Santos et al. BMC Genomics (2020) 21:35
https://doi.org/10.1186/s12864-019-6426-2

BMC Genomics

METHODOLOGY ARTICLE Open Access

Custom selected reference genes outperform pre-defined reference genes in transcriptomic analysis

Karen Cristine Gonçalves dos Santos[1,2], Isabel Desgagné-Penix[1,2] and Hugo Germain[1,2*]

Abstract

Background: RNA sequencing allows the measuring of gene expression at a resolution unmet by expression arrays or RT-qPCR. It is however necessary to normalize sequencing data by library size, transcript size and composition, among other factors, before comparing expression levels. The use of internal control genes or spike-ins is advocated in the literature for scaling read counts, but the methods for choosing reference genes are mostly targeted at RT-qPCR studies and require a set of pre-selected candidate controls or pre-selected target genes.

Results: Here, we report an R-based pipeline to select internal control genes based solely on read counts and gene sizes. This novel method first normalizes the read counts to Transcripts per Million (TPM) and then excludes weakly expressed genes using the DAFS script to calculate the cut-off. It then selects as references the genes with lowest TPM covariance. We used this method to pick custom reference genes for the differential expression analysis of three transcriptome sets from transgenic *Arabidopsis* plants expressing heterologous fungal effector proteins tagged with GFP (using GFP alone as the control). The custom reference genes showed lower covariance and fold change as well as a broader range of expression levels than commonly used reference genes. When analyzed with NormFinder, both typical and custom reference genes were considered suitable internal controls, but the custom selected genes were more stably expressed. geNorm produced a similar result in which most custom selected genes ranked higher (i.e. were more stably expressed) than commonly used reference genes.

Conclusions: The proposed method is innovative, rapid and simple. Since it does not depend on genome annotation, it can be used with any organism, and does not require pre-selected reference candidates or target genes that are not always available.

Keywords: Next-generation sequencing, Housekeeping genes for qPCR, R script

Background

RNAseq is a technique used since the pioneer studies of R Lister, RC O'Malley, J Tonti-Filippini, BD Gregory, CC Berry, AH Millar and JR Ecker [1] (*Arabidopsis thaliana*), U Nagalakshmi, Z Wang, K Waern, C Shou, D Raha, M Gerstein and M Snyder [2] (*Saccharomyces cerevisiae*), BT Wilhelm, S Marguerat, S Watt, F Schubert, V Wood, I Goodhead, CJ Penkett, J Rogers and J Bähler [3] (*Schizosaccharomyces pombe*), and A Mortazavi, BA Williams, K McCue, L Schaeffer and B Wold [4] (*Mus musculus*). This technique allows the combination of transcript discovery and expression level quantification in a single assay and has an unlimited dynamic range of detection compared to microarray or RT-qPCR [5, 6].

For differential expression studies, the gene expression values must be comparable between samples, which means that count data should be normalized for sequencing depth and other biases such as transcript length, GC content and transcript coverage. Reads/Fragments per Kilobase per Million (RPKM or FPKM) and Transcripts per Million (TPM) both normalize count data by transcript length and sequencing depth [7], but they may give biased results in the presence of highly expressed genes or when a lot of the genes are expressed in only one sample [8]. This is because one differentially expressed gene shifts

* Correspondence: Hugo.Germain@uqtr.ca
[1]Department of Chemistry, Biochemistry and Physics, Université du Québec à Trois-Rivières, Trois-Rivières, QC G9A 5H7, Canada
[2]Plant Biology Research Group, Université du Québec à Trois-Rivières, Trois-Rivières, QC G9A 5H7, Canada

the sequencing effort distributed to the others and all genes appear to be differentially expressed [9–11]. Other methods such as relative log expression (DESeq2) and trimmed mean of M-values (edgeR) can work with the carry-over effect of highly expressed genes [10].

The comparison of different softwares for RNAseq analysis is a recurrent subject in the literature [12–14] and many authors argue over the benefits of using housekeeping genes or spike-in controls to scale the count data, yet the evaluation of the reference genes used for RNAseq data analysis is not as common. When using internal or external control genes, the normalization is first performed on the controls and the result is used to normalize the other genes. The use of external spike-ins is advocated for introducing little error into the read counts, allowing identification of global shifts in gene expression [15–17]. However, reports have shown mixed performances with different normalization methods [18], resulting in high false discovery rates and false positive rates [19]. These may show differences in amplification depending on the type of tissue studied or the protocol for mRNA enrichment [20].

One alternative for external spike-ins is the use of internal control genes, as it is done in qPCR studies. Typical control genes are actin, tubulin, elongation factor 1, polyubiquitin and ribosomal RNAs, though the stability of expression of several of those is dependent on the conditions studied [21]. To solve this issue, different algorithms were proposed to find stably expressed genes, mostly for qPCR applications, but they need a set of predefined genes of interest (RefGenes, T Hruz, O Laule, G Szabo, F Wessendorp, S Bleuler, L Oertle, P Widmayer, W Gruissem and P Zimmermann [22]) or a set of preselected candidate reference genes (geNorm, J Vandesompele, K De Preter, F Pattyn, B Poppe, N Van Roy, A De Paepe and F Speleman [23]; NormFinder, CL Andersen, J Ledet-Jensen and T Ørntoft [24]; BestKeeper, MW Pfaffl, A Tichopad, C Prgomet and TP Neuvians [25]). The most frequent approach is to take previously identified stably expressed genes, as done by B Zhuo, S Emerson, JH Chang and Y Di [11], this however does not ensure that the selected genes will show stable expression in the studied organism and conditions.

