sciendo

DOI: 10.2478/bile-2022-0009

Biometrical Letters
Vol. 59 (2022), No. 2, 127-139

Application of edgeR and DESeq2 methods in plant experiments based on RNA-seq technology

Grażyna Niedziela, Alicja Szabelska-Beręsewicz, Joanna Zyprych-Walczak, Małgorzata Graczyk

Department of Mathematical and Statistical Methods, Poznan University of Life Sciences, Wojska Polskiego 28, Poznań, 60-637, Poland,
e-mail: grazyna.niedziela@up.poznan.pl, alicja.szabelska-beresewicz@up.poznan.pl, joanna.zyprych-walczak@up.poznan.pl, malgorzata.graczyk@up.poznan.pl

Summary

We compared two of the most common methods for differential expression analysis in the RNA-seq field: edgeR and DESeq2. We evaluated these methods based on four real RNA-seq plant datasets. The results indicate that there is a large number of joint differentially expressed genes between the two methods. However, depending on the research goal and the preparation of an experiment, different approaches to statistical analysis and interpretation of the results can be suggested. We focus on answering the question: what workflow should be used in the statistical analysis of the datasets under consideration to minimize the number of falsely identified differentially expressed genes?

Key words: edgeR, DESeq2, RNA-seq, statistical methods for differential analysis

1. Introduction

RNA-seq technology based on transcriptome sequencing is particularly used in differential expression studies. This technology allows the investigation of thousands of genes at the same time. It can supply information about the expression profile of those genes. Many analyses using RNA-seq technologies relate to medical or animal data. There are far fewer plant, especially agricultural, datasets in publicly available RNA-seq data repositories. Thanks to transcriptome studies (Birney et al., 2007), knowledge about genomes, DNA structure and the mechanisms accompanying its modifications is developing rapidly. The sequencing method was first proposed by Sanger and

Coulson (Sanger and Coulson, 1975; Sanger et al., 1977). It was a milestone in the development of sequencing techniques. The Sanger method, called first-generation sequencing, was a predecessor to the next-generation technology (NGS). This consists of tools that enable transcriptome analysis by RNA-seq sequencing (Metzker, 2010). The methodology of research based on RNA-seq technology enabled significant development in science in the field of genomic structure and mechanisms related to gene transcription in prokaryotic (Van Vliet, 2010) and eukaryotic (Wang et al., 2009) organisms. Next-generation sequencing can deliver the complete sequence of transcripts present in the cell (Pepke et al., 2009; Martin et al., 2010). There are two main steps in the sequencing experiment: preliminary and secondary data analysis (Marioni et al., 2008). For secondary analysis, according to Garber et al. (2011), there are three main classes of software appropriate to the type of biological problems addressed. These are tools used to find the differential expression of genes (DEG), find new expression regions and discover the structure of transcripts. The approach of looking for differential expression very often uses aggregated data resulting from counts of reads matched to a given gene (Marioni et al., 2008). The methodology used for calculation counts is similar to the analysis of data from microarray experiments using appropriate probability distributions. The most commonly used distribution is the Pascal distribution (Oshlack et al., 2010; Anders and Huber, 2010; Hardcastle and Kelly, 2010). There are also methods based on other distributions, e.g. Poisson (Auer and Doerge, 2011; Auer et al., 2012) or methods using a non-parametric approach (Li and Tibshirani, 2011). Apparently, there exist many statistical methods used for seeking genes that differentiate between two or more conditions, causing scientists often to ask which method to use. There are many publications comparing different methods. For instance in Li et al. (2022) the authors compare eight methods: edgeR, DESeq, DESeq2, baySeq, EBSeq, NOISeq, SAMSeq and Voom, using both simulated and real datasets. The comparisons were across different scenarios with either equal or unequal library sizes, different distribution assumptions and sample sizes. The conclusion of those studies was that different methods gave the best results depending on the parameters taken into account. A similar framework is found in a study by Shahjaman et al. (2020). There, the authors describe a comparison of seven selection methods: DEseq, DEseq2, SAMseq, Bayseq, limma (voom), edgeR and edgeR robust, using both simulated and real datasets. From this study the authors conclude, based on the presence of outliers in relation to the effectiveness of the presented

