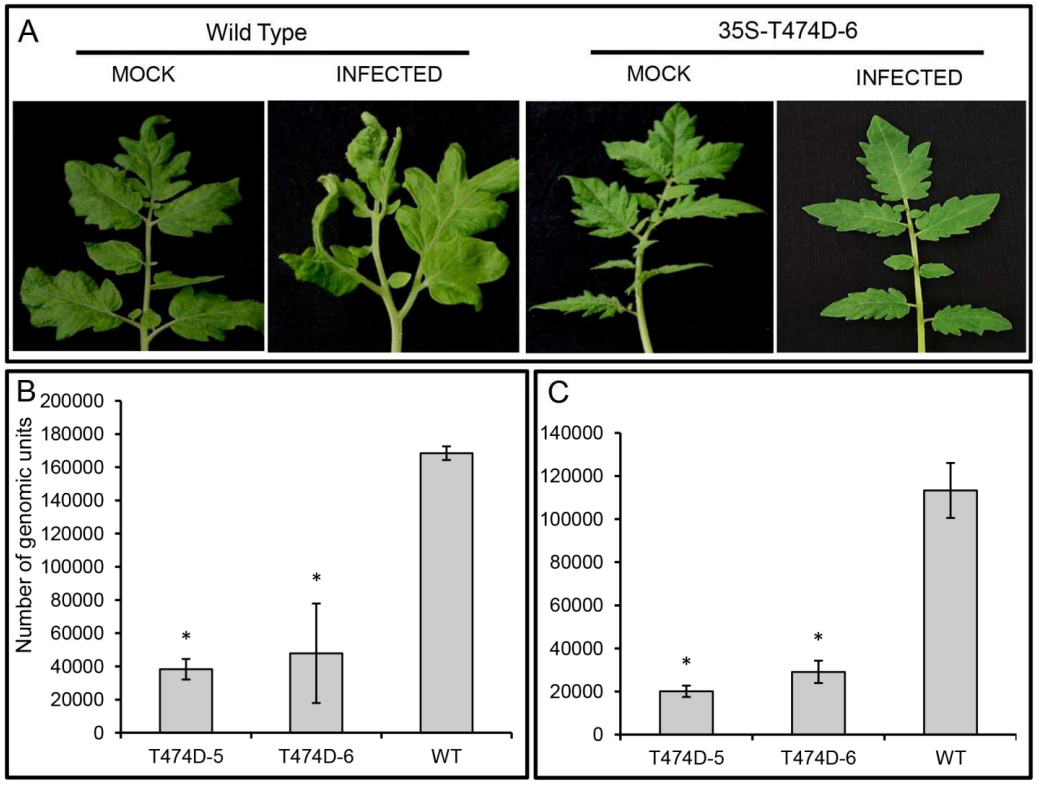



Figure 13.The T474D-6 overexpressing line is also tolerant to ToSRV infection.

(A) Symptoms associated with ToSRV infection in the 35S-T474D-6 line. The T474D-6 line was infected with ToSRV, and photographs were taken at 21 DPI. (B) and (C) Viral DNA accumulation in the T474D-6- and T474-5-overexpressing lines at 14 DPI (B) and 28 DPI (C). Prior to performing real-time PCR, infected leaves were diagnosed by standard PCR. Subsequently, total DNA extracted from systemically infected leaves at 14 DPI (B) or 28 DPI (C) was used as a template for quantitative PCR using ToYSV DNA-A-specific primers. The fold variation (± SD, n = 3 biological replicates) is shown as copy units of the viral genome. The asterisks indicate significant differences with $P < 0.05$ according to a Student's t-test.

## 4 – Discussion

Begomoviruses are one of the largest and most successfully groups of plant viruses and cause severe diseases in major crops worldwide, inflicting significant economic losses in many dicotyledonous crops. The tomato-infecting begomoviruses have become an even greater threat to the tomato cultivation due to the emergence of new species along with the recent introduction in South America of a new biotype of the whitefly vector *Bemisia tabaci* that colonizes tomato plants with high efficiency (Castillo-Urquiza et al., 2008; Albuquerque et al., 2012). Current climate changes are expected to alter more the whitefly distribution along the globe posing a serious threat to agriculture worldwide. Here we described a novel strategy to control begomovirus infection. By constitutively activating the NIK-mediated antiviral signaling we succeeded in developing a tolerant crop, tomato, which is inflicted by a diverse begomovirus complex, making engineered tolerance/ resistance an even more difficult task to accomplish. Very importantly, the T474D-overexpressing tomato transgenic lines were tolerant to ToYSV and ToSRV, which display highly divergent genomic sequences and hence are phylogenetically separated within the two major groups of begomoviruses found in Brazil (Albuquerque et al., 2012). These observations may indicate the potential of a sustained NIK-mediated defense pathway to confer broad-spectrum tolerance to begomoviruses in distinct plant species.

A comparison between the transcriptomes induced by virus infection in wild type lines and by ectopic expression of the T474D gain-of-function mutant in transgenic lines indicated that virus infection was the trigger of the NIK-mediated antiviral pathway and the T474D hyperactive receptor could support a sustained antiviral response. Therefore, upon perception of virus infection the plant cells may activate or synthesize an unknown molecule signal to trigger the NIK-mediated antiviral defense.

The T474D-overexpressing lines and infected WT share some similar up- and down-regulated changes that may account for the defense response of the NIK-mediated antiviral signaling pathway. This interpretation is supported by the observation that mock-inoculated T474D-overexpressing lines showed a constitutive infected wild type transcriptome (Figure 4A). We provided several other lines of evidence that support this interpretation. Firstly, we showed that

three independently tomato displayed identical gene expression profiles induced by ectopic expression of T474D mutant receptor. Secondly, while 90% of the ribosomal protein genes were down-regulated; several resistant protein-like genes and pathogenesis-related genes were up-regulated in T474D transgenic lines. These results indicate that the expression of ribosomal genes and immune system-related genes were coordinately reprogrammed in the transgenic lines. This coordinate regulation of subnetworks of gene sets, which are involved in the same cellular processes (translation and cell defense), is likely a result from activation of a master regulator (in this case, the immune receptor NIK) rather than an artefact of ectopic expression of the gain-of-function mutant. Very importantly, the T474D-induced transcriptome explained the tolerant phenotype of both transgenic species against distinct begomoviruses. Thirdly, although virus infection and a sustained NIK pathway by ectopically expressing T474D may induce similar defense responses, the intensity of the output is expected to be higher in T474D transgenic lines because T474D bypasses viral NSP inhibition. Consistent with this prediction, both down-regulation of the translational machinery genes (see Figure 2D for comparison) and up-regulation of immune-related genes (RNA sequencing data) were more pronounced in T474D-overexpressing lines than in infected WT. Finally, expression of the T474D mutant potentiated the NIK-mediated response, as it would be expected from expression of a constitutively activated defense receptor NIK. The down-regulation of ribosomal gene expression (Figure 2C) and up-regulation of immune-related genes were more extreme in the T474D lines than in NIK lines. Very likely, the induction of the immune system and activation of a resistance-like response, as observed in T474D lines, require a sustained NIK signaling, because overexpression of normal NIK in tobacco does not cause induction of PR genes (Carvalho et al., 2008). Accordingly, the ectopic expression of the T474D gain-of-function mutant was more effective against begomovirus infection than overexpression of the NIK defense receptor in both Arabidopsis and tomato transgenic lines.

