A Novel, Improved, Application for the Normalization of RNA-seq Expression Data in Complex Polyploids

Dyfed Lloyd Evans[1,2]

[1]South African Sugarcane Research Institute, 170 Flanders Drive, Mount Edgecombe, Durban, South Africa, 4302

[2]Cambridge Sequence Services (CSS), Waterbeach, Cambridge, CB2 9TB, UK5

Abstract

Much of the work on the normalization of RNA-seq data has been performed on human, notably cancer tissue. Little work has been done in plants, particularly polyploids and those species with incomplete or no genomes. We present a novel implementation of GeTMM (Gene Length Corrected TMM) that accounts for GC bias and works at the transcript level. The algorithm also employs transcript length as a factor, allowing for incomplete transcripts and alternate transcripts. This significantly improves overall normalization. The GCGeTMM methodology also allows for simultaneous determination of differentially expressed transcripts (and by extension genes) and stably expressed genes to act as references for qRT-PCR and microarray analyses.

Introduction

As RNA-seq analysis has come to dominate transcriptomic analyses (a switch from the previous front-runner of microarrays) due to its massively sequencing of cDNA (Mortazavi et al. 2008). RNA-seq also allows for the study of novel transcripts along with a better range of detection and lower technical variability (Zhao et al 2014) and provides a high degree of agreement with the currently accepted 'gold standard' in transcriptomic expression, qRT-PCR at both the absolute and relative expression analysis levels (Su and Mason 2014). A typical RNA-seq experiment involves several steps to enable data analysis: trimming (Williams et a1 2016), alignment (mapping) (Borozan et al. 2013; Yang et al. 2015), read counting, data normalization and analysis (Lin et al. 2015; Li et al. 2017).

As much of the developmental work in RNA-seq analyses derives from the human genome, sequence reads are aligned to a reference genome, and the number of reads mapping to that feature are proportional to the length and abundance of the feature — with the 'gene' feature being a surrogate for all the transcripts transcribed from that gene.

Little work has been done on error correction of RNA-seq data for the discovery of stably-expressed transcripts/genes and differentially expressed genes in plants, particularly complex polyploids though there are some whole genome analyses (Park et al 2019; Gupta et al 2012). However, most of the developmental work in this area has been performed on human data, particularly comparing cancer and normal tissue datasets. Typical normalization methods for RNA-seq data typically allow for either intersample comparisons (for differentially expressed genes) or intrasample comparison (for discovery and/or validation of gene signatures) (Smid et al. 2018).

For other organisms, where no genome is available a transcriptome assembly may be substituted for the complete genome. Where both the genome and transcriptome are incomplete reads may be mapped g to a subset that only includes the transcripts/genes of interest (Peri et al 2020). However, as depth of RNA-seq data can vary, normalization has to be performed to correct for differences between sequencing runs (e.g. library size and relative abundances) prior to any downstream analyses.

The most commonly used RNA-seq normalization methods are TMM, as implemented in edgeR (Robinson et al. 2008) and RLE, implemented in DESeq2 (Anders and Huber 2010; Love et al. 2014). However, neither of these methods employ any gene length normalization (their aim being to identify differentially-expressed genes between samples and thus they assume that the gene length is constant across samples. TPM (Transcripts Per kilobase Million) normalization (Li et al. 2008) extends the previously used RPKM (Reads Per Kilobase per Million reads) for single-end sequencing protocols (Mortazavi et al 2008) and its paired-end counterpart, FPKM (Fragments Per Kilobase per Million reads) (Trapnell et al. 2008), as both RPKM and FPKM proved to be inadequate and biased (Bullard et a1. 2010; Olshack et al, 2009; Wagner et al 2012). TPM employs a simple normalization scheme, where the raw read counts of each gene are divided by the gene length in kb and the total sum of all RPK is considered the library size of that sample. Thus TPM can be used for fully-elucidated genomes and for partial transcriptomes. Finally, the library size is divided by a million, and that number is employed as the scaling factor to scale each genes' RPK value.

