UMYU Scientifica, Vol. 5 NO. 1, March 2026, Pp 090 – 100
https://doi.org/10.56919/usci.2651.008
ISSN: 2955 – 1145 (print); 2955 – 1153 (online)
A periodical of the Faculty of Natural and Applied Sciences, UMYU, Katsina

ORIGINAL RESEARCH ARTICLE

Comparative Evaluation of Filtering Strategies in Differential Gene Expression Analysis of RNA Sequencing Data

Abdulazeez Giwa[1]*, Barakat Oladipupo[1], Oluwafunmito Ishola[1], Mubaraq Abdulrahmon[1], Zainab Abdulrahman-Giwa[1], Oluwadamilola Ogunmolu[1]
[1]Department of Zoology and Environmental Biology, Lagos State University, Lagos, Nigeria

ABSTRACT
Differential gene expression (DGE) analysis identifies genes expressed at varying levels between conditions, offering valuable insights into affected biological processes. RNA Sequencing (RNA-Seq) DGE analysis usually includes a filtering step to remove genes having low expression from the count data matrix. This study assesses the impact of different filtering strategies on DGE analysis. RNA-Seq read counts of the GSE150706 (n = 72) and TARGET (Therapeutically Applicable Research to Generate Effective Treatments) neuroblastoma (n = 84) datasets were used for analysis. DGE analysis was performed between the Pulled and Close-out groups in GSE150706 and between the *MYCN*-amplified and non-amplified groups in the TARGET neuroblastoma datasets. The effect of filtering strategies (filterByExpr, count, minimal, and no filtering) was assessed on the count data matrix, the number of low-count genes, the number of differentially expressed genes (DEGs) identified, and enrichment analysis. An adjusted p-value < 0.05 was set as the significance threshold for DGE analysis and enrichment analysis. For the GSE150706 dataset, 222, 288, 289, and 208 DEGs were identified from the filterByExpr, none, minimal, and count filtered matrices, respectively, while for the neuroblastoma dataset, 1662, 2059, 2075, and 1579 DEGs were identified from the filterByExpr, none, minimal, and count filtered matrices, respectively. FilterByExpr and count filtering returned no outliers and low counts at the end of DGE analysis. The filtering strategy also influenced enrichment analysis results. Filtering is an important step in DGE analysis with a significant impact on DGE output and downstream analysis. It is recommended to use filterByExpr or count filtering in DGE analysis of RNA-Seq data.

ARTICLE HISTORY
Received December 14, 2025
Accepted March 07, 2026
Published March 15, 2026

KEYWORDS
Transcriptomics, Differential Gene Expression, Filtering, RNA-Seq, Differentially Expressed Genes

INTRODUCTION
Transcriptomics is the study of the transcriptome, the complete set of RNA transcripts produced by the genome, under specific conditions or in a specific cell (Dong & Chen, 2013). Understanding the transcriptome is crucial for elucidating the molecular constituents of cells and tissues and for understanding developmental and disease processes (Wang *et al.*, 2009). Comparative transcriptomic analyses enable the identification of genes differentially expressed across biological conditions, thereby providing insight into regulatory processes and disease-associated pathways.

RNA Sequencing (RNA-Seq) is the current technology of choice used to profile gene expression (Zhao *et al.*, 2014; Rao *et al.*, 2019; van der Kloet *et al.*, 2020). It allows the detection of low-abundance and novel transcripts and enables transcriptome analysis for organisms lacking a reference genome (Levin *et al.*, 2009; Zhao *et al.*, 2014; Rai *et al.*, 2018). RNA-Seq output undergoes processing to infer biological meaning, involving the application of various statistical analyses, bioinformatics methods, and tools. Processing and analysis of RNA-Seq data involve the use of bioinformatics methods and tools implemented in programming languages, commonly Python and R. Python is primarily used to process raw RNA-Seq data, while R is primarily used to analyse already processed RNA-Seq data. The sequential steps of RNA-Seq data analysis include quality control, read alignment to the reference genome, quantification, and differential gene expression (DGE) analysis (Van Verk *et al.*, 2013; Conesa *et al.*, 2016). Variation in these steps involves transcript quantification using the pseudoalignment method, which does not require alignment to a reference genome (Bray *et al.*, 2016; Patro *et al.*, 2017). Following quantification, statistical modeling is applied to identify genes whose

Correspondence: Abdulazeez Giwa. Department of Zoology and Environmental Biology, Lagos State University, Lagos, Nigeria. ✉ abdulazeez.giwa@lasu.edu.ng.
How to cite: Giwa, A., Oladipupo, B., Ishola, O., Abdulrahmon, M., Abdulrahman-Giwa, Z., & Ogunmolu, O. (2026). Comparative Evaluation of Filtering Strategies in Differential Gene Expression Analysis of RNA Sequencing Data. *UMYU Scientifica*, 5(1), 90 – 100. https://doi.org/10.56919/usci.2651.008

expression differs significantly between experimental groups.