methods, that most of the existing DEG selection methods produce lower accuracies and higher false discoveries in the presence of outliers. There are some robust approaches such as edgeR robust and DEseq2 that perform well in the presence of outliers for large-sample cases, but they have weak performance for small-sample cases in the presence of outliers. In Liu et al. (2021) a detailed protocol is provided to identify DEGs between cholangiocarcinoma (CHOL) and normal tissues using limma, DESeq2 and EdgeR. The authors conclude that the three methods are partly similar. The choice of the method should depend on the data, since all three methods have their own advantages.
A similar conclusion is reached in Lin et al. (2016). Those authors argued that the algorithms employed in DESeq2 and edgeR are at some points different, but the sets of DEG were in good agreement. In an attempt to test this statement in the field of plant RNA-seq data, the two most popular methods of differential analysis will be presented: edgeR (Robinson et al., 2010) and DESeq2 (Anders and Huber, 2010; Love et al., 2014).

This study is confined to methods that are implemented and available within the R framework (R Core Team, 2020) and that are applicable to count tables. We further focus on finding genes that are differentially expressed between two conditions only, since this is arguably the most commonly encountered application.

The choice of method may change the results and conclusions of the research; therefore the main aim of this work is to compare the two most frequently chosen methods of statistical analysis for differential expression, namely edgeR and DESeq2.

2. Materials and methods

2.1. Materials

To compare the effectiveness of the edgeR and DESeq2 methods on different datasets, in our analysis we consider data that contain only two classes: the test group and the control group.

We considered four sequencing experiments performed based on: Bread Wheat (Cantu et al., 2011), Arabidopsis (Pass et al., 2017), Tomato Plant (Frąckowiak et al., 2022) and Wheat (Wielkopolan et al., 2022). More details about the datasets are presented in Table 1.

The first dataset, Bread Wheat (*Triticum aestivum*), is divided into two groups as follows: wheat growing naturally without any interference,

Table 1. Information about the four considered RNA-seq datasets

Dataset	Source	No. genes	No. groups	No. samples
Bread Wheat	Cantu et al. (2011)	30470	2	7
Arabidopsis	Pass et al. (2017)	28598	2	8
Tomato	Frąckowiak et al. (2022)	35768	2	8
Wheat	Wielkopolan et al. (2022)	56531	2	6

and wheat in which the Grain Protein Content (GPC) transcription factor had a reduced expression. Increasing the concentration of nutrients in wheat grains is important to alleviate nutritional deficiencies in many parts of the world.
The proteins and nutrients in the wheat grain come largely from the remobilization of degraded leaf particles during the monocarpic aging process. Monocarpic senescence is an active process during which the plant must disassemble complex molecules, increase active transport mechanisms and maintain functional conductive tissues, while coordinating the programmed death of depleted leaf cells.

The second considered dataset consists of material coming from Arabidopsis (*Arabidopsis thaliana*). The first part of the dataset consisted of plants that were grown in limited light, and the second group of radishes that were intensely lit. The biological material in this experiment was the tissues of the cut plant parts. The question asked was: Which genes are activated when the intensity of plant irradiation is changed? The analysis of the Arabidopsis genome shows that the resolution of the size of the chromatin particles by differential digestion with MNase allows the detection of sensitive features of the chromatin structure. Simultaneous analysis of transcription levels reveals the influence of transient exogenous factors in modifying the subnucleosomal landscape in association with transcriptional changes, leading to the binding of transcriptional factors.

The third dataset consists of material concerning Tomato Plants (*Solanum lycopersicum L.*). The first group is a control group in which the plant was watered only with water. In the second case, the tomato was treated with the resistance inducer benzothiadiazole (BTH). The biological material was the leaves of the tomato plant. Using differential analysis methods, genes were sought that activate when the plant is treated with the inducer BTH.