Under ideal conditions, a normalization methodology should account for all the major sources of error in a sample and should yield a dataset on which both between-sample and within-sample analyses can be

performed. Smid et al. (2018) aimed to do this in their implementation of GeTMM (Gene length corrected trimmed mean of M-values) which combines gene-length correction with the normalization procedure TMM. GeTMM performs similarly TPM in intersample analyses but has clear advantages in intrasample comparisons (Smid et al. 2018).

Recent studies have shown that, in plants, gene expression is highly tissue specific and often varies with tissue age and the physiological status of the plant and the exact experimental conditions and, to date, there has not been any clear report of universal reference genes (Kozera and Rapacz 2013; Joseph et al 2018; Hong et al. 2008; Gutierrez et al. 2010). Coupled with recent findings that commonly-employed housekeeping genes may be far more variable in their expression than previously realized. Historically, selection of reference genes for qPCR studies has typically been arbitrary, with genes such as *25S* and *18S* rRNAs, *GAPDH*, and *Actin* commonly being selected without experimental validation (this being true for both plant and animal studies). In concert, these genes were often employed with the assumption that they are stably expressed across tissues. However, we now know that in many instances these commonly used RGs exhibit tissue and treatment specific variability (Chari et al., 2010; De Jonge et al., 2007). A previous preliminary study on a number of human cell lines and tumour versus matched normal tissue samples demonstrated that inappropriate choice of RGs may lead to errors when interpreting experiments involving quantification of gene expression (Janssens et al. 2004).

In an attempt to correct for over-expressed genes with variable expression edgeR employs the Trimmed Means of M-values (TMM) (Robinson and Olshack 2010) in which highly expressed genes and those that have a large variation of expression are excluded, whereupon a weighted average of the subset of genes is used to calculate a normalization factor. In the edgeR implementation precision (inverse of variance) weights are used to account for the fact that log fold changes from genes with higher read counts have lower variance on the logarithm scale. This typically excludes very highly expressed transcripts such as 25S and 18S ribosomal RNAs along with other very highly expressed transcripts from the initial transcript pool.

Many normalization methodologies assume that, depending on whether genes or transcripts are the base unit, that the length of this base unit is the same between samples. In diploids, this can be assumed to be correct at the gene level. However in organisms with high ploidy allelic variants of genes make this unlikely. If t he experiment is being performed at the transcript level, due to alternate transcripts (particularly tissue specific alternate transcripts) the assumption does not hold at all. Thus corrections for transcript/gene lengths are required.

Panicum virgatum (switchgrass) is a tall, upright, bunchgrass that is a feature of North American prairies. *Panicum virgatum* is seen as potentially being an important bioenergy crop. It is an outcropping species that is an hybrid of two ancestral species (Lovell et al. 2021). Tetraploid P. virgatum plants have variously hybridized to generate disparate octoploid forms (Triplett et al 2012).

We extend the GeTMM implementation by removing over and under-expressed transcripts with edgeR and develop a novel PERL implementation to apply effective transcript length (including alternate transcript variants), GC skew and library lengths as additional normalization variables and apply this to public leaf RNA-seq datasets in *Panicum virgatum* cultivars. Being highly polyploid and with a newly available high quality genome sequence (Lovell et al 2021) and numerous high depth RNA-seq datasets available, many with multiple replicates Panicum virgatum makes an excellent reference test species for the software.

Materials and Methods

Identification of Switchgrass Datasets

Analyses on algorithm efficiency were performed using the tetraploid *Panicum virgatum* AP13 v5.1 (Lovell et al. 2021) genome as a reference. All transcripts (not just primary transcripts) were exported from Phytozome v 13 (Goodstein et al. 2012). Two datasets with three replicates apiece were employed for the analyses of differential expression. These were *Panicum virgatum* cv Alamo leaf day 9: SRA accessions SRR12851488; SRR12851477; SRR12851469 and *Panicum virgatum* cv Alamo leaf 5 months: SRA accessions SRR6485351; SRR6485352; SRR6485353 (Chen et al 2020).

