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

Karen Cristine Gonçalves Dos Santos
Universite du Quebec a Trois-Rivieres
Isabel Desgagné-Penix
Universite du Quebec a Trois-Rivieres
Hugo Germain (hugo.germain@uqtr.ca)
Universite du Quebec a Trois-Rivieres https://orcid.org/0000-0002-7046-6194

Methodology article

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

Posted Date: May 14th, 2019

DOI: https://doi.org/10.21203/rs.2.9587/v1

Version of Record: A version of this preprint was published on January 10th, 2020. See the published version at https://doi.org/10.1186/s12864-019-6426-2.

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 script 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 DAFT 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 stable. geNorm produced a similar result in which most custom selected genes ranked higher (i.e. were more stable) 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.

Background

RNAseq is a technique used since the pioneer studies of Lister et al. [1] (*Arabidopsis thaliana*), Nagalakshmi et al. [2] (*Saccharomyces cerevisiae*), Wilhelm et al. [3] (*Schizosaccharomyces pombe*), and Mortazavi et al. [4] (*Mus musculus*). This technique allows the combination of transcripts discovery and expression levels quantification in a single assay and has an unlimited dynamic range of detection compared to microarrays 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, 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. This is because one differentially expressed gene shifts the sequencing effort distributed to the others and all genes appear to be differentially expressed [7-9]. 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 [8].

The comparison of different softwares for RNAseq analysis is a recurrent subject in the literature [10-12] 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 then 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 [13-15]. However, reports have shown mixed performances with different normalization methods [16], resulting in high false discovery rates and false positive rates [17]. These may show differences in amplification depending on the type of tissue studied or the protocol for mRNA enrichment [18].

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 [19]. 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, Hruz et al. [20]) or a set of pre-selected candidate reference genes (geNorm, Vandesompele et al. [21]; NormFinder, Andersen et al. [22]; BestKeeper, Pfaffl et al. [23]). The most frequent approach is to take previously identified stable genes, as done by Zhuo et al. [9] this however does not ensure that the selected genes will be stably expressed 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. Our method alleviates the problem inherent to predefined reference genes, which may not be stable across experimental set-ups, are applicable to a single species and does not necessitate spike-ins.

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 [26], 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 100nt.

Libraries were trimmed using Trimmomatic [28]and then aligned to the TAIR10 assembly of the genome of *A. thaliana* with TopHat v2.0.14 [29] in Galaxy [30]. The general information of the sequencing results and mapping data is presented in Additional file 3, the dataset was deposited in NCBI-SRA under BioProject PRJNA528094. Further analyses were done using R software v.3.2.5. Genomic ranges of *Arabidopsis* transcripts were obtained from Ensembl plants [31] with GenomicFeatures and overlaps of sequencing reads with the transcripts were counted using GenomicAlignments [32], using options for paired-end reads and union mode.

We transformed the counts into TPM [24] and calculated the cutoff for active genes with DAFS [25]. We considered as reference the 0.5% of the active genes with the lowest covariance (R script in Additional file 1). Next, we used DESeq2 [33] to confirm that the selected genes were not deregulated. Finally, we compared the custom selected reference genes against a list of 14 commonly used housekeeping reference genes (Table 1) using NormFinder [22] and geNorm [21], using TPM values for the expression levels.

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 stable *Arabidopsis* genes proposed by Zhuo et al. [9]. The first set of reference genes was assessed for stability in three different permutations of the transcriptome sets as shown in Figure 1A (panel 1: Control vs Mlp37347, panel 2: Control vs Mlp 124499, panel 3: Mlp37347 vs Mlp124499). In each case, high levels of covariance, ranging from 4.4% (NDUFA8 in Control vs Mlp37347) to 178% (eIF4A in Control vs Mlp124499) were obtained. Next, we performed the same analysis using the 104 genes proposed by Zhuo et al. [9]. For the three permutations of the transcriptome sets, important fluctuations in the covariance were observed ranging from 2.3% to 48% (Figure 1B). These results demonstrate that neither the commonly used reference genes presented in Table 1 nor the 104 stable reference genes proposed by Zhuo et al. [9] were stable genes 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 [24] and calculate the average TPM and covariance for each gene. We then used the DAFS function [25] 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 script in Additional file 1). This pipeline is thereafter referred to as the custom selection script.