The last experiment was carried out based on Wheat (*Triticum aestivum*). The biological material was plant leaves. In the first group, bio-

logical material was collected from naturally growing plants without any interference. In the second, the plants were treated with an insect without a microbiome. The purpose of the research was to find genes that are activated when a plant is treated with an insect without a microbiome.

Principal component analysis was used to check whether the grouping of samples was consistent with the assumptions of the experiments. The first two principal components for the analyzed datasets are presented in Figure 1. For all datasets, the first two components explain at least 80% of the variability.

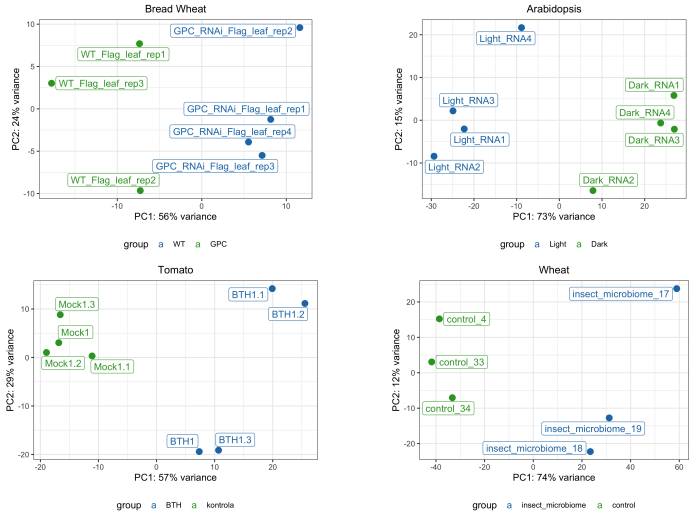


Figure 1. Principal component analysis for datasets

2.2. Methods for differential analysis: EdgeR and DESeq2

The aim of differential analysis is to determine which genes from the database have significantly different expression levels between the groups considered. In this study, we analyze four datasets using two parametric methods: edgeR (Robinson et al. (2010)) and DESeq2 (Love et al. (2014)).

Let Y_{ij}^g denote the number of counts for the gene g in the j-th sample of the i-th group, where $i \in \{A, B\}$, $j = 1, \ldots, n_i$, $g = 1, \ldots, G$, n_i is the number of samples in the i-th group, and G denotes the total number of genes. Let us assume that the variable Y_{ij}^g follows a negative binomial

distribution with the parameters:

$$Y_{ij}^g \sim NB(\mu_{ij}^g, \phi_g), \tag{1}$$

where $E(Y_{ij}^g) = \mu_{ij}^g$ and ϕ_g is the dispersion parameter of gene g. Let us assume that λ_i^g is the true expression level of the gene g in the i-th group, and d_{ij} is the normalizing factor for the j-th sample in the i-th group. Then the relationship between the expected value and expression level can be given as:

$$\mu_{ij}^g = d_{ij}\lambda_i^g. \tag{2}$$

The tested hypotheses are of the form:

$$\begin{cases} H_0 & : \lambda_A^g = \lambda_B^g \\ H_1 & : \sim H_0 \end{cases}. \tag{3}$$

Following the approach given by Robinson and Smyth (2008) let $Z_i^g = \sum_{j=1}^{n_i} Y_{ij}^g \sim NB(n_i d^* \lambda_i^g, \phi_g n_i^{-1})$ be a random variable representing the expression level for gene g in the i-th group, and d^* be a geometric mean of library sizes.
Then:

$$\widehat{\lambda_i^g} = Z_i^g \tag{4}$$

is an estimator of the expression level of gene g in the groups under consideration.
As stated in Robinson and Smyth (2008), if all of the library samples were the same, the dispersion parameter can be estimated using the conditional maximum likelihood estimator (CMLE) which maximizes the following function:

$$\ell_{Y_{ij}^g | Z_i^g = z_i^g}^g(\phi_g) = \sum_{i=1}^{2} \left[\sum_{j=1}^{n_i} \log\Gamma(y_{ij}^g + \phi_g^{-1}) - n_i \log\Gamma(\phi_g^{-1}) - \psi_i^g \right], \tag{5}$$

where

$$\psi_i^g = \log\Gamma(z_i^g + n_i\phi_g^{-1}) + \log\Gamma(n_i\phi_g^{-1}). \tag{6}$$

For experiments with unequal library sizes Robinson and Smyth (2008) proposed the quantile-adjusted conditional maximum likelihood (qCML) procedure to obtain a dispersion estimator.