Read Pre-processing and Mapping

Prior to mapping, the reads to the reference transcripts polyA tails were manually clipped from transcripts (where they occurred).

The 9-day leaf was employed as control and the 5-month leaf samples were the test samples. For the SRA datasets, following Corchete et al. (2020) adapter removal only was performed with Trimmomatic 0.39 (LEADING:4 TRAILING:4 SLIDINGWINDOW:4:20MINLEN:50) (Bolger et al. 2014). However, instead of directly performing quality trimming with Trimmomatic reads were next error corrected with the error correction pipeline of SPAdes v 3.15.1 (Prjibelski et al. 2020). For paired end data, subsequent to SPAdes error correction the paired end data only was passed to Trimmomatic for quality trimming. For single end data all error corrected reads were employed as input for Trimmomatic.

Trimmed and error corrected reads were mapped to individual transcript sequences padded with Ns using HISAT2 v2.2.2.1 (Kim et al. 2019) mapped reads were enumerated with HTSeq v 0.12.4 (Anders et al. 2015) using a custom the Union approach. HISAT2's native SAM output format was piped to SAMtools (Danecek et al. 2021) and output in BAM format. Mapped reads were analyzed for completeness of the transcript and the presence of missing exons, before being employed as input to our novel normalization algorithm.

Data Export

Based on transcript sequences and read mappings, the following data were collected for input into the GeGC-TMM methodology: read lengths and insert sizes from the bam mapping file using samtools and picard tools; transcript sequence length and GC content using Emboss infoseq and geecee (Rice et al. 2000); count of mapped reads using Samtools.

Algorithm Implementation

GeGC-TMM methodology for normalizing RNA-seq data

Transcripts were trimmed of adapter sequences and low-quality sequence regions using Trimmomatic (ref). Trimmed sequences were error corrected using the error correction portion of the SPAdes (ref) assembler pipeline.

GC content is another major factor that requires normalization. Risso et al. (2011) demonstrated that full quantile normalization is the most appropriate approach. For this methodology genes are stratified according to GC-content, with the normalized expression measures defined as:

$$y_j = y_j - T(y_j : j' \in k(j) + T(y_1 \ldots . y_j)) \quad 1$$

where *k*(*j*) denotes the GC-content stratum to which gene *j* belongs and *T* denotes the upper-quantile for control genes. The quantiles of the read count distributions are then matched between GC-bins, by sorting counts within bins and then taking the median of quantiles across bins.

Prior to GeTMM normalization, GC normalization was performed in the Bioconductor R package EDASeq package (Risso et al. 2011) using the withinLaneNormalization and betweenLaneNormalization methods for inter-sample and intra-sample normalization, respectively.

Outputs from EDASeq were input into the GeGC-TMM application, which is described below:

For normalization, the overall methodology is based on the work of Smid et al. (2018) for data normalization and the work of Corchete et al. (2020) for the stability ranking of genes. Below is a full mathematical treatment:

Define Y_{gk} as the observed count of mapped raw reads for gene *g* in library *k*. μ_{gk} is the true and *unknown* expression level (total number of transcripts) and L_g is the length of gene *g* and N_k is the total number of reads in library *k*.

For sequencing data, the gene-wise log-fold changes are defied as:

$$M_g = \log_2\left(\frac{Y_{gk}/M_k}{Y_{gk'}/N_{k'}}\right) \quad (2)$$

the corresponding absolute expression levels are:

$$A_g = \frac{1}{2}\log_2\left(\frac{Y_{gk}}{N_k} \bullet \frac{Y_{gk'}}{N_{k'}}\right) \text{ for } Y_{g\cdot} \neq 0 \quad (3)$$

M and A values are employed to trim those genes/transcripts with too high (high expression and high variance) and too low (incomplete coverage). In our implementation we trimmed M_{gk}^r (sample k relative to sample r (where r is the reference set) for gene g) by 20% and the absolute expression, Ag, by 5%.