The experiments presented here shed light on the response underlying NIK-mediated antiviral defenses. We showed that constitutive activation of NIK in the T474D lines impaired global translation, and such activation might constitute an excellent strategy for fighting begomovirus infection in host cells. Nevertheless, antiviral translation defense systems have not been identified in

plants. In fact, plant viruses from diverse families use different gene expression strategies, such as cap-independent translation strategies, to bypass host regulatory mechanisms and sustain the translation of viral proteins (Kneller el al. 2006). Furthermore, the majority of plant viruses are (+) single-stranded RNA viruses that are not known to globally inhibit host translation. In the case of geminiviruses, however, which rely completely on the plant translation machinery and cannot circumvent host translational regulation, a global repression of translation is expected to significantly affect virus infection, as observed in the T474D-overexpressing lines. In fact, by assessing directly viral transcripts, we showed that the loading of coat protein mRNA into actively translating polysomes was significantly reduced in systemic infected leaves of T474D-overexpressing as compared to that of Col-O and NIK1-overexpressing lines (Figure 12). This indicates that suppression of global protein synthesis may effectively protect plant cells against DNA viruses.

The current mechanistic model for activation of the NIK-mediated antiviral signaling pathway holds that upon an unknown stimulus, the NIK LRR extracelullar domain undergoes oligomerization with itself or another receptor, allowing the intracellular kinase domains to transphosphorylate and to activate one another (Mariano et al., 2004). Activation of NIK by phosphorylation on the crucial threonine residue at position 474 leads to a regulated relocation of rpL10 to the nucleus to propagate the defense signaling cascade that impairs virus replication and/or movement. Our current data add some relevant insights towards understanding this layer of defense. They indicate that virus infection is the trigger of the NIK-mediated antiviral signaling and the output of transducing the defense signal consists in suppression of global translation and up-regulation of the immune system. Compelling evidence in the literature has revealed a fundamental role for members of the Arabidopsis LRRII-RLK subfamily as co-receptors for transducing defense and development signals (Chinchilla et al., 2009; Postel et al., 2010). As a member of the LRR-RLKII subfamily, NIK may serve as a co-receptor for independent branches of LRR-RLK-mediated signaling pathways in immunity and translational control. However, a LRR-RLK partner of NIK that could function in the plant innate immunity response has yet to be identified. Another question that remains unanswered is how NIK activation would mediate suppression of global translation. Yeast rpL10A is required for joining the 40S and 60S subunits

(Eisinger et al., 1997) and for the nuclear export of the large 60S subunit (Gadal et al., 2001). By analogy with the yeast rpL10A homologue, perturbation of Arabidopsis rpL10A nucleocytoplasmic trafficking by NIK1 would interfere with ribosome subunit assembly and 60S subunit export from the nucleus, which would affect general translation and impair virus infection. Nevertheless, our data indicate that suppression of global translation induced by constitutive activation of NIK is associated with a massive down-regulation of expression of translational machinery-related genes. Hence, it is possible that the NIK-mediated nuclecytoplasmis trafficking of rpL10 may be linked to regulation of gene expression. The possible extraribosomal functions of rpL10 associated with transcriptional factor regulation (Oh et al., 2002; Imafuku et al., 1999; Monteclaro and Vogt 1993) may serve as a potential target to assess this hypothesis. We have recently isolated a MYB domain-containing transcriptional factor that interacts with rpL10 *in vitro* and *in vivo* (data not shown). The identification of functional targets for the rpL10 partner will be crucial to elucidate the underlying mechanism for the NIK-mediated suppression of global translation.

## 5 – References


Albuquerque, Leonardo C, Arvind Varsani, Fernanda R Fernandes, Bruna Pinheiro, Darren P Martin, Paulo de Tarso Oliveira Ferreira, Thaís Oliveira Lemos, and Alice K Inoue-Nagata. 2012. "Further Characterization of Tomato-Infecting Begomoviruses in Brazil." *Archives of Virology* 157 (4): 747–52. doi:10.1007/s00705-011-1213-7.

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): R106. doi:10.1186/gb-2010-11-10-r106.

Anders, Simon, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. 2013. "Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor." *Nature Protocols* 8 (9): 1765–86. doi:10.1038/nprot.2013.099.

Ashburner, M, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nature Genetics* 25 (1): 25–29. doi:10.1038/75556.

Au, Kin Fai, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. 2010. "Detection of Splice Junctions from Paired-End RNA-Seq Data by SpliceMap." *Nucleic Acids Research* 38 (14): 4570–78. doi:10.1093/nar/gkq211.

Bullard, James H, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. 2010. "Evaluation of Statistical Methods for Normalization and Differential Expression in mRNA-Seq Experiments." *BMC Bioinformatics* 11: 94. doi:10.1186/1471-2105-11-94.

Carvalho, Claudine M, Anésia A Santos, Silvana R Pires, Carolina S Rocha, Daniela I Saraiva, João Paulo B Machado, Eliciane C Mattos, Luciano G Fietto, and Elizabeth P B Fontes. 2008. "Regulated Nuclear Trafficking of rpL10A Mediated by NIK1 Represents a Defense Strategy of Plant Cells against Virus." *PLoS Pathogens* 4 (12): e1000247. doi:10.1371/journal.ppat.1000247.

Castillo-Urquiza, Gloria P, José Evando A Beserra Jr, Fernanda P Bruckner, Alison T M Lima, Arvind Varsani, Poliane Alfenas-Zerbini, and F Murilo Zerbini. 2008. "Six Novel Begomoviruses Infecting Tomato and Associated Weeds in Southeastern Brazil." *Archives of Virology* 153 (10): 1985–89. doi:10.1007/s00705-008-0172-0.