Here we propose a simple and fast method to identify the most stably expressed genes for each experimental condition. Our method is aimed at differential expression studies and represents a simple way to select custom reference genes for any species or any type of experiments, so they can be used in the normalization step of differential expression analysis algorithms, and does not necessitate spike-ins. It alleviates the problem inherent to predefined reference genes, which may not be stably expressed across experimental set-ups and are applicable to a single species.

Results

Initially three RNAseq transcriptomes were generated using *Arabidopsis* transgenic plants expressing GFP alone (control) or GFP-fused to fungal effector genes (*Mlp37347* and *Mlp124499*). We tested the normalization of our RNAseq data using two sets of reference genes: commonly used reference genes (Table 1) and the 104 stably expressed *Arabidopsis* genes proposed by B Zhuo, S Emerson, JH Chang and Y Di [11]. The first set of reference genes was assessed for stability in three different permutations of the transcriptome sets as shown in Fig. 1a (panel 1: Mlp37347 vs Control, panel 2: Mlp124499 vs Control, panel 3: Mlp124499 vs Mlp37347). In each case, high levels of covariance, ranging from 4.9% (NDUFA8 in Mlp124499 vs Mlp37347) to 41.5% (tubulin 6 in Mlp124499 vs Mlp37347) were obtained. Next, we performed the same analysis using the 104 genes proposed by B Zhuo, S Emerson, JH Chang and Y Di [11]. For the three permutations of the transcriptome sets, important fluctuations in the covariance were observed ranging from 2.9 to 49% (Fig. 1b). Finally, we did the same for the set of 30 genes selected by T Czechowski, M Stitt, T Altmann, MK Udvardi and W-R Scheible [26] for several plant tissues (Additional file 1). These results demonstrate that neither the commonly used reference genes, nor the 104 reference genes proposed by B Zhuo, S Emerson, JH Chang and Y Di [11] were stably expressed in our conditions.

In order to search for more stably expressed genes, we developed a custom method to select reference genes using only one's own RNAseq data. We first used a R function to transform the count data into Transcripts per Million [27] and calculate the average TPM and covariance for each gene. We then used the

Table 1 Common reference genes used in this study for comparison against custom selected reference genes

Symbol	Name	ATG
Actin 2	ACT2	AT3G18780
Actin 7	ACT7	AT5G09810
Actin 8	ACT8	AT1G49240
Adenine phosphoribosyltransferase 1	APT1	AT1G27450
Elongation factor 1-α	EF1α	AT5G60390
Eukaryotic translation initiation factor 4A-1	eIF4A	AT3G13920
NADH-ubiquinone oxidoreductase 19-kDa subunit	NDUFA8	AT5G18800
Tubulin β-2/β-3 chain	TUB2	AT5G62690
β-tubulin 6	TUB6	AT5G12250
Tubulin β-9 chain	TUB9	AT4G20890
Polyubiquitin	UBQ4	AT5G20620
Ubiquitin extension protein	UBQ5	AT3G62250
Polyubiquitin	UBQ10	AT4G05320
Polyubiquitin	UBQ11	AT4G05050