Thus, the final set of genes G* is a subset of the initial set G (ie $G \quad \in G$). The above is implemented in the edgeR package (Robinson and Oshlack 2010) and this was employed for the initial stages of analysis.

The normalization factor for a sample k relative to a sample r is obtained as:

$$\log_2(TMM_k^{(r)}) = \frac{\sum_{g\in G} \quad \omega_{gk}^r M_{gk}^r}{\sum_{g\in G} \quad \omega_{gk}^r} \quad (4)$$

Where:

$$M_{gk}^r = \frac{\log_2\left(\frac{Y_{gk}}{N_k}\right)}{\log_2\left(\frac{Y_{gr}}{N_r}\right)} \quad (5)$$

And:

$$\omega_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}} \quad (6)$$

There is an additional implicit trimming here as Y_{gk} and Y_{gr} must be greater than 0.

For GeTMM normalization, mapped read data are first converted to RPK (reads per kilobase) and are scaled with the TMM scale factor, as defined above.

Converting raw read counts to RPK is as simple as dividing a gene (or transcript's) raw read mapping count by the gene length in kilobases.

$$g_{RPK} = \frac{Y_{gk}}{L_g(kbp)} \quad 7$$

However, rather than the full length of the gene (or transcript of interest) it is better to use $\tilde{l}_g$, the Effective

Length of the feature of interest which is defined thus:

$$\tilde{l}_g = l_i - \mu_{FLD} + L \quad 8$$

l_i is the length of the feature of interest, μ_{FLD} is the mean of the fragment length distribution, as determined from the mapped reads and $\mathscr{L}$ is the sequence bias (if the mapping technique provides it). If $\mathscr{L}$ is not provided, it is typically set to 1. For species without a reference genome and only an incomplete transcriptome

RPK scale factor is given as:

$$RPKScaling(RPK_S) = \frac{\sum_g^n g_{RPK} \times 2^{\log_2(TMM_k^{(r)})}}{10^6} \quad 9$$

Where n is the total number of genes in G** (G** being the set of all genes after edgeR and GeTMM trimming).

The normalized read count for gene g thus becomes:

$$\tilde{g} = \frac{g_{RPK}}{RPK_S} \quad 10$$

Where $\tilde{g}$ represents the GeTMM normalized gene count.

Identification of Stably-expressed Genes

Corchete et al (2020) in their analysis of best in breed methodologies for RNA-seq based procedures for gene expression quantitative analyses advocated the use of coefficient of variance for ranking gene expression stability.

By their definition,

$$CoV = \frac{MAD}{med(g)} \quad 11$$

Where med(g) is the median value for gene 'g' and MAD is the median absolute deviation as given by:

$$MAD = med(|X_i - med(X)|) \quad 12$$

Where X is the xth gene and 'i' is the i'th sample and, in terms of algorithmic implementation:

$$|X_i - med(X)| = \sqrt{(X_i - med(X))^2} \quad 13$$

By calculating the CoV for each gene, the genes can be ranked in order of stability.

The full methodology and implementation methodology with how to run the code is given in the code itself (open source) and the accompanying documentation.

Differential Expression Analysis

After running the RNA-seq mapped data through the GeGC-TMM application fold change was estimated from the edgeR (Robinson et al. 2010) regression model fit. To compare the cumulative effects of the different normalization protocols each method was applied in sequence and compared with the results of the full pipeline.