Chinchilla, Delphine, Libo Shan, Ping He, Sacco de Vries, and Birgit Kemmerling. 2009. "One for All: The Receptor-Associated Kinase BAK1." *Trends in Plant Science* 14 (10): 535–41. doi:10.1016/j.tplants.2009.08.002.

Cumbie, Jason S, Jeffrey A Kimbrel, Yanming Di, Daniel W Schafer, Larry J

Wilhelm, Samuel E Fox, Christopher M Sullivan, et al. 2011. “GENE-Counter: A Computational Pipeline for the Analysis of RNA-Seq Data for Gene Expression Differences.” *PloS One* 6 (10): e25279. doi:10.1371/journal.pone.0025279.

Day, A G, E R Bejarano, K W Buck, M Burrell, and C P Lichtenstein. 1991. “Expression of an Antisense Viral Gene in Transgenic Tobacco Confers Resistance to the DNA Virus Tomato Golden Mosaic Virus.” *Proceedings of the National Academy of Sciences of the United States of America* 88 (15): 6721–25.

Dillies, Marie-Agnès, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, et al. 2012. “A Comprehensive Evaluation of Normalization Methods for Illumina High-Throughput RNA Sequencing Data Analysis.” *Briefings in Bioinformatics*. doi:10.1093/bib/bbs046.

Dutka, Alan F., and Howard H. Hanson. 1989. *Fundamentals of Data Normalization*. Addison-Wesley.

Edelbaum, Dagan, Rena Gorovits, Sonoko Sasaki, Masato Ikegami, and Henryk Czosnek. 2009. “Expressing a Whitefly GroEL Protein in Nicotiana Benthamiana Plants Confers Tolerance to Tomato Yellow Leaf Curl Virus and Cucumber Mosaic Virus, but Not to Grapevine Virus A or Tobacco Mosaic Virus.” *Archives of Virology* 154 (3): 399–407. doi:10.1007/s00705-009-0317-9.

Eisinger, D P, F A Dick, and B L Trumpower. 1997. “Qsr1p, a 60S Ribosomal Subunit Protein, Is Required for Joining of 40S and 60S Subunits.” *Molecular and Cellular Biology* 17 (9): 5136–45.

Fauquet, C M, R W Briddon, J K Brown, E Moriones, J Stanley, M Zerbini, and X Zhou. 2008. “Geminivirus Strain Demarcation and Nomenclature.” *Archives of Virology* 153 (4): 783–821. doi:10.1007/s00705-008-0037-6.

Fontes, Elizabeth P B, Anesia A Santos, Dirce F Luz, Alessandro J Waclawovsky, and Joanne Chory. 2004. “The Geminivirus Nuclear Shuttle Protein Is a Virulence Factor That Suppresses Transmembrane Receptor Kinase Activity.” *Genes & Development* 18 (20): 2545–56. doi:10.1101/gad.1245904.

Fu, Xing, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, et al. 2009. “Estimating Accuracy of RNA-Seq and Microarrays with Proteomics.” *BMC Genomics* 10: 161. doi:10.1186/1471-2164-10-161.

Gadal, O, D Strauss, J Kessl, B Trumpower, D Tollervey, and E Hurt. 2001. “Nuclear Export of 60s Ribosomal Subunits Depends on Xpo1p and Requires a Nuclear Export Sequence-Containing Factor, Nmd3p, That Associates with the Large Subunit Protein Rpl10p.” *Molecular and Cellular Biology* 21 (10): 3405–15. doi:10.1128/MCB.21.10.3405-3415.2001.

Garber, Manuel, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. 2011a. “Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq.” *Nature Methods* 8 (6): 469–77. doi:10.1038/nmeth.1613.

Grant, Gregory R, Junmin Liu, and Christian J Stoeckert Jr. 2005. “A Practical False Discovery Rate Approach to Identifying Patterns of Differential Expression in Microarray Data.” *Bioinformatics (Oxford, England)* 21 (11): 2684–90. doi:10.1093/bioinformatics/bti407.

Grant, Gregory R, Elisabetta Manduchi, and Christian J Stoeckert Jr. 2007. “Analysis and Management of Microarray Gene Expression Data.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 19: Unit 19.6. doi:10.1002/0471142727.mb1906s77.

Grigelionis, Bronius. 1990. *Probability Theory and Mathematical Statistics: Proceedings of the Fifth Vilnius Conference, June 25-July 1, 1989*. VSP.

Hansen, Kasper D, Rafael A Irizarry, and Zhijin Wu. 2012. “Removing Technical Variability in RNA-Seq Data Using Conditional Quantile Normalization.” *Biostatistics (Oxford, England)* 13 (2): 204–16. doi:10.1093/biostatistics/kxr054.

Hardcastle, Thomas J, and Krystyna A Kelly. 2010. “baySeq: Empirical Bayesian Methods for Identifying Differential Expression in Sequence Count Data.” *BMC Bioinformatics* 11: 422. doi:10.1186/1471-2105-11-422.

Hartigan, John. 1975. *Clustering Algorithms*. Books on Demand.

Hashmi, Jamil A, Yusuf Zafar, Muhammad Arshad, Shahid Mansoor, and Shaheen Asad. 2011. “Engineering Cotton (Gossypium Hirsutum L.) for Resistance to Cotton Leaf Curl Disease Using Viral Truncated AC1 DNA Sequences.” *Virus Genes* 42 (2): 286–96. doi:10.1007/s11262-011-0569-9.

Havelda, Zoltán, Eva Várallyay, Anna Válóczi, and József Burgyán. 2008. “Plant Virus Infection-Induced Persistent Host Gene Downregulation in Systemically Infected Leaves.” *The Plant Journal: For Cell and Molecular Biology* 55 (2): 278–88. doi:10.1111/j.1365-313X.2008.03501.x.

Horan, Kevin, Charles Jang, Julia Bailey-Serres, Ron Mittler, Christian Shelton, Jeff F Harper, Jian-Kang Zhu, John C Cushman, Martin Gollery, and Thomas Girke. 2008. “Annotating Genes of Known and Unknown Function by Large-Scale Coexpression Analysis.” *Plant Physiology* 147 (1): 41–57. doi:10.1104/pp.108.117366.

Hubbard, Stevan R, and W Todd Miller. 2007. “Receptor Tyrosine Kinases: Mechanisms of Activation and Signaling.” *Current Opinion in Cell Biology* 19 (2): 117–23. doi:10.1016/j.ceb.2007.02.010.