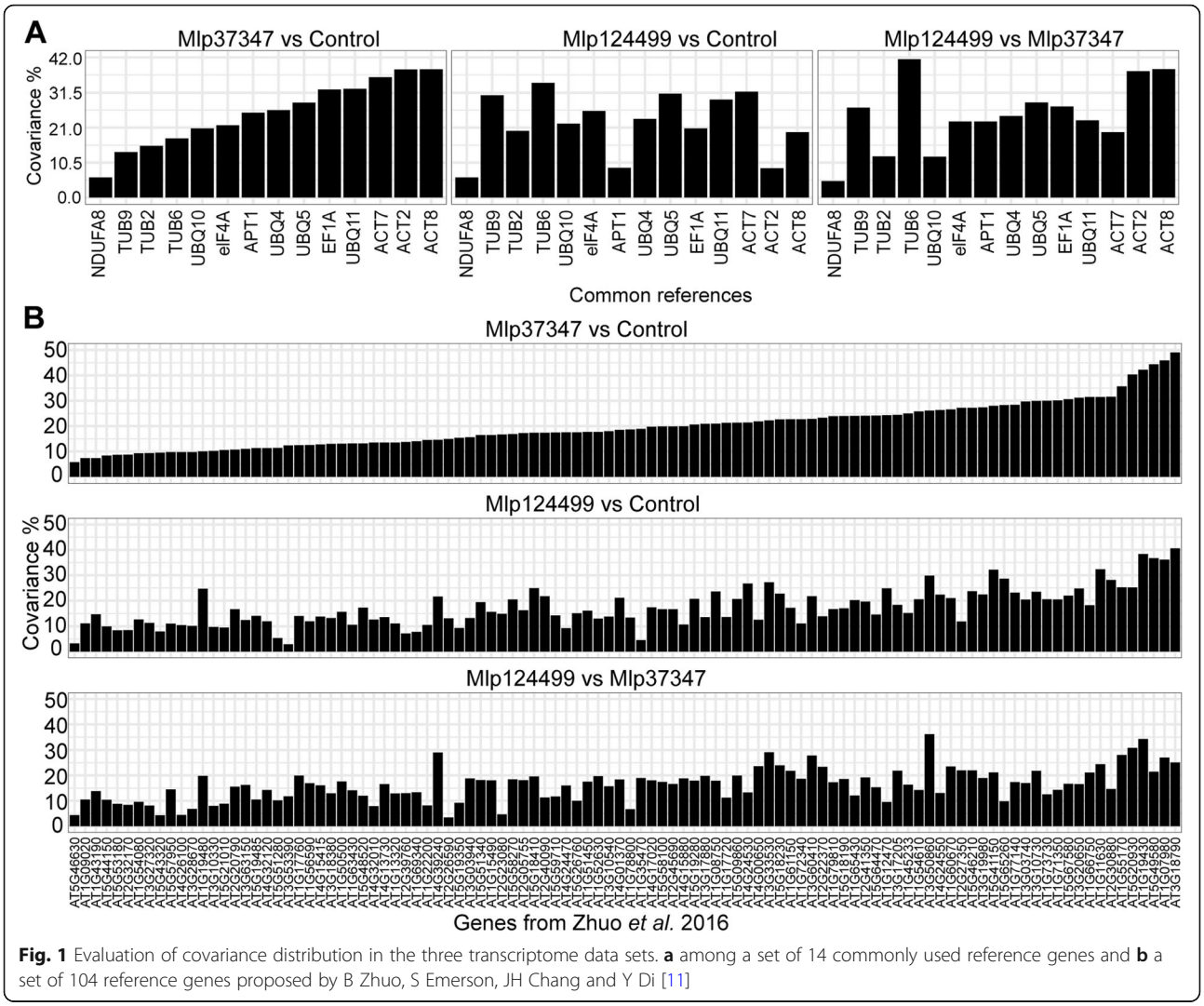

Fig. 1 Evaluation of covariance distribution in the three transcriptome data sets. **a** among a set of 14 commonly used reference genes and **b** a set of 104 reference genes proposed by B Zhuo, S Emerson, JH Chang and Y Di [11]

DAFS function [28] to calculate a cut-off for the exclusion of weakly expressed genes. Finally, the 0.5% remaining genes with lowest covariance were selected as reference genes (R-package "CustomSelection" [29]). This pipeline is thereafter referred to as the custom selection script.

To test the developed method, we used the same transcriptome sets described in Fig. 1 (the list of selected genes for each analysis is available in Table 1, Additional file 2). For each transcriptome set, we show in Fig. 2 the average expressing in $\log_2$ TPM and covariance of the common reference genes (Common), the set of 30 genes from T Czechowski, M Stitt, T Altmann, MK Udvardi and W-R Scheible [26] (Czechowski et al. 2005), the set of 104 genes from B Zhuo, S Emerson, JH Chang and Y Di [11] (Zhuo et al. 2016) and the genes selected using the CustomSelection package [29] (Custom script). In all pairings the custom selected reference genes show broader range of expression levels and lower covariance (Fig. 2) than the other sets. Next, we performed a differential expression analysis with DESeq2 [30] without control genes. We show in Fig. 3 the $\log_2$-transformed fold change by the $-\log_{10}$-transformed adjusted *p*-value for each gene set. We can see that the set of genes selected with the custom script shows lower fold change in all cases. We also compared the results of DESeq2 using no reference gene or the four sets indicated above for each permutation. As is shown in Table 2, in all the permutations the analysis without the use of references gives higher number of up-regulated genes than the analyses that use any of the reference sets while resulting in a lower number of down-regulated genes, possibly indicating a shift to downregulation that is not detected without reference genes.

To further test the stability of the custom reference genes in our experiment, we used NormFinder [24] and