Results

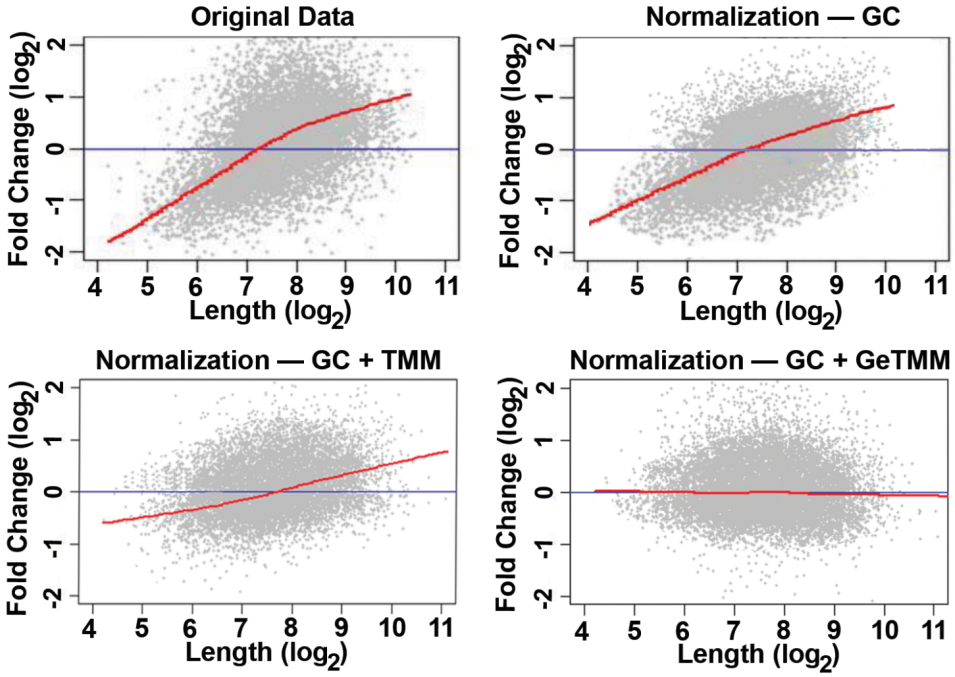
As the implementation of the GC-GeTMM is step-wise output from the individual steps can be exported for analysis. The RNA-seq data for nine day old and five month old plants were normalized independently. Each dataset had three replicates. Normally the normalized replicates would be averaged prior to determination of the coefficient of variation.

In this case, all datapoints were size sorted and the $\log_2$ of the length was determined. Fold change was determined with edgeR and $\log_2$ of fold change was plotted against $\log_2$ of transcript length (Figure 1). For each transcript length the mean fold change was plotted in red.

Each step in the normalization process (GC, TMM, GeTMM) yields improvements in the data quality as the midline curve (red in Figure 1) tends towards the X-axis.

Subsequent to normalization, the replicates were merged and CoV values were determined (Table 1).

Figure 1



Analysis of the effects of different normalization methodologies on the quality of differential expression studies for *Panicum virgatum* 9 day old and 5 month old leaves. Top left: comparison of fold change against transcript length for the original data. Top right: effect of GC normalization on data quality. Bottom left: effect of TMM normalization on the GC normalized data. Bottom right: effect of combining GC normalization with GeTMM normalization.

CoVs were determined for all the transcripts at the genome (apart from those transcripts that failed TMM analyses as being too highly and unstably expressed). Focussing at the transcript level also allows differentially expressed and orthologous transcripts from the two founding genomes of *P. virgatum* to be analyzed. The results of the analysis are presented in Table 1.