Imafuku, I, T Masaki, M Waragai, S Takeuchi, M Kawabata, S Hirai, S Ohno, et al. 1999. “Presenilin 1 Suppresses the Function of c-Jun Homodimers via

Interaction with QM/Jif-1." *The Journal of Cell Biology* 147 (1): 121–34.

Jansen, R, and M Gerstein. 2000. "Analysis of the Yeast Transcriptome with Structural and Functional Categories: Characterizing Highly Expressed Proteins." *Nucleic Acids Research* 28 (6): 1481–88.

Jean, Géraldine, André Kahles, Vipin T Sreedharan, Fabio De Bona, and Gunnar Rätsch. 2010. "RNA-Seq Read Alignments with PALMapper." *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* Chapter 11: Unit 11.6. doi:10.1002/0471250953.bi1106s32.

Jeanmougin, Marine, Aurelien de Reynies, Laetitia Marisa, Caroline Paccard, Gregory Nuel, and Mickael Guedj. 2010. "Should We Abandon the T-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies." *PloS One* 5 (9): e12336. doi:10.1371/journal.pone.0012336.

Jiang, Hui, and Wing Hung Wong. 2009. "Statistical Inferences for Isoform Expression in RNA-Seq." *Bioinformatics (Oxford, England)* 25 (8): 1026–32. doi:10.1093/bioinformatics/btp113.

Kadota, Koji, Tomoaki Nishiyama, and Kentaro Shimizu. 2012. "A Normalization Strategy for Comparing Tag Count Data." *Algorithms for Molecular Biology: AMB* 7 (1): 5. doi:10.1186/1748-7188-7-5.

Kanehisa, Minoru. 2013. "Molecular Network Analysis of Diseases and Drugs in KEGG." *Methods in Molecular Biology (Clifton, N.J.)* 939: 263–75. doi:10.1007/978-1-62703-107-3_17.

Kawaji, Hideya, and Yoshihide Hayashizaki. 2008. "Genome Annotation." *Methods in Molecular Biology (Clifton, N.J.)* 452: 125–39. doi:10.1007/978-1-60327-159-2_6.

Kent, W James. 2002. "BLAT--the BLAST-like Alignment Tool." *Genome Research* 12 (4): 656–64. doi:10.1101/gr.229202. Article published online before March 2002.

Kneller, Elizabeth L Pettit, Aurélie M Rakotondrafara, and W Allen Miller. 2006. "Cap-Independent Translation of Plant Viral RNAs." *Virus Research* 119 (1): 63–75. doi:10.1016/j.virusres.2005.10.010.

Langmead, Ben, Kasper D Hansen, and Jeffrey T Leek. 2010. "Cloud-Scale RNA-Sequencing Differential Expression Analysis with Myrna." *Genome Biology* 11 (8): R83. doi:10.1186/gb-2010-11-8-r83.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25. doi:10.1186/gb-2009-10-3-r25.

Levin, Joshua Z, Moran Yassour, Xian Adiconis, Chad Nusbaum, Dawn Anne

Thompson, Nir Friedman, Andreas Gnirke, and Aviv Regev. 2010. “Comprehensive Comparative Analysis of Strand-Specific RNA Sequencing Methods.” *Nature Methods* 7 (9): 709–15. doi:10.1038/nmeth.1491.

Li, Bo, Victor Ruotti, Ron M Stewart, James A Thomson, and Colin N Dewey. 2010. “RNA-Seq Gene Expression Estimation with Read Mapping Uncertainty.” *Bioinformatics (Oxford, England)* 26 (4): 493–500. doi:10.1093/bioinformatics/btp692.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Li, Heng, Jue Ruan, and Richard Durbin. 2008. “Mapping Short DNA Sequencing Reads and Calling Variants Using Mapping Quality Scores.” *Genome Research* 18 (11): 1851–58. doi:10.1101/gr.078212.108.

Lin, Ching-Yi, Wen-Shi Tsai, Hsin-Mei Ku, and Fuh-Jyh Jan. 2012. “Evaluation of DNA Fragments Covering the Entire Genome of a Monopartite Begomovirus for Induction of Viral Resistance in Transgenic Plants via Gene Silencing.” *Transgenic Research* 21 (2): 231–41. doi:10.1007/s11248-011-9523-9.

Lunter, Gerton, and Martin Goodson. 2011. “Stampy: A Statistical Algorithm for Sensitive and Fast Mapping of Illumina Sequence Reads.” *Genome Research* 21 (6): 936–39. doi:10.1101/gr.111120.110.

Mariano, Andrea C, Maxuel O Andrade, Anésia A Santos, Sonia M B Carolino, Marli L Oliveira, Maria Cristina Baracat-Pereira, Sergio H Brommonshenkel, and Elizabeth P B Fontes. 2004. “Identification of a Novel Receptor-like Protein Kinase That Interacts with a Geminivirus Nuclear Shuttle Protein.” *Virology* 318 (1): 24–31. doi:10.1016/j.virol.2003.09.038.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. 2008. “RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays.” *Genome Research* 18 (9): 1509–17. doi:10.1101/gr.079558.108.

McCall, Matthew N, and Rafael A Irizarry. 2008. “Consolidated Strategy for the Analysis of Microarray Spike-in Data.” *Nucleic Acids Research* 36 (17): e108. doi:10.1093/nar/gkn430.

McCarthy, Davis J., Yunshun Chen, and Gordon K. Smyth. 2012. “Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation.” *Nucleic Acids Research* 40 (10): 4288–97. doi:10.1093/nar/gks042.

McQuitty, Louis L. 1966. “Similarity Analysis by Reciprocal Pairs for Discrete

and Continuous Data." *Educational and Psychological Measurement* 26 (4): 825–31. doi:10.1177/001316446602600402.

Monaco, Marcela K, Joshua Stein, Sushma Naithani, Sharon Wei, Palitha Dharmawardhana, Sunita Kumari, Vindhya Amarasinghe, et al. 2013. "Gramene 2013: Comparative Plant Genomics Resources." *Nucleic Acids Research*. doi:10.1093/nar/gkt1110.

Monteclaro, F S, and P K Vogt. 1993. "A Jun-Binding Protein Related to a Putative Tumor Suppressor." *Proceedings of the National Academy of Sciences of the United States of America* 90 (14): 6726–30.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods* 5 (7): 621–28. doi:10.1038/nmeth.1226.