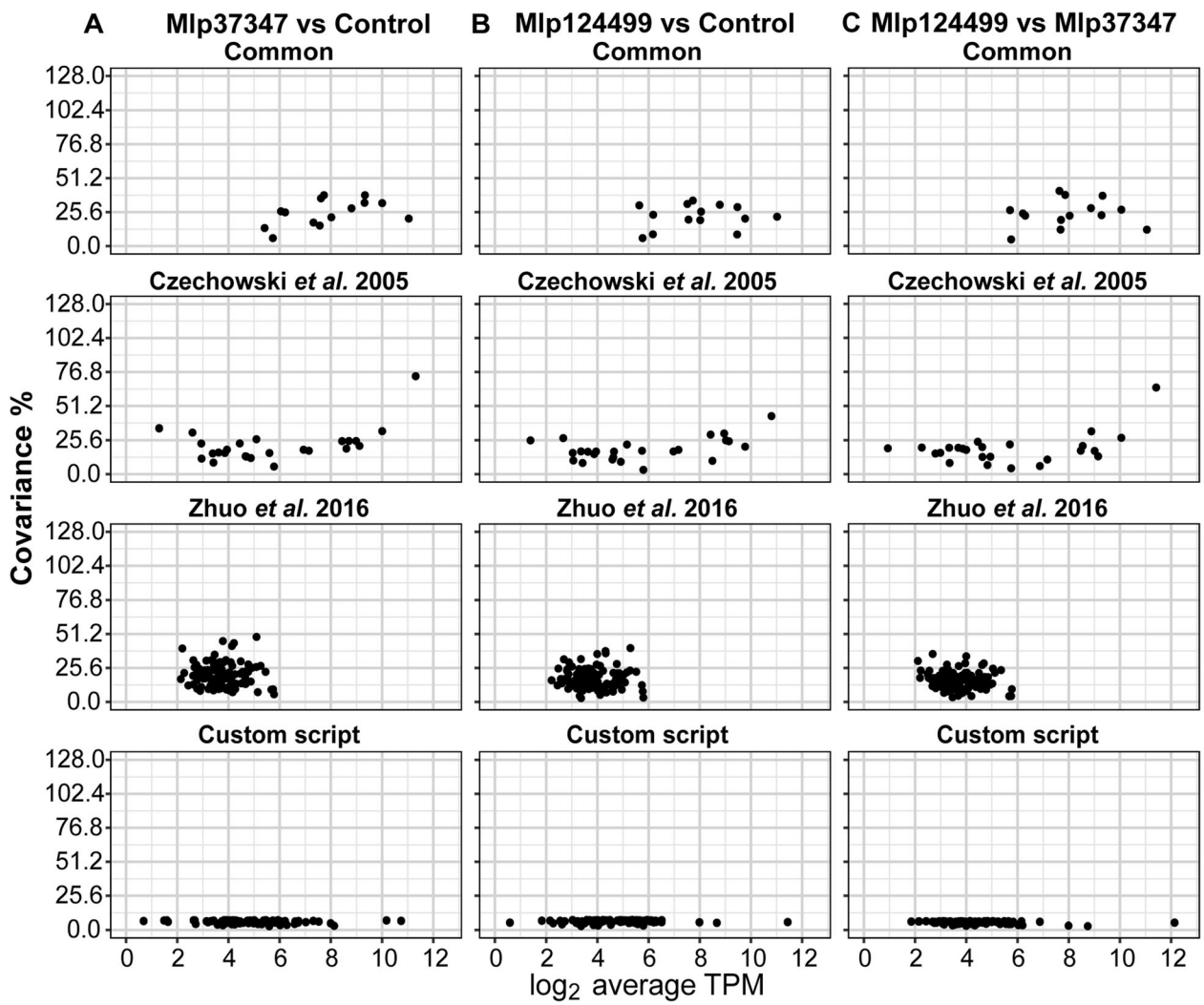

Fig. 2 Comparison the four sets of reference genes in relation to covariance level and $\log_2$ TPM for **a** Mlp37347 vs Control, **b** Mlp124499 vs Control and **c**) Mlp124499 vs Mlp37347

geNorm [23] to compare the four sets of reference genes using $\log_2$ transformed TPM values. The complete result is presented in the Tables S3-S5 of the Additional file 2. We present in Fig. 4 the comparison of the set of common reference genes against the custom selected reference genes. The gene AT5G18800 (NDUFA8) which is in the set of common references was selected by the custom script in all three permutations and is shown with a purple border. Both sets of genes (custom and common refences) were under the stability threshold of NormFinder (0.5), meaning that the software considers them suitable references genes, however the custom selected genes (shown with a blue border) were more stable than the commonly used genes (shown in red, Fig. 4). This was also the case for most genes tested with geNorm.

Discussion

The use of reference genes in RNAseq studies is suggested in the literature [15–17], yet the methods for the selection of these genes are designed for qPCR data and require a set of pre-selected reference or target genes or the selection of conditions similar to that of one's own experiment [22–25], which are not always available. As there is no previous transcriptomic study of plants constitutively expressing fungal effectors and since the information available on these effectors is scarce [31], it is not possible to know a priori their function and which host genes are impacted by the presence of these fungal proteins. For these reasons, we propose a new R-package which enables the selection of custom reference genes regardless of the organisms used or of the experimental conditions.