The test procedure is performed with an exact test that is similar to Fisher's test, where the hypergeometric distribution is replaced with the negative binomial distribution.

According to the approach given by Anders and Huber (2010), Love et al. (2014) the DESeq2 method is described. Similarly to the edgeR algorithm, it is assumed that the number of counts follows the negative binomial distribution with parameters consistent with equations (1) and (2).

To determine the dispersion parameter, the additional assumption is made that the local relationship between the level of expression of genes and their variance is linear, i.e.:

$$\phi_g(Y_{ij}^g) = \mu_{ij}^g + d_{ij}^2 v_i^g, \tag{7}$$

where v_i^g is the variance of gene g in the i-th group.

It is worth noting that both methods use the negative binomial distribution to model the data. In many RNA-seq experiments the data consist of thousands of genes and only a few samples ($g >> n$). In such cases, there is a need to reduce the high dimensionality. One way to overcome this problem is to first use a test procedure for each gene separately. Next, correction for multiple comparisons is applied by calculation of the adjusted p-values. In both methods edgeR and DESeq2, the False Discovery Rate (FDR) introduced by Benjamini and Hochberg (1995) is used in the second step of the testing procedure to enable comparison of the results between genes and to verify whether expression varies significantly between groups.

2.3. Statistical analysis

In the statistical analysis, the data were first preprocessed. All genes with very low expression (below 10 counts) across the examined samples were discarded. The counts for every gene were normalized with the TMM method introduced by Robinson and Oshlack (2010). Differential expression analysis was performed using linear models implemented in the edgeR (version 3.38.4) and DESeq2 (version 1.36.0) packages. The set of differentially expressed genes (DEGs) was determined based on the adjusted p-value (FDR), set to be smaller than 0.05. All figures in this paper were created using the ggplot2 (version 3.3.6) R package. All statistical analyses were carried out using R software 4.2.1 (R Core Team, 2020).

3. Results

Based on differential analysis performed with the edgeR and DESeq2 methods, the sets of DEGs were obtained. In Table 2, the numbers of DEGs for each dataset together with percentages of shared genes with respect to each method are presented. Based on the results given in Table 2 it can be said that all of the considered experiments yield large sets of differentiating genes (however, edgeR has a smaller number of DEGs in almost every case) and a large proportion of these genes are shared by both methods.

Table 2. Sizes of lists of DEGs obtained from the datasets by the edgeR and DESeq2 methods

Dataset	Method			
	edgeR		DESeq2	
	# DEG	% of shared genes	# DEG	% of shared genes
Bread Wheat	3195	60.38	2039	94.61
Arabidopsis	2490	98.92	3198	77.02
Tomato	785	99.24	970	80.31
Wheat	4596	98.35	5284	85.54

A relationship between statistical significance (measured as $-\log FDR$) and the magnitude of changes in expression ($\log_2 FC$) can be presented in the form of a scatterplot called a volcano plot. The volcano plots for the individual experiments are presented in Figure 2. In this plot, the most up-regulated genes are towards the right (representing genes with increased expression in the study group), the most down-regulated genes are towards the left (representing genes with reduced expression in the study group), and the most statistically significant genes are towards the top. Genes are called significant if they have $FDR < 0.05$. From the biological point of view, one may also be interested in genes with the greatest fold-changes (FC), both up- and down-regulated. Therefore, an additional restriction is often used to filter out genes with too small FC. The green color marks the genes for which $FDR < 0.05$ and $|\log_2 FC| < 1.5$. The blue color marks the points representing the genes for which $FDR < 0.05$ and $|\log_2 FC| > 1.5$. In all experiments, there is a higher number of up-regulated genes with high significance. This tendency applies to both the edgeR and DESeq2 methods.