To test the developed method, we used the same transcriptome sets described in Figure 1 (the list of selected genes for each analysis is available in Additional file 2). For each transcriptome set, Figure 2 displays a comparison in $\log_2$ Fold Changes (left) and a $\log_2$ average TPM comparison (right) between the genes selected using the custom selection script (Custom References) and the Common Reference Genes (Table 1). In all pairings the custom selected reference genes show lower $\log_2$ Fold Change (Figure 2A, B, and C), broader range of expression levels and lower covariance (Figure 2D, E, and F) than the commonly used reference genes.

To further test the stability of the custom reference genes in our experiment, we used NormFinder [22] and geNorm [21] to compare commonly used control genes and custom selected reference genes from the proposed pipeline using $\log_2$ transformed TPM values. Both sets of genes were under the stability threshold of NormFinder (0.5), meaning that the software considers them suitable references genes, however the custom selected genes (shown in black) were more stable than the commonly used genes (shown in red) (Figure 3). 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 [13-15], yet the methods for the selection of these genes are designed for qPCR data and require a set of preselected reference or target genes or the selection of conditions similar to that of one's own experiment [20-23], 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 [26], 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-based function which enables the selection of custom reference genes regardless of the organisms used or of the experimental conditions.

The method developed here only requires information available from the RNAseq analyses. It uses Transcripts per Million as a proxy for the expression level and the DAFS algorithm [25] to exclude genes with low counts, which may be inactive [27]. We first assessed whether the most commonly used reference genes (Table 1) or a set a

published stable reference genes for *Arabidopsis* [9] were indeed stable in our experimental conditions. As demonstrated in Figure 1, both sets of reference genes show a high level of covariance 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 causes the loss of some differentially expressed genes which show modest variation in gene expression [19]. 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 Figure 2, in all the pairings of the data used, the custom selected reference genes outperformed the common reference genes in their stability of expression, presenting lower fold changes and lower covariances. Our method also allows the selection of more reference genes (the final number is user defined), giving more reference points, hence more robustness, to the normalization of genes expressed at different levels.

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 stable in one's own experimental conditions.

Abbreviations

RNA: Ribonucleic Acid

RT-qPCR: Reverse Transcription quantitative Polymerase Chain Reaction

TPM: Transcripts per Million

DAFS: Data-Adaptive Flag Method for RNA-Sequencing Data

GFP: Green Fluorescent Protein

RPKM: Reads per Kilobase per Million

FPKM: Fragments per Kilobase per Million

RNAseq: RNA sequencing

mRNA: messenger Ribonucleic Acid

qPCR: quantitative Polymerase Chain Reaction

ACT2: Actin 2

ACT7: Actin 7

ACT8: Actin 8

APT1: Adenine phosphoribosyltransferase 1

EF1α: Elongation factor 1-α

eIF4A: Eukaryotic translation initiation factor 4A-1

NDUFA8: Nicotinamide adenine dinucleotide-ubiquinone oxidoreductase 19-kiloDalton subunit

TUB2: Tubulin β-2/β-3 chain

TUB6: β-Tubulin 6

TUB9: Tubulin β-9 chain

UBQ4: Polyubiquitin

UBQ5: Ubiquitin extension protein

UBQ10: Polyubiquitin

UBQ11: Polyubiquitin

ABRC: *Arabidopsis* Biological Resources Center

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

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

Competing interests

The authors declare that they have no competing interests.

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.

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.

Acknowledgements

We thank Melodie B. Plourde for revising the manuscript.

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):523-536.
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; 320:1344-1349.
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; 453:1239-1245.
4. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods 2008; 5(7):621-628.
5. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 2009; 10(1):57-63.
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. Journal. 2014; doi:10.1371/journal.pone.0078644.
7. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Journal. 2010; doi:10.1186/gb-2010-11-3-r25.
8. Wolf JBW. Principles of transcriptome analysis and gene expression quantification: an RNA-seq tutorial. Molecular ecology resources 2013; 13(4):559-572.
9. Zhuo B, Emerson S, Chang JH, Di Y. Identifying stably expressed genes from multiple RNA-Seq data sets. Journal. 2016; doi:10.7717/peerj.2791.
10. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Briefings in Bioinformatics 2018; 19(5):776-792.
11. 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. Journal. 2013; doi:10.1186/gb-2013-14-9-r95.
12. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. Journal. 2013; doi:10.1186/1471-2105-14-91.
13. 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(3):476-482.
14. Lutzmayer S, Enugutti B, Nodine MD. Novel small RNA spike-in oligonucleotides enable absolute normalization of small RNA-Seq data. Journal. 2017; doi:10.1038/s41598-017-06174-3.
15. 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):53-61.

16. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nature Biotechnology 2014; 32(9):896-902.
17. Paepe KD. Comparison of methods for differential gene expression using RNA-seq data. Dissertation. Gand: Universiteit Gent; 2015.
18. 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. Science China Life Sciences 2013; 56(2):134-142.
19. 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 Biotechnology Journal 2008; 6(6):609-618.
20. 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. Journal. 2008; doi:10.1155/2008/420747.
21. 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. Journal. 2002; doi:10.1186/gb-2002-3-7-research0034.
22. 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 Research 2004; 64:5245-5250.
23. 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. Biotechnology letters 2004; 26(6):509-515.
24. Slowikowski K. Counts_to_tpm.R [https://gist.github.com/slowkow/c6ab0348747f86e2748b/ea6b1a870ca99e68717a22b8cf78ab35e642f0ec]. 2016. Accessed 21-11-2018.
25. 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.
26. 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. Molecular plant pathology 2018; *19*:191-200.
27. 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.
28. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30(15):2114-2120.
29. 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. Journal. 2013; doi:10.1186/gb-2013-14-4-r36.
30. 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. Journal. 2016; doi:10.1093/nar/gkw343.
31. 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):3439-

3440.

32. Huber MLW, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan M, Carey VJ. Software for computing and annotating genomic ranges. Journal. 2013; doi:10.1371/journal.pcbi.1003118.
33. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Journal. 2014; doi:10.1186/s13059-014-0550-8.

Tables

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	elF4A	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

Figures

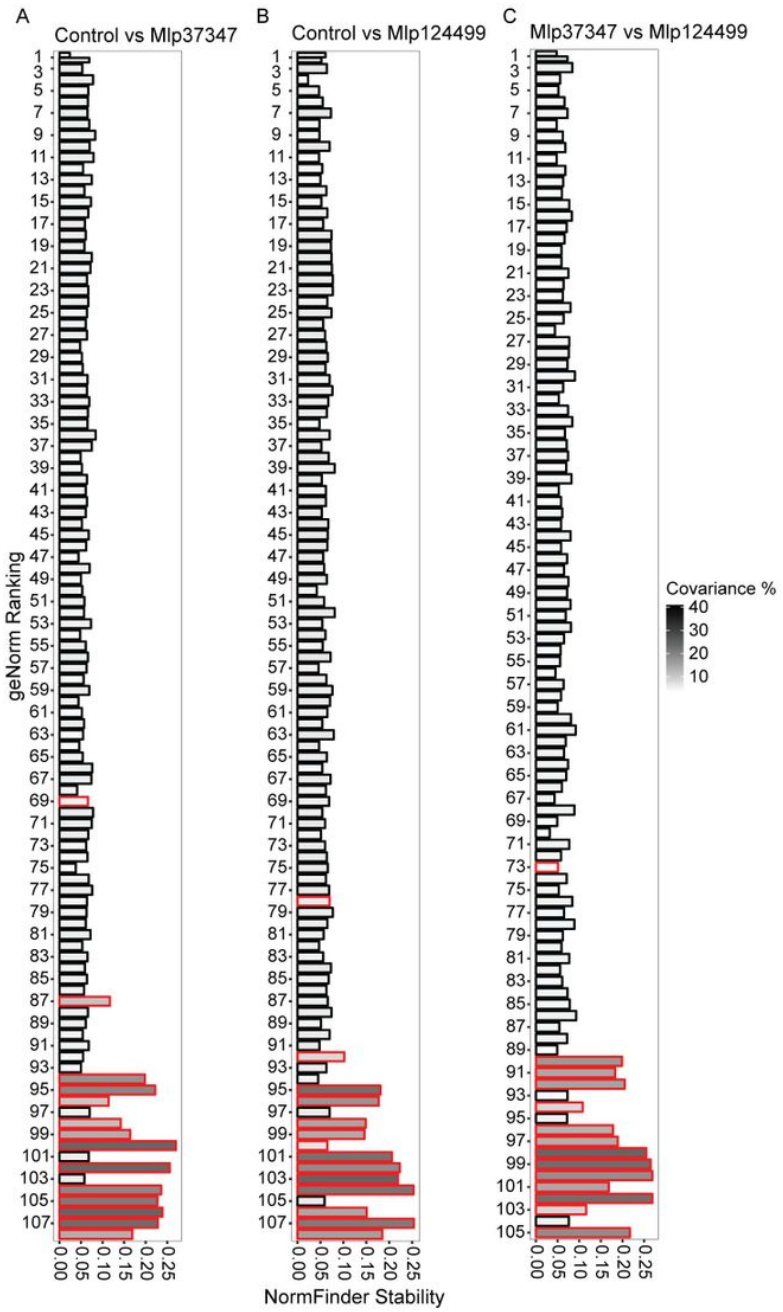

Figure 1

Evaluation of covariance distribution the three transcriptome data sets. A) among a set of 14 commonly used reference genes B) a set of 104 reference genes proposed by Zhuo et al. [9].