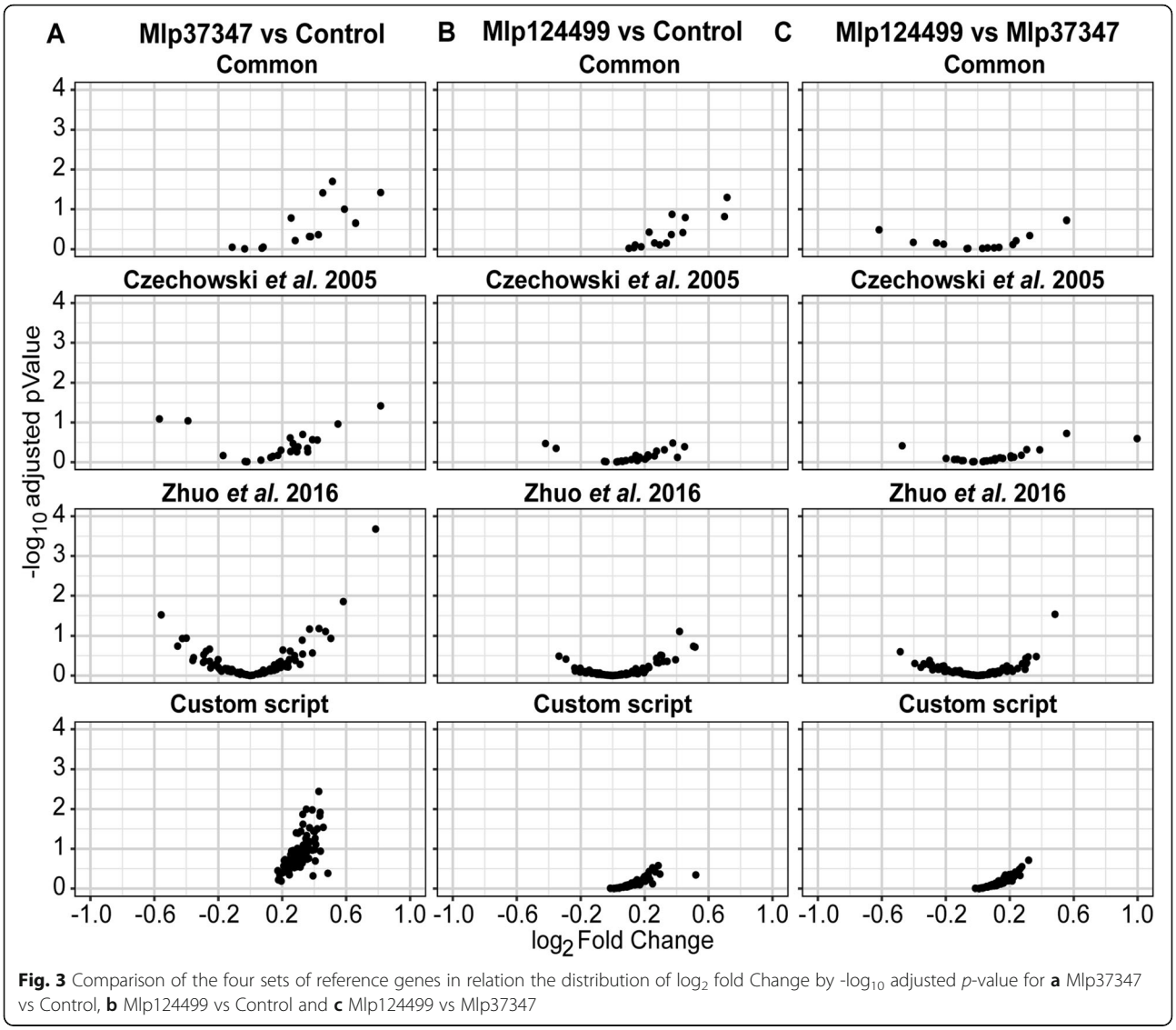

Fig. 3 Comparison of the four sets of reference genes in relation the distribution of $\log_2$ fold Change by $-\log_{10}$ adjusted *p*-value for **a** Mlp37347 vs Control, **b** Mlp124499 vs Control and **c** Mlp124499 vs Mlp37347

The method developed here only requires information available from the RNAseq analyses. It uses Transcripts per Million [27] as a proxy for the expression level and the DAFS algorithm [28] to exclude genes with low counts, which may be inactive [32]. We first assessed whether the most commonly used reference genes (Table 1) or two sets of published reference genes for *Arabidopsis* [11, 26] were indeed stably expressed in our experimental conditions. As demonstrated in Fig. 1 and Additional file 1, three sets of reference genes show a high level of co-variance in our experimental conditions, indicating that they were not suitable reference genes for our differential expression analysis.

Having a high level of variability in the expression of the reference genes results in skewed quantitative analysis and may cause the loss of some differentially expressed genes which show modest variation in gene expression [21]. In relation to the reference gene sets, there is minimal overlap between sets published and the ones selected in this article (maximum of 5 genes shared between our set and the set of B Zhuo, S Emerson, JH Chang and Y Di [11] and 2 genes shared between our set and the set of T Czechowski, M Stitt, T Altmann, MK Udvardi and W-R Scheible [26], shown in Additional file 2 Table S3, S4, S5 column J). However, there is extensive overlap in the deregulated genes (up- and down-regulated as shown in Additional file 2: Table S2). This fact demonstrates that all three sets perform well in detecting deregulated genes, however having a references gene set with lower co-variance results in the finding of more de-regulated genes (Additional file 2: Table S2 downregulated) since more subtle deregulation can be detected.