The histograms presented in Figure 3 demonstrate the distribution of differentiating genes for each method and each dataset.

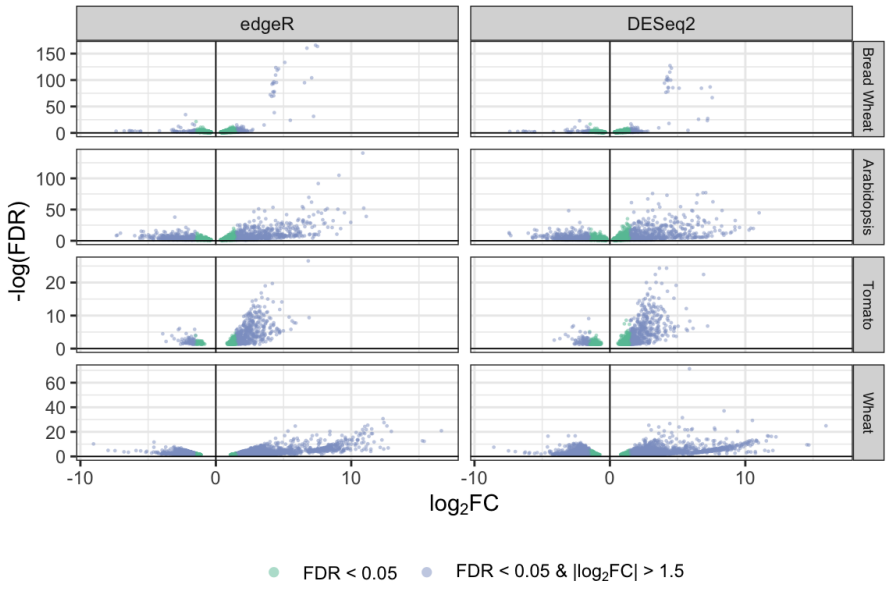


Figure 2. Volcano plots of DEGs for the datasets and the two methods

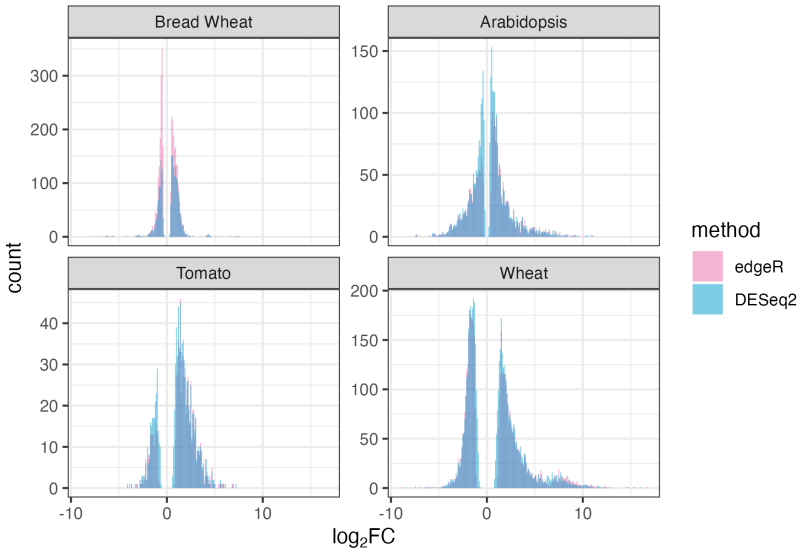


Figure 3. The histograms for DEGs for the datasets and the two methods

In general, both methods show a similar distribution of differentiating genes. It is worth noting that for the Bread Wheat dataset more differentially expressed genes were determined with the edgeR method. However, for the other experiments, more differentiating genes were selected using the DESeq2 method. A summarization of the joint portion of differentially expressed genes is presented in Figure 4. Venn diagrams make it possible to compare the overlap of DEGs in two or more statistical comparison tracks. In Figure 4 the number of up- and down-regulated genes is shown for each experiment and each method.