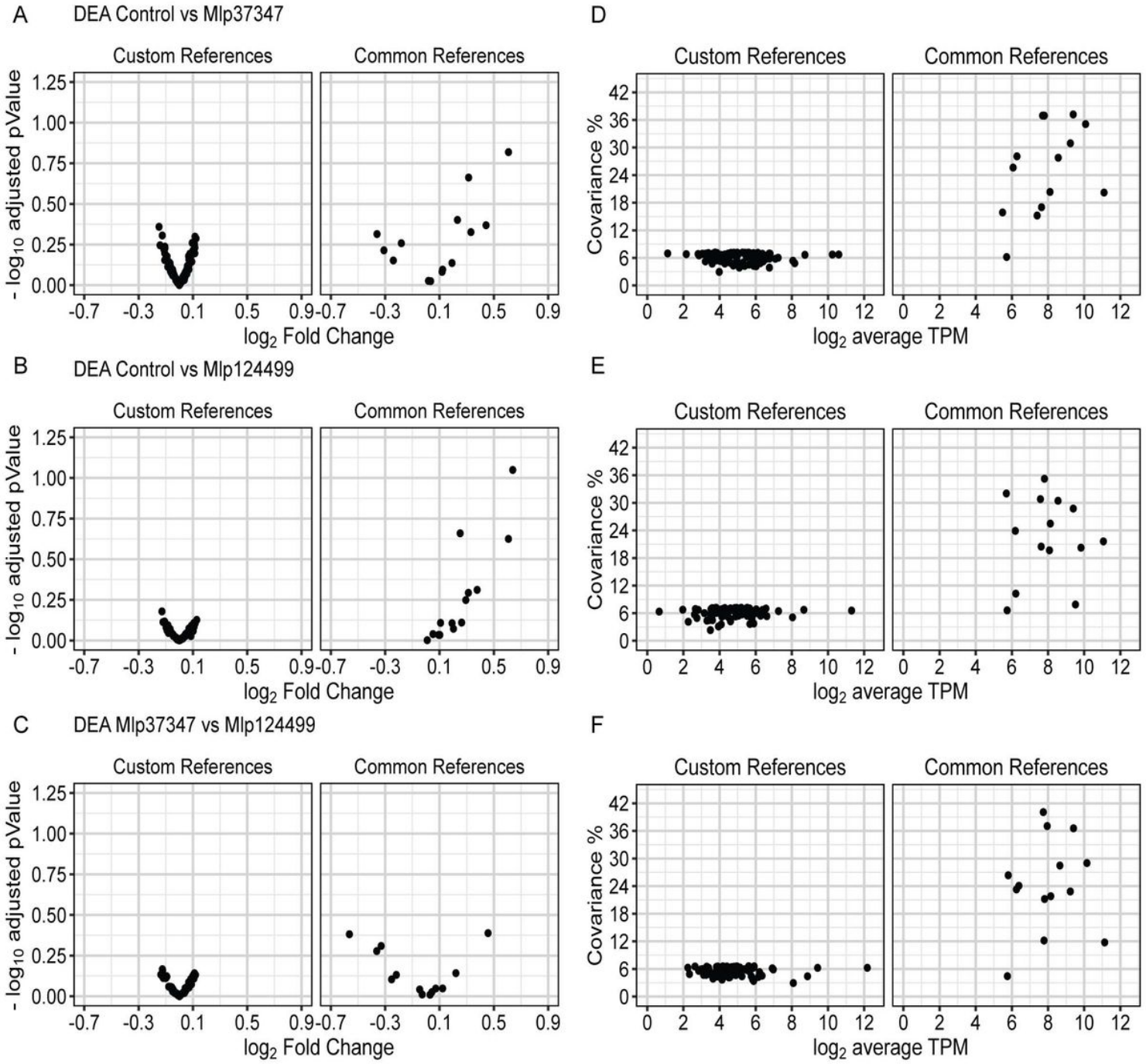

Figure 2

Comparison of custom selected reference genes with commonly used reference genes for three sets of data. A, B and C show the distribution of log2 Fold Change by -log10 adjusted pValue for A) Control vs Mlp37347, B) Control vs Mlp124499 and C) Mlp37347 vs Mlp124499. D, E and F depict the distribution of log2 average TPM values by covariance for D) Control vs Mlp37347, E) Control vs Mlp124499 and F) Mlp37347 vs Mlp124499