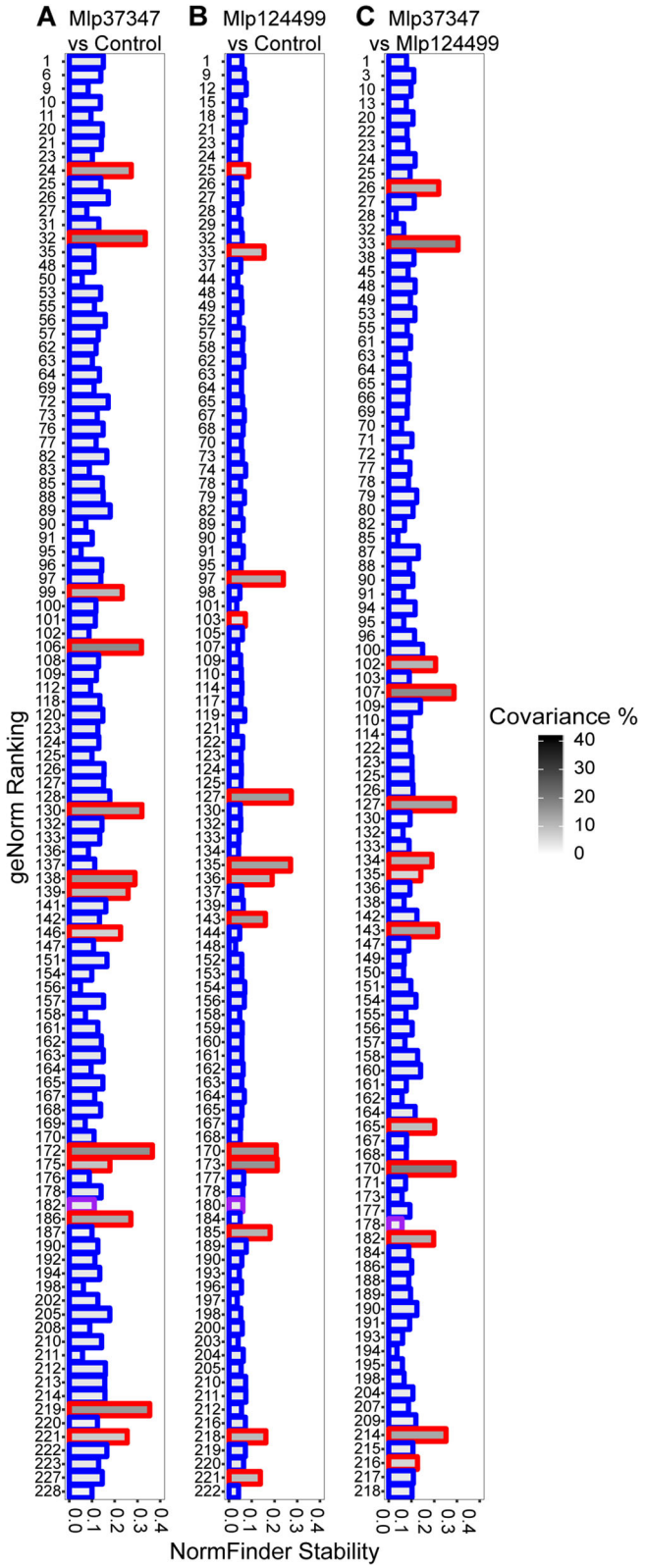

Fig. 4 Comparison of custom selected reference genes (blue border) and commonly used reference genes (red border) with geNorm ranking, NormFinder stability index and covariance for **a** Mlp37347 vs Control, **b** Mlp124499 vs Control and **c** Mlp124499 vs Mlp37347. The bar with purple border indicates the gene (NDUFA8) selected with the custom script that is also present in the common references

Thus, to alleviate the bias inherent to the use of inappropriate reference genes, we devised a R-based pipeline to select custom reference genes for one's own experimental data. As presented in Figs. 2 and 3, in all the pairings of the data used, the custom selected reference genes outperformed the other sets of reference genes in their expression stability, presenting lower fold changes and lower covariances. Our method allows the selection of genes more stably expressed and the selection of more genes as references (the final number is user defined, with the default setting being 0.5% of the expressed genes), giving more reference points, hence more robustness, to the normalization of genes expressed at different levels. The advantage of having a user-defined threshold is that when there is extensive variation in the data, a stringent threshold may result in the selection of few or no genes as references. On the contrary, extremely homogenous data would result in a very large reference gene set, for this reason a user-defined threshold is preferable.

Conclusions

Our results show the need for a new R-based pipeline for the selection of custom reference genes in transcriptomic studies. Our method can be applied to any organism and to any type of experimental conditions, and can easily be implemented or modified in R. This tool provides an alternative to spike-in controls and represents an improvement over pre-defined reference genes which may not be stably expressed in one's own experimental conditions.

Methods

Initial *Arabidopsis thaliana* Columbia-0 were obtained from *Arabidopsis* Biological Resources Center (ABRC). *Arabidopsis* transgenic plants expressing GFP alone (Control) or fused to a candidate secreted effector protein of the fungus *Melampsora larici-populina* (Mlp37347 or Mlp124499), obtained in our laboratory [31], were used for the transcriptome analysis.

RNA was extracted from pooled aerial tissue of 2-week-old soil-grown plants, doing four replicates per genotype, with the Plant Total RNA Mini Kit (Geneaid) using RB buffer following manufacturer's protocol. The samples were treated with DNAse, then RNA quality was assessed using agarose gel electrophoresis. Libraries were generated with the NeoPrep Library Prep System (Illumina) using the TruSeq Stranded mRNA Library Prep kit (Illumina) and 100 ng of total RNA following manufacturer's recommendations. The libraries were then sequenced with Illumina HiSeq 4000 Sequencer paired-end reads of 100 nt.