Oberg, Ann L., Brian M. Bot, Diane E. Grill, Gregory A. Poland, and Terry M. Therneau. 2012. "Technical and Biological Variance Structure in mRNA-Seq Data: Life in the Real World." *BMC Genomics* 13 (1): 304. doi:10.1186/1471-2164-13-304.

Oh, Hyung Suk, Haeyoung Kwon, Suk Kyun Sun, and Chul-Hak Yang. 2002. "QM, a Putative Tumor Suppressor, Regulates Proto-Oncogene c-Yes." *The Journal of Biological Chemistry* 277 (39): 36489–98. doi:10.1074/jbc.M201859200.

Postel, Sandra, Isabell Küfner, Christine Beuter, Sara Mazzotta, Anne Schwedt, Andrea Borlotti, Thierry Halter, Birgit Kemmerling, and Thorsten Nürnberger. 2010. "The Multifunctional Leucine-Rich Repeat Receptor Kinase BAK1 Is Implicated in Arabidopsis Development and Immunity." *European Journal of Cell Biology* 89 (2-3): 169–74. doi:10.1016/j.ejcb.2009.11.001.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, et al. 2011. "The Pfam Protein Families Database." *Nucleic Acids Research* 40 (D1): D290–D301. doi:10.1093/nar/gkr1065.

Qin, Li-Xuan, Richard P Beyer, Francesca N Hudson, Nancy J Linford, Daryl E Morris, and Kathleen F Kerr. 2006. "Evaluation of Methods for Oligonucleotide Array Data via Quantitative Real-Time PCR." *BMC Bioinformatics* 7: 23. doi:10.1186/1471-2105-7-23.

Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content Normalization for RNA-Seq Data." *BMC Bioinformatics* 12: 480. doi:10.1186/1471-2105-12-480.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics (Oxford, England)* 26 (1): 139–40. doi:10.1093/bioinformatics/btp616.

Robinson, Mark D, and Alicia Oshlack. 2010. "A Scaling Normalization Method

for Differential Expression Analysis of RNA-Seq Data." *Genome Biology* 11 (3): R25. doi:10.1186/gb-2010-11-3-r25.

Robinson, Mark D, and Gordon K Smyth. 2007. "Moderated Statistical Tests for Assessing Differences in Tag Abundance." *Bioinformatics (Oxford, England)* 23 (21): 2881–87. doi:10.1093/bioinformatics/btm453.

Robinson, Mark D., and Gordon K. Smyth. 2008. "Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data." *Biostatistics* 9 (2): 321–32. doi:10.1093/biostatistics/kxm030.

Rocha, Carolina S, Anésia A Santos, João Paulo B Machado, and Elizabeth P B Fontes. 2008. "The Ribosomal Protein L10/QM-like Protein Is a Component of the NIK-Mediated Antiviral Signaling." *Virology* 380 (2): 165–69. doi:10.1016/j.virol.2008.08.005.

Rojas, Maria R, Charles Hagen, William J Lucas, and Robert L Gilbertson. 2005. "Exploiting Chinks in the Plant's Armor: Evolution and Emergence of Geminiviruses." *Annual Review of Phytopathology* 43: 361–94. doi:10.1146/annurev.phyto.43.040204.135939.

Sakamoto, Tetsu, Michihito Deguchi, Otávio J B Brustolini, Anésia A Santos, Fabyano F Silva, and Elizabeth P B Fontes. 2012. "The Tomato RLK Superfamily: Phylogeny and Functional Predictions about the Role of the LRRII-RLK Subfamily in Antiviral Defense." *BMC Plant Biology* 12: 229. doi:10.1186/1471-2229-12-229.

Santos, Anésia A, Claudine M Carvalho, Lilian H Florentino, Humberto J O Ramos, and Elizabeth P B Fontes. 2009. "Conserved Threonine Residues within the A-Loop of the Receptor NIK Differentially Regulate the Kinase Function Required for Antiviral Signaling." *PloS One* 4 (6): e5781. doi:10.1371/journal.pone.0005781.

Santos, Anésia A, Kênia V G Lopes, Jorge A C Apfata, and Elizabeth P B Fontes. 2010. "NSP-Interacting Kinase, NIK: A Transducer of Plant Defence Signalling." *Journal of Experimental Botany* 61 (14): 3839–45. doi:10.1093/jxb/erq219.

Shedden, Kerby, Wei Chen, Rork Kuick, Debashis Ghosh, James Macdonald, Kathleen R Cho, Thomas J Giordano, et al. 2005. "Comparison of Seven Methods for Producing Affymetrix Expression Scores Based on False Discovery Rates in Disease Profiling Data." *BMC Bioinformatics* 6: 26. doi:10.1186/1471-2105-6-26.

Shimodaira, Hidetoshi. 2002. "An Approximately Unbiased Test of Phylogenetic Tree Selection." *Systematic Biology* 51 (3): 492–508. doi:10.1080/10635150290069913.

Smith, T F, and M S Waterman. 1981. "Identification of Common Molecular Subsequences." *Journal of Molecular Biology* 147 (1): 195–97.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee,

Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. “Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15545–50. doi:10.1073/pnas.0506580102.

Suzek, Baris E, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. 2007. “UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters.” *Bioinformatics (Oxford, England)* 23 (10): 1282–88. doi:10.1093/bioinformatics/btm098.

Tomato Genome Consortium. 2012. “The Tomato Genome Sequence Provides Insights into Fleshy Fruit Evolution.” *Nature* 485 (7400): 635–41. doi:10.1038/nature11119.

Trapnell, Cole, David G Hendrickson, Martin Sauvageau, Loyal Goff, John L Rinn, and Lior Pachter. 2013. “Differential Analysis of Gene Regulation at Transcript Resolution with RNA-Seq.” *Nature Biotechnology* 31 (1): 46–53. doi:10.1038/nbt.2450.

Trapnell, Cole, Lior Pachter, and Steven L Salzberg. 2009. “TopHat: Discovering Splice Junctions with RNA-Seq.” *Bioinformatics (Oxford, England)* 25 (9): 1105–11. doi:10.1093/bioinformatics/btp120.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15. doi:10.1038/nbt.1621.

Vu, Tien Van, Nirupam Roy Choudhury, and Sunil Kumar Mukherjee. 2013. “Transgenic Tomato Plants Expressing Artificial microRNAs for Silencing the Pre-Coat and Coat Proteins of a Begomovirus, Tomato Leaf Curl New Delhi Virus, Show Tolerance to Virus Infection.” *Virus Research* 172 (1-2): 35–45. doi:10.1016/j.virusres.2012.12.008.