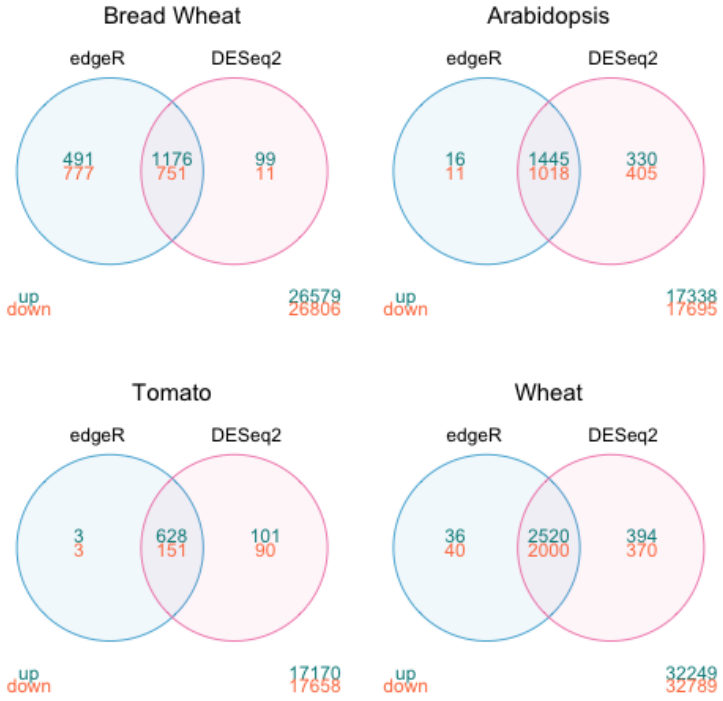


Figure 4. Venn diagram for edgeR and DESeq2 for datasets based on the significant genes

From these figures, we can see that in most cases, the DESeq2 method obtained a greater number of differentiating genes than the edgeR method. The reverse is true only for the Bread Wheat experiment. It should be noted that the genes that edgeR selects as differentiating are in most cases also selected by DESeq2.

4. Conclusions

In this paper, we have compared the performance of two different test statistics in selecting genes for two-class RNA-seq expression data. The results of the experiments show that a large proportion of the DEGs are common to both methods; however, edgeR has more of its selected genes in common with DESeq2 than vice versa. In most cases, the DESeq2 method indicates a greater number of DEGs than edgeR. This analysis of four different datasets showed that despite the large similarities in the reporting of DEGs, there are also some differences that should be noted in the cases when in the experiment we did not observe many DEGs. A good practice, in this case, would be to use the edgeR and DESeq2 methods during the statistical analysis, and to give as the final result the intersection of these two lists of differentiating genes. Attention should also be paid to the need to use diagnostic plots (such as volcano plots, MA plots, and density plots), which can show the advantages and disadvantages of using a given method in a specific dataset.

Acknowledgements

The high-performance calculations were carried out at the Poznan Supercomputing Networking Centre, project 392.

References

Anders S., Huber W. (2010): Differential expression analysis for sequence count data. Genome Biology 11(10): R106.

Auer P.L., Doerge R.W. (2011): A two-stage poisson model for testing rna-seq data. Statistical Applications in Genetics and Molecular Biology 10(1).

Auer P.L., Srivastava S., Doerge R. (2012): Differential expression—the next generation and beyond. Briefings in Functional Genomics 11(1): 57–62.

Benjamini Y., Hochberg Y. (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57(1): 289–300.

Birney E., Stamatoyannopoulos J.A., Dutta A., Guigó R., Gingeras T.R., Margulies E.H., Weng Z., Snyder M., Dermitzakis E.T., Thurman R.E. et al. (2007): Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. Nature 447(7146): 799–816.

Cantu D., Pearce S.P., Distelfeld A., Christiansen M.W., Uauy C., Akhunov E., Fahima T., Dubcovsky J. (2011): Effect of the down-regulation of the high Grain

Protein Content (GPC) genes on the wheat transcriptome during monocarpic senescence. BMC Genomics 12(1): 492.