Libraries were trimmed using Trimmomatic [33] (LEADING:4 TRAILING:4 SLIDINGWINDOW:4:20 MINLEN:20) and then the surviving paired reads were aligned to the TAIR10 assembly of the genome of *A. thaliana* with TopHat v2.0.14 [34] in Galaxy [35] (default options, with average mate inner distance varying for each replicate (Additional file 2: Table S6) and standard deviation of mate inner distance of 50 base pairs). The general information of the sequencing results and mapping data is presented in Additional file 2: Table S6, the dataset was deposited in NCBI under BioProject PRJNA528094. Further analyses were done using R software v.3.2.5. Genomic ranges of *Arabidopsis* transcripts were obtained from Ensembl plants [36] with GenomicFeatures and overlaps of sequencing reads with the transcripts were counted using GenomicAlignments [37], using options for paired-end reads and union mode.

We transformed the counts into TPM [27] and calculated the cutoff for active genes with DAFS [28]. We considered as reference the 0.5% of the active genes with the lowest covariance (R package "CustomSelection" [29]). Next, we used DESeq2 [38] to confirm that the selected genes were not deregulated. Finally, we used geNorm [23] and NormFinder [24] to compare the custom selected reference genes against three sets of genes (a list of 14 commonly used housekeeping reference genes (Table 1), the reference genes selected by T Czechowski, M Stitt, T Altmann, MK Udvardi and W-R Scheible [26] and the 104 reference genes selected by B Zhuo, S Emerson, JH Chang and Y Di [11]), using TPM values for the expression levels.

Description of the R-package. This package has 4 functions, "Counts_to_tpm" (to convert read counts into TPM values using a named vector with gene lengths) and the read count data frame with the samples as the column names and the genes as row names, "DAFS" (uses the data frame of TPM values, first object of the result from "Counts_to_tpm" to get the threshold for expressed genes), "gene_selection" (uses the data frame of TPM and the result from "DAFS" output a data frame with the selected reference genes, their average TPM and the covariance of the TPM values) and "customReferences" (calculates internally "Counts_to_tpm", "DAFS" and "gene_selection" outputs the result from "gene_selection"). The package also includes to datasets for testing: a data frame of counts created with the data used in this article and a named vector with the lengths of genes from Arabidopsis. A Wiki, which is the file README.md of this package, describes a workflow to get the read counts from raw read files.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12864-019-6426-2.

Additional file 1. Covariance level for each of the 30 genes selected by T Czechowski, M Stitt, T Altmann, MK Udvardi and W-R Scheible [26] for each permutation (A: Mlp37347 vs Control; B: Mlp124499 vs Control; C: Mlp124499 vs Mlp37347).

Additional file 2: Table S1. TAIR IDs of custom selected references for each transcriptome permutation. **Table S2.** DESeq2 results summary of analysis without reference genes or with different reference sets (Custom selected, from T Czechowski, M Stitt, T Altmann, MK Udvardi and W-R Scheible [26], from B Zhuo, S Emerson, JH Chang and Y Di [11] or Commonly used references). Table presents the number of genes found up- and down-regulated in **Table S3** to **S5.** Summary of the results of several analyses for all the genes evaluated in this article: Column A: TAIR ID; Column B: ranking calculated with geNorm with the function "selectHKs" from the R package "NormqPCR"; Column C: average TPM value; Column D: covariance of the TPM values; Column E: the difference of expression of a gene between two samples calculated with NormFinder; Column F: the common standard deviation of the expression of a gene between two samples calculated with NormFinder; Column G: stability measure from NormFinder; Column H: log2-transformed fold change of each gene calculated with DESeq2 without using reference genes; Column I: adjusted *p* value of the gene deregulation calculated with DESeq2 without using reference genes; Column J: sources that identified the gene as a reference, when more than one source selected the gene as reference they are separated by a ";". **Table S3.** Permutation Mlp37347 vs Control; **Table S4.** Permutation Mlp124499 vs Control; **Table S5.** Permutation Mlp124499 vs Mlp37347. **Table S6.** Metadata of samples used; replicate identification, number of sequenced reads, average length of the separation between two paired reads, number of reads after trimming and filtering and number of aligned reads for each of the 4 replicates of the three samples used in this study.

Abbreviations

ABRC: *Arabidopsis* Biological Resources Center; ACT2: Actin 2; ACT7: Actin 7; ACT8: Actin 8; APT1: Adenine phosphoribosyltransferase 1; DAFS: Data-Adaptive Flag Method for RNA-Sequencing Data; EF1α: Elongation factor 1-α; eIF4A: Eukaryotic translation initiation factor 4A-1; FPKM: Fragments per Kilobase per Million; GFP: Green Fluorescent Protein; mRNA: messenger Ribonucleic Acid; NDUFA8: Nicotinamide adenine dinucleotide-ubiquinone oxidoreductase 19-kDa subunit; qPCR: quantitative Polymerase Chain Reaction; RNA: Ribonucleic Acid; RNAseq: RNA sequencing; RPKM: Reads per Kilobase per Million; RT-qPCR: Reverse Transcription quantitative Polymerase Chain Reaction; TPM: Transcripts per Million; TUB2: Tubulin β-2/β-3 chain; TUB6: β-Tubulin 6; TUB9: Tubulin β-9 chain; UBQ10: Polyubiquitin; UBQ11: Polyubiquitin; UBQ4: Polyubiquitin; UBQ5: Ubiquitin extension protein

Acknowledgements

We thank Melodie B. Plourde for revising the manuscript.