Wang, Kai, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, et al. 2010. “MapSplice: Accurate Mapping of RNA-Seq Reads for Splice Junction Discovery.” *Nucleic Acids Research* 38 (18): e178. doi:10.1093/nar/gkq622.

Wang, Likun, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. 2010. “DEGseq: An R Package for Identifying Differentially Expressed Genes from RNA-Seq Data.” *Bioinformatics (Oxford, England)* 26 (1): 136–38. doi:10.1093/bioinformatics/btp612.

Wang, Xiaofeng, Michael B Goshe, Erik J Soderblom, Brett S Phinney, Jason A Kuchar, Jia Li, Tadao Asami, Shigeo Yoshida, Steven C Huber, and Steven D Clouse. 2005. “Identification and Functional Analysis of in Vivo Phosphorylation Sites of the Arabidopsis BRASSINOSTEROID-INSENSITIVE1 Receptor Kinase.” *The Plant Cell* 17 (6): 1685–1703.

doi:10.1105/tpc.105.031393.

Ward, Joe H. 1963. “Hierarchical Grouping to Optimize an Objective Function.” *Journal of the American Statistical Association* 58 (301): 236–44. doi:10.1080/01621459.1963.10500845.

Wu, Thomas D, and Colin K Watanabe. 2005. “GMAP: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.” *Bioinformatics (Oxford, England)* 21 (9): 1859–75. doi:10.1093/bioinformatics/bti310.

Yang, Ei-Wen, Thomas Girke, and Tao Jiang. 2013. “Differential Gene Expression Analysis Using Coexpression and RNA-Seq Data.” *Bioinformatics (Oxford, England)*. doi:10.1093/bioinformatics/btt363.

# CHAPTER II

## READS: a friendly RNA-seq platform to perform multiple counting methods for Differential Gene Expression (DGE)

Otávio J B Brustolini, Fabyano F. Silva, Elizabeth P. B. Fontes

**Abstract**

The highvolume of RNA sequencing data provided by many high-throughput techniques, such as the next generation RNA sequencing technology (RNA-seq), is now accessible to any research laboratory and is quickly becoming established as a key research tool in any global gene expression experiment. In an RNA-seq workflow, the reads (short sequences generated by the RNA-seq technology) are mapped (aligned) to a reference sequence data set (transcriptome or genome). A counting table (number of reads per gene) can be set up, and further downstream analysis can be performed to recover the biological meaning of the experiment. However, this protocol can be arduous for anyone who is not an R experienced user. To determine which genes are significantly different in the expression profile among treatments and to evaluate the biological meaning though annotation, many R scripts must be created and correctly executed. The variety of options for differential gene expression (DGE) methodology available on the R/Bioconductor makes this task even more difficult. Our platform was designed to fill these gaps and make these iterations faster and easier. However, it achieves still more by allowing combinations of the p-values generated by the DGE methods: edgeR, DESeq2, baySeq and NBPSeq. Inspired by the meta analysis, we used a combined p-value calculated by the Fisher's method, weighted Z-test, truncated product method, binomial test or a simple intersection (average or median) to assist in identifying of the statistically significantdifferentially expressed (DE) genes based on these multiple DGE methods. To accomplish this goal, the friendly interface interacts with low-level R scripts to perform an RNA-seq analysis without directly using a large number of scripts. Indeed, the analysis will be completely performed within

a few steps. A directory under the project name will be used to save all of the generated files and store a SQLite database containing the DE genes with their expression values and annotation. As the program intended to be completely free and open to contributions, all the program's methods have been designed to be transparent to the end user. This platform is completely free.

## 1 - INTRODUCTION

Recent advances in transcriptome studies have been accelerated by the development of high-throughput data technologies. The microarray hybridization method and the next generation RNA sequencing technology (RNA-seq) are examples of high-throughput techniques that have been applied in global gene expression experimental designs. The RNA-seq platform addresses a multitude of applications, including relative expression analyses, alternative splicing, discovery of novel transcripts and isoforms, RNA editing, allele-specific expression and the exploration of non-model-organism transcriptomes (Anders et al., 2013). Given the resulting enormous amount of data, a high-level analysis should be performed to extract the maximum biological significance between and within treatments (Garber et al., 2011). After the raw data processing that is generally provided by the filtering software, it is necessary to map the reads to a reference genome or transcriptome database. To accomplish this task, many mapping programs have been proposed with many types of improvements in their alignment algorithms (Garber et al., 2011; Mortazavi et al., 2008; Kanehisa 2013). The outcome of the aligner program can be given in a SAM format file, which can be used to generate the reads counting table (Li et al., 2009). The counting table file enables the differential gene expression (DGE) approach. The differentially expressed genes between conditions are predicted using advanced statistical tests, which associate each differential gene expression with corrected p-values. The enormousnumber of tested genes provided by the counting table are associated with p-values, which are the results from the multiple simultaneous tests of many hypotheses. Therefore, these p-values require corrections because a large proportion of false positives can occur. Then, these corrected p-values can be used as a decision value to determine what gene expression level is statistically significant between conditions. Using a corrected p-value cut-off, the groups of genes that display statistically significant differential gene expression in a particular contrast are identified. A variety of methods are available to accomplish the DGE analysis. Based on the nature of the counting reads in RNA-seq data, a Poisson distribution would be obvious choice. However, because of the variability reported by the biological replicates, the mean and variance are not equal;instead, the observed variance

is much greater than the mean. Hence, several studies have reported that this distribution does not account for biological variability across the samples (Robinson and Smyth 2007; Robinson and Smyth 2008; Anders et al., 2013). A more appropriate alternative extension is the Negative Binomial (NB) distribution. Therefore, the majority of DGE methods use tests based on the NB distribution, which are distinguished by their approaches to estimating the dispersion or the normalization factors.

The most widely used DGE methods are presented in the R/Bioconductor (http://www.bioconductor.org) software (Gentleman et al., 2004). Given the great variety of DGE methods, selecting the appropriate DGE method for a given set of RNA-seq data has become a problem. Using a collection of multiple DGE methods could be an interesting alternative. Because no consensus about the best choice for DGE methods exists, exploring more than one DGE methods would enhance our opportunities to find statistically significant differentially expressed genes. The number of accepted genes could differ greatly among the DGE methods. An intersection among all DGE results could be too rigorous to find the DE genes because the variation among the methods could be so high that no genes would be retrieved. Thus, we have proposed a software platform to allow the user to choose which methodologies are more suitable to explore the data. We have still implemented five methods to combine a multiple p-values into a single value: Fisher's method, the weighted Z-test, the truncated product method, the binomial test and simple intersection (average or median). The combined p-values can help the user to identify the significant DE genes. To facilitate the interactions between the user and the software, a graphical interface has been developed based on the graphical user interface (GUI) framework QT 4.8 (http://qt-project.org) using the language C++.