Several statistical tools have been developed to identify differentially expressed genes, including DESeq2, edgeR, and limma-voom. These tools use mathematical models designed to specifically analyze RNA-Seq count data and determine whether differences in gene expression between conditions are statistically significant (Anders & Huber, 2010; Love *et al.*, 2014; Law *et al.*, 2014; Robinson *et al.*, 2010). Comparative benchmark studies have demonstrated that DGE results can vary substantially depending on modeling assumptions, dispersion estimation strategies, and preprocessing decisions (Soneson & Delorenzi, 2013; Costa-Silva *et al.*, 2017). Furthermore, replication structure and variance modeling have been shown to significantly influence statistical power and false discovery rates in RNA-Seq experiments (Schurch *et al.*, 2016).

A common yet underexamined preprocessing step in RNA-Seq DGE analysis is filtering of low-expression genes prior to statistical testing. Low-count genes often exhibit high relative variance and may contribute disproportionately to dispersion estimation, thereby influencing model stability and multiple testing correction. Although some tools, such as edgeR, provide built-in filtering functions (e.g., filterByExpr), others rely on user-defined thresholds. The lack of standardized filtering criteria has led to diverse filtering practices across published studies (Chen *et al.*, 2016; Love *et al.*, 2016; Law *et al.*, 2016; Love *et al.*, 2018).

Previous work has suggested that removing low-expression genes may improve detection sensitivity and reduce noise-driven differential expression (Sha *et al.*, 2015). However, broader methodological evaluations have also highlighted concerns regarding reproducibility and false discoveries when statistical control is insufficient (Squair *et al.*, 2021). Despite these observations, the direct impact of alternative filtering strategies on DEG identification consistency and downstream functional interpretation remains insufficiently characterized.

Notably, prior studies such as Giwa *et al.* (2020), Eshibona *et al.* (2022), and Giwa & Giwa (2022) applied edgeR filtering before downstream DGE analysis with DESeq2, illustrating the practical variability in filtering-tool combinations. This raises important methodological questions regarding how pre-filtering decisions interact with downstream dispersion modeling and independent filtering procedures. Since researchers use different filtering methods, which can affect the final results, it is necessary to carefully compare these strategies to understand their impact. The aim of this study is therefore to assess the effect of different filtering strategies on differential gene expression analysis and downstream enrichment outcomes.

MATERIALS AND METHODS

Datasets

Two datasets were used in this study for analysis including the GSE150706 and the TARGET neuroblastoma dataset. These were selected because they are from different groups, i.e., Animal (GSE150706) and human (TARGET neuroblastoma), show contrasting DGE analysis outcomes, and were convenient to access. The GSE150706 (Sun *et al.*, 2020) gene expression count data were downloaded from the Gene Expression Omnibus (GEO). This dataset is a gene expression dataset derived from the blood tissue of 24 beef cattle with Bovine Respiratory Disease (BRD) at three different stages, including entry (arrival of the animal in feedlot), pulled (with identified sickness), and close-out stage (at recovery, healthy animal), for a total of 72 samples. These stages represented subclinical, clinical and healthy states, respectively (Sun *et al.*, 2020). The TARGET neuroblastoma dataset was downloaded from the Xena browser. The TARGET neuroblastoma dataset in the Xena database comprises high-risk neuroblastoma samples with available clinical information. Gene expression RNA-Seq read counts of the TARGET neuroblastoma dataset (dataset ID: TARGET-NBL.htseq_counts.tsv) were obtained from the Genomic Data Commons (GDC) hub in Xena browser using the xenaPython package. The fields used in querying the dataset are outlined in Table 1. Application of the query with the fields outlined in Table 1 returned 84 neuroblastoma samples; 25 of which had *MYCN* amplification, while 59 samples had no *MYCN* amplification.