Table 1

Panicum Gene ID	Description	**CoV**
Pavir.7KG325700.1	PROLYL 4-HYDROXYLASE ALPHA SUBUNIT	0.05619045672458
Pavir.5NG379000.1	Ca2+-dependent phospholipid-binding protein	0.104374467
Pavir.6KG285300.1	4-coumarate--CoA ligase / 4-coumaryl-CoA synthetase	0.107080581736534
Pavir.2KG453700.1	translation initiation factor 1 (EIF1, SUI1)	0.112553429
Pavir.1NG433900.1	AP-2 complex subunit mu-1 (AP2M1)	0.119966947
Pavir.3NG140751.1	actin-related protein 1	0.122978337
Pavir.5KG391700.1	WD40 repeat-containing protein	0.12303473
Pavir.5KG148400.1	RNA polymerase II transcription mediators	0.125587916269135
Pavir.9KG227366.1	TIP41-like family protein	0.128201355417764
Pavir.2NG416900.1	CDK9 kinase-activating protein cyclin T	0.131326736530266
Pavir.1KG485600.5	Endomembrane protein 70 protein family	0.133528993616846
Pavir.9KG031100.1	DELLA protein (DELLA8)	0.133608997
Pavir.1NG511300.1	SERINE INCORPORATOR // SERINC-DOMAIN CONTAINING SERINE AND SPHINGOLIPID BIOSYNTHESIS PROTEIN	0.134423823005138
Pavir.2NG432600.1	Uncharacterized conserved protein	0.134894278622994
Pavir.4KG408400.1	rab geranylgeranyl transferase like protein	0.135144550730235
Pavir.5NG560301.1	SCY1-like protein 1 (SCYL1)	0.135881204515391
Pavir.1KG194700.1	Protein phosphatase 2A-2	0.136308640241443
Pavir.9KG032081.1	Cytochrome P450 CYP4/CYP19/CYP26 subfamilie	0.136568020956609
Pavir.5NG626600.1	Splicing factor U2AF, large subunit (RRM superfamily	0.137697229433054
Pavir.1KG097800.1	Uncharacterized conserved protein	0.139371804123243
Pavir.2NG477100.1	Serine/threonine protein kinase	0.140096820985049
Pavir.1NG540200.1	Kelch motif (Kelch_1) DCD (Development and Cell Death) domain protein	0.141345511421789
Pavir.6KG381900.1	ARM repeat superfamily protein	0.141650316655419
Pavir.5NG159400.3	m3G-cap-specific nuclear import receptor (Snurportin1	0.141921598182036
Pavir.5KG143700.1	Scd6-like Sm domain (LSM14)	0.142188152694473
Pavir.9NG382200.2	Pavir.9NG382200.2	0.142536947561398
Pavir.3KG496500.2	TBP-associated factor 8	0.143578790730331
Pavir.9KG020200.1	CDK inhibitor P21 binding protein	0.143948580886646
Pavir.9NG181500.1	elongation factor 1 alpha-like protein (HBS1)	0.145170087677981
Pavir.4KG212500.3	mediator of RNA polymerase II transcription subunit 6 (MED6)	0.146550633770169
Pavir.2KG279800.1	Putative u4/u6 small nuclear ribonucleoprotein	0.147489049610693
Pavir.5KG437500.1	Vacuolar sorting protein VPS36	0.147708960475672
Pavir.5KG548300.1	AAA-type Paste family protein	0.148404758593064
Pavir.2KG218200.1	mRNA cleavage and polyadenylation factor I/II complex, subunit Pcf11	0.148506533385334
Pavir.1KG100300.1	WI/SNF-related matrix-associated actin-dependent regulator of chrommatin	0.148535647781652
Pavir.7KG338200.1	Trypsin-like peptidase domain (Trypsin_2)	0.144855315650112
Pavir.1NG562600.1	GLUTATHIONE REDUCTASE, MITOCHONDRIAL	0.148836769
Pavir.9KG458870.1	TRANSMEMBRANE PROTEIN ADIPOCYTE-ASSOCIATED 1	0.149615523807001
Pavir.4KG109292.1	AAA-type ATPase family protein	0.14991651892742

Results of CoV analysis for the Panicum virgatum transcriptomes. Only those stably-expressed transcripts between the two tissues of interest with a CoV better than 15% are shown.

Discussion

The quest for stably expressed genes in multiple tissues and developmental stages is essential for the normalization of gene expression analyses by qRT-PCR and microarray studies. qRT-PCR is currently accepted as the gold standard for expression analyses, but is generally expensive and time-consuming. Microarray analyses are very high throughput but can suffer from issues of sensitivity. RNA-seq can afford a middle way. Many datasets are readily available from NCBI's sequence read archive (SRA) and individual RNA-seq datasets can be generated for a few thousand dollars.

RNA-seq allows rapid mapping of reads to transcripts and the quick ranking of stable transcripts. Thus the GCGe-TMM application was developed to extend the Ge-TMM to correct GC bias along with transcript length bias. The application presented in this paper can simultaneously error correct RNA-seq data for import into other applications for differential expression analysis as well as ranking transcripts in terms of covariant of expression.