## 2 - IMPLEMENTATION

As the reads were implemented based on a graphical interface, the modules of the system were separated by their functions. The core that interacts with the R scripts was implemented in a separated module using standard C++ and the R language. The interface modules send the parameters to the system core, and the functions parse the generated results. The main module controls all the user/system core interactions. The system core creates the R scripts, run them and parses the outcome. The results are the formatted data retrieved by the DGE methods (edgeR, DESeq2, baySeq and NBPSeq) outputs. The combined p-value module was implemented in C++ using the equations provided in the supplementary materials (S1). The graphics for cluster analysis, Smear Plots, heatmap and volcano plot are implemented in R script embedded in C++ core code. Using SeqAN library (https://www.seqan.de), the bam and GFF files were parsed and formatted for addition to the SQLite database. Figure 1A illustrates the implementation diagram of the software interface and core layers.

## 3 - RESULTS

An example data set from a real *Arabidopsis thaliana* experiment will illustrate the potential of the software. TheArabidopsis experiment was designed using four biological replicates with two samples for wild type (WT) and two for the mutant NIK T474D lines (Santos et al., 2009). The reads were obtained from the mRNA-Seq Illumina HiSeq 2000 experiment. First, the reads were mapped on the *Arabidopsis* transcriptome TAIR10 (http://www.arabidopsis.org) database using the Bowtie 2.1.0 aligner. The mapping output was generated as four SAM format files (Li et al., 2009). Our software parses the content of the SAM files and generates the counting table in comma separated values (csv) format. If the counting table in csv has already been generated by another program, it can still be read using the program counter option. For the loading of this counting table, the experimental design must be determined. An easy interface was created to guide the user toward a suitable experimental design. However, it is possible to sendthe R type formula directly to the factorial design interface. Because in our example we have a single comparison of T474D vs WT plants, we chose the pairwise option. The set up options were enabled, and the multiple DGE methods were tested. A combined p-value method was used. For illustration, we used all possible options. Figure 1B shows the DE genes based on the options selected in the set up. At the low level, the DGE algorithms interact with the R/Bioconductor packages through scripts presented at the core of the program, which runs using the R command. The DGE method comparisons using the combined p-value intersection could be more balanced regarding the results of the consensus outcomes. However, for the example data the combined p-value shows greater confidence than the individual p-values from the DGE results. The low number of DE genes identified by the Fisher's method is explained by its naïve assumption of the independence of the DGE methods (Fig 1B). For a more realistic assumption, the combined p-values based on the methods might be more appropriate. Nevertheless, all these combined p-valuemethods can be tested by the user. After the DGE analysis, the downstream section is enabled, and the annotation of the genes can be recovered. Moreover, some graphics options becmone avalailable. A more advanced analysis such as clusters can also be performed.

All of the applied scripts will be presented in the working subdirectory

forvisualization by the user. In addition, the producecd results will be presented in the same subdirectory. However, the program divides all of the results into separated groups. Our goal is to make the system transparent to the user, i.e.,to allow the user to know how the program works internally, meaning that all the scripts and source code used can be retrieved, analyzed and changed by the user. Furthermore, our application was designed to offer a friendly user interface with the R commands at the low level of the program. A user with limmited experience in the R environment will be able to perform a complete analysis without knowing any R commands, while an experienced R user could also use our program to make his analysis faster or to customize his own analysis at the low level.

The DE genes will be stored in the SQLite database. The annotation added to the DE genes could also be exported to csv tables. A more complex experimental design could be exploited if main goal was to detect the effects of an experiment. This approach is implemented in all the DGE methods used in R/Bioconductor. At the end of the analysis, the outcomes of the DEG methods are distinguished for each methodology used, and the graphics created by the user will be in the project directory.

## 4 - CONCLUSION

Many programs,including online programs, have been presented to analyze RNA-seq data. However, the main problem of these programsis the lack of flexibility to compare or combine multiple DGE methods, specially those presented in R/Bioconductor. However, these analyses could be easily performed using the terminal by an experienced R user. Meanwhile, the majority of researchers need fast answers without spending extended time to learn terminal commands. The installation of this program on a local computer contributes to the confidentiality of the data and avoids the server overload in the case of many accessions. The flexibility, easy usability and fast interaction make our system the right choice to perform advanced analyses without spendindescessive resources ona commercial solution.

ACKNOWLEDGEMENTS

This research was supported by the CNPq grants 573600/2008-2 and 470287/2011-0. O.J.B. received a graduate fellowship from CNPq

## 5. REFERENCES


Anders, Simon, Davis J McCarthy, Yunshun Chen, Michal Okoniewski, Gordon K Smyth, Wolfgang Huber, and Mark D Robinson. 2013. “Count-Based Differential Expression Analysis of RNA Sequencing Data Using R and Bioconductor.” *Nature Protocols* 8 (9): 1765–86. doi:10.1038/nprot.2013.099.

Garber, Manuel, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. 2011. “Computational Methods for Transcriptome Annotation and Quantification Using RNA-Seq.” *Nature Methods* 8 (6): 469–77. doi:10.1038/nmeth.1613.

Gentleman, Robert C., Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, et al. 2004. “Bioconductor: Open Software Development for Computational Biology and Bioinformatics.” *Genome Biology* 5 (10): R80. doi:10.1186/gb-2004-5-10-r80.

Jiang, Hui, and Wing Hung Wong. 2009. “Statistical Inferences for Isoform Expression in RNA-Seq.” *Bioinformatics (Oxford, England)* 25 (8): 1026–32. doi:10.1093/bioinformatics/btp113.

Kanehisa, Minoru. 2013. “Molecular Network Analysis of Diseases and Drugs in KEGG.” *Methods in Molecular Biology (Clifton, N.J.)* 939: 263–75. doi:10.1007/978-1-62703-107-3_17.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Marioni, John C, Christopher E Mason, Shrikant M Mane, Matthew Stephens, and Yoav Gilad. 2008. “RNA-Seq: An Assessment of Technical Reproducibility and Comparison with Gene Expression Arrays.” *Genome Research* 18 (9): 1509–17. doi:10.1101/gr.079558.108.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. “Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq.” *Nature Methods* 5 (7): 621–28. doi:10.1038/nmeth.1226.