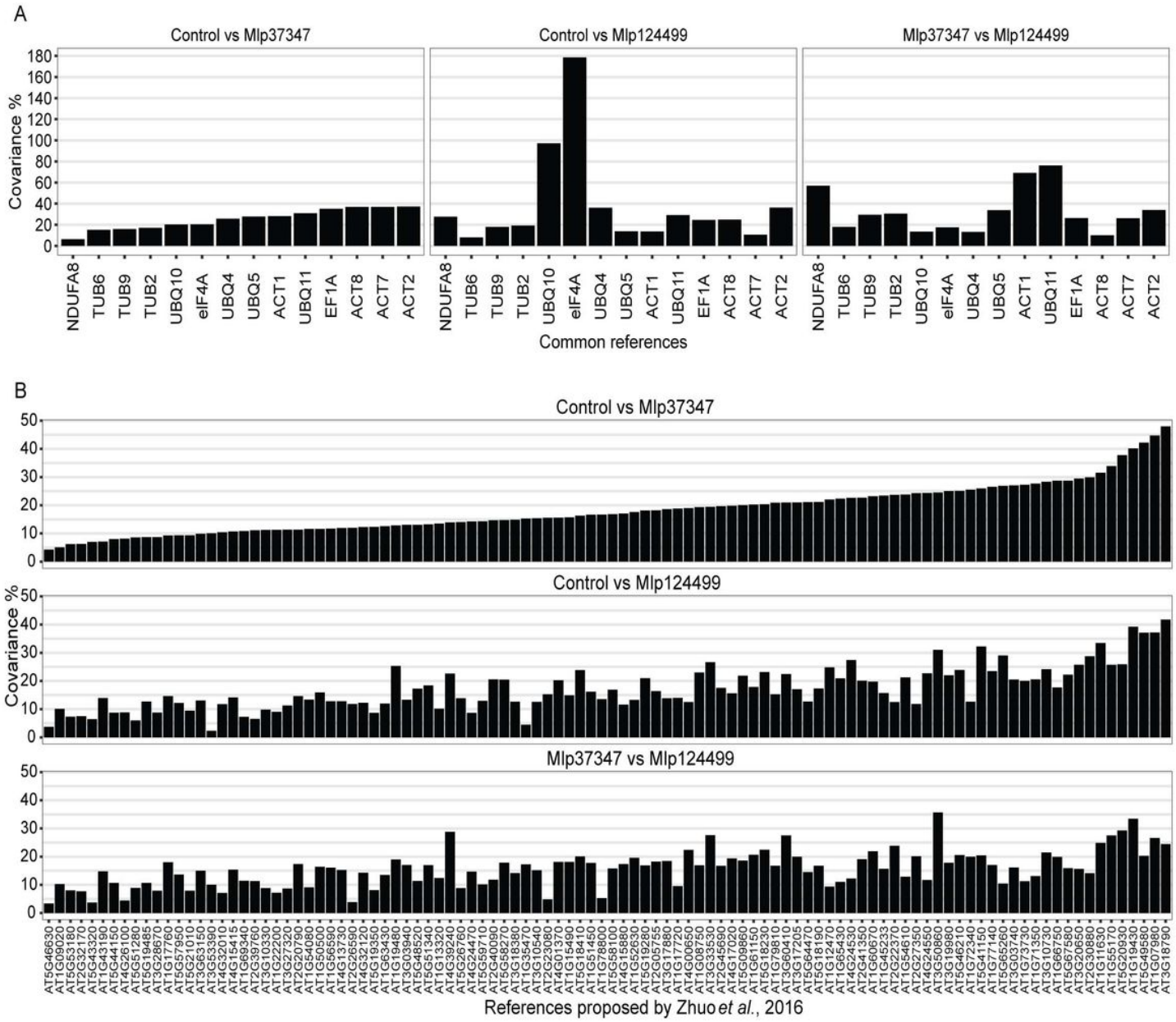

Figure 3

Comparison of custom selected reference genes (black border) and commonly used reference genes (red border) with geNorm ranking, NormFinder stability index and covariance for three sets of data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplement1.r
- supplement1.docx
- supplement3.docx