Figure 1 demonstrates that the three steps of normalization employed (GC bias, TMM and length bias) all significantly improve the error profile of the data, making it suitable for further analyses. For whole genome analyconses (Table 1) a conservative cutoff of 15% was chosen (typical cutoffs range between 30 and 40% (refs). As the methodology for TMM excludes highly-expressed transcripts but unstably expressed transcripts. This excludes some of the common highly-expressed transcripts (25S rRNA, 16S rRNA, GAPDH, TATA binding protein (Thellin et al 1999; Vandesompele et al 2002).

Improved normalization leads to improved and more reliable differential expression analyses.

Conclusion

The GCGeTMM methodology presented in this paper affords improvements over the Ge-TMM implementation, particularly for monocot plants, with their generally higher GC content (Li and Du 2014).

The software described herein enables the simultaneous identification of stably expressed genes/transcripts and the identification of differentially-expressed transcripts at a whole genome and a gene/transcript level. It is applicable to species with complete genomes, but can also be used for those species without a genome (but with a transcriptome) as such it can be employed for orphan and under-funded species as it relies on transcript rather than gene level analyses. It can also be employed for subsets of transcripts (particularly for the analysis of stably expressed transcripts.

As such, the application presents a major step forwards for the analysis of stably expressed genes and differentially expressed genes in RNA-seq based gene expression analysis for plants, polyploids and those species with only partial or unsequenced genomes.

Competing Interests

The author declares there are no competing interests. However, for transparency DLlE is a co-founder and non-renumerated senior scientist of CSS a non-profit organization promulgating improvements in sequencing and sequence analysis.

Code and Data Availability

All data employed in this paper are publicly available. All code has been deposited in GitHub: https://github.com/gwydion1/expression.

References:

Anders S, Huber W, 2010. Differential expression analysis for sequence count data. *Nature Precedings*, https://doi.org/10.1038/npre.2010.4282.1

Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*, 31:166-169.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30:2114-2120.

Borozan I, Watt SN, Ferretti V. 2013. Evaluation of alignment algorithms for discovery and identification of pathogens using RNA-Seq. *PloS one*, 8:p.e76935.

Bullard JH, Purdom E, Hansen KD, Dudoit S. 2010. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics*, 11:1-13.

Chari R, Lonergan KM, Pikor LA, Coe BP, Zhu CQ, Chan TH, MacAulay CE, Tsao MS, Lam S, Ng RT, Lam WL. 2010. A sequence-based approach to identify reference genes for gene expression analysis. *BMC medical genomics*, 3:1-11.

Chen P, Chen J, Sun M, Yan H, Feng G, Wu B, Zhang X, Wang X, Huang L. 2020. Comparative transcriptome study of switchgrass (Panicum virgatum L.) homologous autopolyploid and its parental amphidiploid responding to consistent drought stress. *Biotechnology for biofuels*, 13:1-18.

Corchete LA, Rojas EA, Alonso-López D, De Las Rivas J, Gutiérrez NC, Burguillo FJ. 2020. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific reports*, 10:1-15.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience*, 10:p.giab008.

De Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, ter Elst A. 2007. Evidence based selection of housekeeping genes. *PloS one*, 2:p.e898.

Gupta R, Dewan I, Bharti R, Bhattacharya A. 2012. Differential expression analysis for RNA-Seq data. *International Scholarly Research Notices*, *2012*.

Gutierrez L, Mauriat M, Guénin S, Pelloux J, Lefebvre JF, Louvet R, Rusterucci C, Moritz T, Guerineau F, Bellini C, Van Wuytswinkel O. 2008. 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*, 6*:*609-618.

Hong S-Y, Seo PJ, Yang M-S, Xiang F, Park C-M. 2008. Exploring valid reference genes for gene expression studies in *Brachypodium distachyon* by real-time PCR. *Bmc Plant Biology* 8:1-11, https://doi.org/10.1186/1471-2229-8-112

Janssens N, Janicot M, Perera T. Bakker A. 2004. Housekeeping genes as internal standards in cancer research. *Molecular Diagnosis*, 8:107-113.