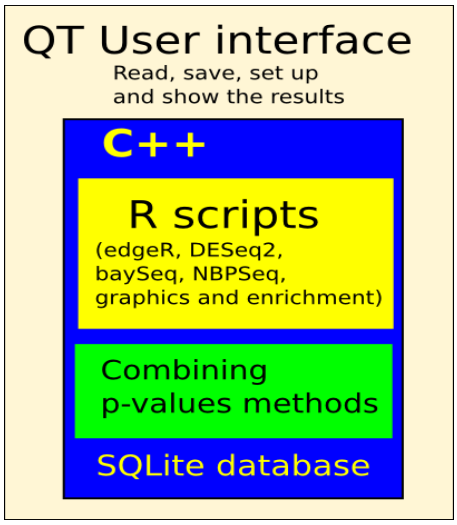
Robinson, Mark D, and Gordon K Smyth. 2007. “Moderated Statistical Tests for Assessing Differences in Tag Abundance.” *Bioinformatics (Oxford, England)* 23 (21): 2881–87. doi:10.1093/bioinformatics/btm453.

Robinson, Mark D., and Gordon K. Smyth. 2008. “Small-Sample Estimation of Negative Binomial Dispersion, with Applications to SAGE Data.” *Biostatistics* 9 (2): 321–32. doi:10.1093/biostatistics/kxm030.

Santos, Anésia A, Claudine M Carvalho, Lilian H Florentino, Humberto J O

Ramos, and Elizabeth P B Fontes. 2009. “Conserved Threonine Residues within the A-Loop of the Receptor NIK Differentially Regulate the Kinase Function Required for Antiviral Signaling.” *PloS One* 4 (6): e5781. doi:10.1371/journal.pone.0005781.

Trapnell, Cole, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. 2010. “Transcript Assembly and Quantification by RNA-Seq Reveals Unannotated Transcripts and Isoform Switching during Cell Differentiation.” *Nature Biotechnology* 28 (5): 511–15. doi:10.1038/nbt.1621.



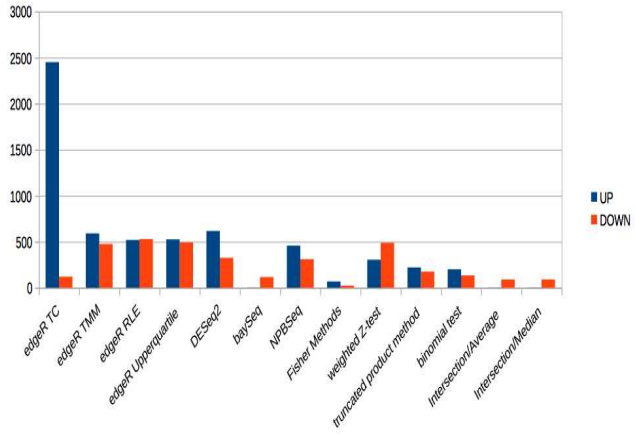




**Fig. 1 – A – Diagram of the internal structure of the reads: the QT interface, the C++ classes and the R scripts. B – Comparison of the accepted gene expression in the Arabidopsis experiment, which is significantly different for each of the DGE methods and for the combed p-value methods.**

**R scripts used for differential expression in Chapter I**

```
#edgeR

library(edgeR)
#Import the data to dataframe
raw<- read.csv2("differential_expressions_raw.csv",header=T, row.name=1)

trats<- read.table ("Trats.txt", header=F)
trats<- as.vector (trats[,1])
contrasts<- unique (trats)
n<- length(contrasts)

d<- DGEList (counts = raw , group = trats)
#Set up normalization factors: none (TC), TMM (default), RLE or Upperquartile
d<- calcNormFactors(d, method="TMM")
d<- estimateCommonDisp(d)
d<- estimateTagwiseDisp(d, prior.df = getPriorN(d), grid.length = 1000)

#Perform each contrast by pair
for (i in 1:n)
for (j in 1:n)
        if (i < j){
                cont1 = contrasts[i]
                cont2 = contrasts[j]
                de<-exactTest(d, pair=c(cont1,cont2), dispersion="tagwise")
                dt<- topTags (de, n=20000, adjust.method="fdr")
                tab<- dt$table[ dt$table$FDR < 0.05,]
                write.csv (tab, paste("./results/", cont2, "-",cont1, ".csv", sep="")
        }


#DESeq

library(DESeq)

raw<- read.csv2("differential_expressions_raw.csv", header=T, row.names=1)

cond<- read.table ("Trats.txt",header=F)
cond<- as.vector(cond[,1])
n<- length(cond)
mraw<- as.matrix (raw)

#Preparing the sample object
cds<- newCountDataSet(mraw, cond)
cds<- estimateSizeFactors(cds)

#Estimatingthe dispersion
cds<- estimateDispersions(cds, method="pooled", sharingMode="fit-only",
fitType="local")

#Perform each contrast by pair
```

```
for (i in 1:n){
for (j in 1:n){
        if (i < j){
#The test
                res<- nbinomTest (cds, cond[i], cond[j])
#Save the top counts by padj < 0.05
                res<- na.omit (res)
                top<- res[res$padj < 0.05,]
                write.csv (top, paste ("./resultados/", cond[j], "-",cond[i],".csv",
sep=""),row.names=F)
                }
        }
}

#baySeq

library(baySeq)
raw<- read.csv2("differential_expressions_raw.csv", header=T, row.names=1)

trats<- read.table ("Trats.txt", header=F)[,1]
colnames(raw) <- trats
n <- length (trats)

rep<- c(1,1,2,2)
gr<- list (NDE = c(1,1,1,1), DE = c(1,1,2,2))

#Perform the test by each pair of contrasts
for (i in seq (1, n, 2))
for (j in seq (1, n, 2))
        if (i > j){
                mdata<- as.matrix(raw[,c(i, i+1, j, j+1)])
                cont<- trats[c(i,j)]

#Prepare the data
                lsize<- apply (mdata, 2, sum)

#Prepare the samples
                CD <- new("countData", data=mdata, replicates=rep,
libsizes=as.integer(lsize), groups=gr)

#Model: Negative Binomial
                CDNB <- getPriors.NB (CD, samplesize=10000, cl=NULL)
                CDPostNB <- getLikelihoods.NB(CDNB, pET = "BIC", cl = NULL)
                tabNB<- topCounts (CDPostNB, group=2, FDR=0.05)
                write.csv2 (tabNB, paste ("./results/", cont[1], cont[2],".csv",
sep="")
}
```