Frąckowiak P., Wrzesińska B., Wieczorek P., Sanchez-Bel P., Kunz L., Dittmann A., Obrępalska-Stęplowska A. (2022): Deciphering of benzothiadiazole (bth)-induced response of tomato (solanum lycopersicum l.) and its effect on early response to virus infection through the multi-omics approach. Plant and Soil pages 1–24.

Garber M., Grabherr M.G., Guttman M., Trapnell C. (2011): Computational methods for transcriptome annotation and quantification using rna-seq. Nature Methods 8(6): 469–477.

Hardcastle T.J., Kelly K.A. (2010): bayseq: empirical bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics 11(1): 1–14.

Li D., Zand M.S., Dye T.D., Goniewicz M.L., Rahman I., Xie Z. (2022): An evaluation of rna-seq differential analysis methods. PLOS ONE 17(9): 1–19.

Li J., Tibshirani R. (2011): Finding consistent patterns: a nonparametric approach for identifying differential expression in rna-seq data. Statistical Methods in Medical Research 22(5): 519–536.

Lin Y., Golovnina K., Chen Z.X., Lee H.N., Negron Y.L.S., Sultana H., Oliver B., Harbison S.T. (2016): Comparison of normalization and differential expression analyses using rna-seq data from 726 individual drosophila melanogaster. BMC Genomics 17(1): 1–20.

Liu S., Wang Z., Zhu R., Wang F., Cheng Y., Liu Y. (2021): Three differential expression analysis methods for rna sequencing: limma, edger, deseq2. Journal of Visualized Experiments (175): e62528.

Love M.I., Huber W., Anders S. (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15(12): 550.

Marioni J.C., Mason C.E., Mane S.M., Stephens M., Gilad Y. (2008): Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research 18(9): 1509–1517.

Martin J., Bruno V.M., Fang Z., Meng X., Blow M., Zhang T., Sherlock G., Snyder M., Wang Z. (2010): Rnnotator: an automated de novo transcriptome assembly pipeline from stranded rna-seq reads. BMC Genomics 11(1): 1–8.

Metzker M.L. (2010): Sequencing technologies—the next generation. Nature Reviews Genetics 11(1): 31–46.

Oshlack A., Robinson M.D., Young M.D. (2010): From rna-seq reads to differential expression results. Genome Biology 11(12): 1–10.

Pass D.A., Sornay E., Marchbank A., Crawford M.R., Paszkiewicz K., Kent N.A., Murray J.A.H. (2017): Genome-wide chromatin mapping with size resolution reveals a dynamic sub-nucleosomal landscape in Arabidopsis. PLOS Genetics 13(9): e1006988.

Pepke S., Wold B., Mortazavi A. (2009): Computation for chip-seq and rna-seq

studies. Nature Methods 6(11): S22–S32.

R Core Team (2020): *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing Vienna, Austria.

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

Robinson M.D., Oshlack A. (2010): A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11(3): R25.

Robinson M.D., Smyth G.K. (2008): Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9(2): 321–332.

Sanger F., Coulson A.R. (1975): A rapid method for determining sequences in dna by primed synthesis with dna polymerase. Journal of Molecular Biology 94(3): 441–448.

Sanger F., Nicklen S., Coulson A.R. (1977): Dna sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences 74(12): 5463–5467.

Shahjaman M., Manir Hossain Mollah M., Rezanur Rahman M., Islam S.S., Nurul Haque Mollah M. (2020): Robust identification of differentially expressed genes from rna-seq data. Genomics 112(2): 2000–2010.

Van Vliet A.H. (2010): Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiology Letters 302(1): 1–7.

Wang Z., Gerstein M., Snyder M. (2009): Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10(1): 57–63.

Wielkopolan B., Frąckowiak P., Wieczorek P., Obrępalska-Stęplowska A. (2022): The Impact of Oulema melanopus—Associated Bacteria on the Wheat Defense Response to the Feeding of Their Insect Hosts. Cells 11(15): 2342.