Authors' contributions

KCGS, IDP and HG designed the work; KCGS performed the experiments; KCGS and HG wrote the paper; IDP and HG revised the paper and all authors approved the manuscript.

Funding

Funding for the project was provided by Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grants to HG. The project in HG's laboratory was also partially funded by an institutional Research Chair and a Canada Research Chair held by HG and a Canada Research Chair held by IDP. KCGS was funded by a master's scholarship from the Fondation de l'Université du Québec à Trois-Rivières, an international PhD scholarship from the Fonds de Recherche du Québec sur la Nature et les Technologies (FRQNT) and a graduate fellowship from MITACS.

Availability of data and materials

The dataset used herein was deposited in NCBI-SRA under BioProject PRJNA528094.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 19 March 2019 Accepted: 24 December 2019
Published online: 10 January 2020

References

1. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. Cell. 2008;133(3):536.
2. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320:1349.
3. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453:1245.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5(7):628.
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):63.
6. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.
7. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131:285.
8. Pachter L. Models for transcript quantification from RNA-seq. arXiv preprint. 2011;arXiv:1104.3889.
9. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25.
10. Wolf JBW. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Mol Ecol Resour. 2013;13(4):572.
11. Zhuo B, Emerson S, Chang JH, Di Y. Identifying stably expressed genes from multiple RNA-Seq data sets. PeerJ. 2016;4:e2791.
12. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018;19:792.
13. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14:R95.
14. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14(1):91.
15. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. Revisiting global gene expression analysis. Cell. 2012;151(October):482.
16. Lutzmayer S, Enugutti B, Nodine MD. Novel small RNA spike-in oligonucleotides enable absolute normalization of small RNA-Seq data. Nat Sci Rep. 2017;7:5913.
17. Taruttis F, Feist M, Schwarzfischer P, Gronwald W, Kube D, Spang R, Engelmann JC. External calibration with *Drosophila* whole-cell spike-ins delivers absolute mRNA fold changes from human RNA-Seq and qPCR data. BioTechniques. 2018;62(2):61.
18. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):902.
19. Paepe KD. Comparison of methods for differential gene expression using RNA-seq data. *Dissertation*. Gand: Universiteit Gent; 2015.
20. Qing T, Yu Y, Du T, Shi L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci. 2013;56(2):142.
21. Gutierrez L, Mauriat M, Guénin S, Pelloux J, Lefebvre JF, Louvet R, Rusterucci C, Moritz T, Guerineau F, Bellini C, et al. The lack of a systematic validation

of reference genes: a serious pitfall undervalued in reverse transcription-polymerase chain reaction (RT-PCR) analysis in plants. Plant Biotechnol J. 2008;6(6):618.
22. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L, Widmayer P, Gruissem W, Zimmermann P. Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinforma. 2008; 2008:420747.
23. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. 2002;3(7):research0034.0011.
24. Andersen CL, Ledet-Jensen J, Ørntoft T. Normalization of real-time quantitative RT-PCR data: a model based variance estimation approach to identify genes suited for normalization - applied to bladder- and colon-cancer data-sets. Cancer Res. 2004;64:5250.
25. Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: BestKeeper – excel-based tool using pair-wise correlations. Biotechnol Lett. 2004;26(6):515.
26. Czechowski T, Stitt M, Altmann T, Udvardi MK, Scheible W-R. Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. Plant Physiol. 2005;139(1):17.
27. Counts_to_tpm.R. https://gist.github.com/slowkow/c6ab0348747f86e2748b/ea6b1a870ca99e68717a22b8cf78ab35e642f0ec. Accessed 21 Nov 2018.
28. George NI, Chang C-W. DAFS: a data-adaptive flag method for RNA-sequencing data to differentiate genes with low and high expression. BMC Bioinformatics. 2014;15:92.
29. Santos KCGD, Desgagné-Pénix I, Germain H. CustomSelection: Custom selected reference genes outperform pre-defined reference genes in transcriptomic analysis. In: This package calculates the Transcripts Per Million data frame from the counts matrix, calculates the minimum expresion level for a gene to be considered expressed in each sample and selects as reference genes those with lowest covariance; 2019.
30. Love MI, Anders S, Hu W. Differential analysis of count data – the DESeq2 package. Genome Biol. 2014;15(550):63.
31. Germain H, Joly DL, Mireault C, Letanneur C, Stewart D, Morency MJ, Petre B, Duplessis S, Séguin A. Infection assays in *Arabidopsis* reveal candidate effectors from the poplar rust fungus that promote susceptibility to bacteria and oomycete pathogens. Mol Plant Pathol. 2018;19:200.
32. Hart T, Komori HK, LaMere S, Podshivalova K, Salomon DR. Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics. 2013;14(1):778.
33. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2120.
34. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36.
35. Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Eberhard C, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44(W1):W10.
36. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W. BioMart and bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3440.
37. Lawrence GJ, Huber MLW, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey VJ. Software for computing and annotating genomic ranges. PLoS Comput Biol. 2013;9:e1003118.
38. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

Publisher's Note