Joseph JT, Poolakkalody NJ, Shah JM. 2018. Plant reference genes for development and stress response studies. *Journal of biosciences*, 43:173-187.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology*, 37:907-915.

Kozera, B. and Rapacz, M., 2013. Reference genes in real-time PCR. *Journal of applied genetics*, *54*(4), pp.391-406.

Li XQ, Du D. 2014. Variation, evolution, and correlation analysis of C+ G content and genome or chromosome size in different kingdoms and phyla. *PLoS One*, 9:p.e88339.

Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. 2010. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26:493-500.

Li X, Brock GN, Rouchka EC, Cooper NG, Wu D, O'Toole TE, Gill RS, Eteleeb AM, O'Brien L, Rai SN. 2017. A comparison of per sample global scaling and per gene normalization methods for differential expression analysis of RNA-seq data. *PloS one*, 12:p.e0176185.

Lin Y, Golovnina K, Chen ZX, Lee HN, Negron YLS, Sultana H, Oliver B, Harbison ST. 2016. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. *BMC genomics*, 17:1-20.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15:1-21.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5:621-628.

Oshlack A, Wakefield MJ. 2009. Transcript length bias in RNA-seq data confounds systems biology. *Biology direct*, 4:1-10.

Park YS, Kim SK, Kim S, Kim KM, Ryu CM. 2019. The transcriptome analysis of the Arabidopsis thaliana in response to the Vibrio vulnificus by RNA-sequencing. *PloS one*, 14:p.e0225976.

Peri S, Roberts S, Kreko IR, McHan LB, Naron A, Ram A, Murphy RL, Lyons E, Gregory BD, Devisetty UK, Nelson AD. 2020. Read mapping and transcript assembly: a scalable and high-throughput workflow for the processing and analysis of ribonucleic acid sequencing data. *Frontiers in genetics*, 10:1361.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes de novo assembler. *Current Protocols in Bioinformatics*, 70:p.e102.

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology*, 11:1-9.

Robinson MD, McCarthy DJ, Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26:139-140. doi: 10.1093/bioinformatics/btp616.

Risso D, Schwartz K, Sherlock G, Dudoit S. 2011. GC-content normalization for RNA-Seq data. *BMC bioinformatics*, 12:1-17.

Smid M, van den Braak RRC, van de Werken HJ, van Riet J, van Galen A, de Weerd V, van der Vlugt-Daane M, Bril SI, Lalmahomed ZS, Kloosterman WP, Wilting SM. 2018. Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. *BMC bioinformatics*, 19:1-13.

Su Z, Mason CE. 2014. SEQC/MAQC-III Consortium A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol*, 32:903-914.

Thellin O, Zorzi W, Lakaye B, De Borman B, Coumans B, Hennen G, Grisar T, Igout A, Heinen E. 1999. Housekeeping genes as internal standards: use and limits. J Biotechnol, 75:291–295. doi: 10.1016/S0168-1656(99)00163-7.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology*, 28:511-515.

Triplett JK, Wang Y, Zhong J, Kellogg EA. 2012. Five nuclear loci resolve the polyploid history of switchgrass (Panicum virgatum L.) and relatives. *PLoS One*, 7:p.e38702.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RTPCR data by geometric averaging of multiple internal control genes. Genome Biol, 3:RESEARCH0034. doi: 10.1186/gb-2002-3-7-research0034.

Wagner GP, Kin K, Lynch VJ. 2012. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in biosciences*, 131:281-285.

Williams CR, Baccarella A, Parrish JZ, Kim CC. 2016. Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC bioinformatics*, 17:1-13.

Yang C, Wu PY, Tong L, Phan J, Wang M. 2015, September. The impact of RNA-seq aligners on gene expression estimation. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics* pp. 462-471.

Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one*, 9:p.e78644.