Table 1: Query criteria for *MYCN* amplification (*MYCN* amp) and *MYCN* non-amplification (*MYCN* non-amp) sample selection

Field	*MYCN* amp values	*MYCN* non-amp values
Diagnostic category	Neuroblastoma	Neuroblastoma
INSS stage	Stage 4	Stage 4
COG risk group	High risk	High risk
MYCN amplification	Amplified	Not Amplified

Filtering

The GSE150706 dataset had expression data for 24616 genes, while the TARGET neuroblastoma dataset had for 60483 genes. The TARGET neuroblastoma dataset was in normalized count format and was converted to raw counts for DGE analysis. For GSE150706, the data matrix included the pulled and close-out samples, with 24 samples each, while the neuroblastoma data matrix included the *MYCN*-amplified and non-amplified samples, with 25 and 59 samples, respectively. Filtering was performed to filter out low expressed genes from the

count data matrices. Four filtering strategies were applied: filterByExpr, none, "minimal," and "count." These filtering strategies were selected because they are common filters potentially used during DGE analysis by researchers. filterByExpr filtering involved the use of the filterByExpr function from edgeR (Robinson *et al.*, 2010), a DGE analysis package. For no filtering strategy, no filtering was performed. For minimal filtering, only rows with at least 10 total reads were kept for downstream analysis. For count filtering, rows having a count total of 10 or more in at least 24 (GSE150706) or 25 (neuroblastoma) samples were retained. These numbers, i.e., 24 and 25, represented the sample sizes of the smallest groups in the two datasets. The impact of these filtering strategies on the count data matrices was identified. The filtered count data matrices were then used for downstream DGE analysis.

Differential gene expression (DGE) analysis

DGE analysis was performed between the pulled (24 samples) and close-out (24 samples) groups of the GSE150706 dataset, and between the *MYCN* amplified (25) and non-amplified (59) groups of the neuroblastoma dataset using DESeq2 (Love *et al.*, 2014) in R. An adjusted p-value < 0.05 and a 1.5 fold change threshold were set in the DGE analysis to identify differentially expressed genes (DEGs).

Comparison

The DEGs output from each filtering strategy were compared for similarities and differences using the *comm* and *diff* commands in the Linux operating system (Ubuntu 22.04.2). The DEGs comparison was carried out between filterByExpr & count, minimal & none, filterByExpr & minimal, filterByExpr & none, count & minimal, and count & no filtering. Thereafter, boxplots of several DEGs were generated from normalized expression values using the Seaborn library in Python. R and Python scripts for the analyses conducted in this study can be found at https://github.com/ZEB-LASU/DGE-Filtering.

Gene enrichment analysis

Gene enrichment analysis was performed to functionally annotate the DEGs. The DEGs from the GSE150706 filterByExpr and the minimal filtering DGE analysis were used for enrichment analysis and compared. The DEGs were in Ensembl ID format and were converted to gene symbols using the BioMart tool in the Ensembl browser. These gene symbols were then input into enrichR (Chen *et al.*, 2013; Kuleshov *et al.*, 2016) (https://maayanlab.cloud/Enrichr/) for enrichment analysis. Enriched terms were terms with p-adjusted value < 0.05.

RESULTS

Filtering and DGE analysis

GSE150706

Based on the gene expression levels in the GSE150706 samples, the filterByExpr, none, minimal, and count filtering removed 12368, 0, 7915, and 12572 genes, respectively, from the data matrix. The DGE analysis identified 222, 288, 289, and 208 DEGs between the pulled and close-out groups from the filterByExpr, none, minimal, and count filtering matrices, respectively. The downstream DGE analysis identified 2975 and 324 low counts in the data matrix for the none and minimal filtering, respectively. No low counts were identified for filterByExpr and count filtering. Table 2 shows information on the kept and removed genes, low counts, and the number of DEGs from the filtering strategies applied to the GSE150706 dataset.

Neuroblastoma

Based on the gene expression levels in the neuroblastoma samples, the filterByExpr, none, minimal, and count filtering removed 33640, 0, 11428, and 34140 genes, respectively, from the data matrix. The DGE analysis identified 1662, 2059, 2075, and 1579 DEGs between the *MYCN*-amplified and non-amplified groups from the filterByExpr, none, minimal, and count filtering matrices, respectively. The downstream DGE analysis identified 16268 and 10468 low counts in the data matrix for the none and minimal filtering, respectively. No low counts were identified for filterByExpr and count filtering. Table 2 shows information on retained and removed genes, low counts, and the number of DEGs from the filtering strategies applied to the neuroblastoma dataset.

Comparison

GSE150706

Comparison of the DEG results from filterByExpr and count filtering identified 208 DEGs common to both strategies and 14 DEGs that differed between them. These 14 DEGs were found to be all in filterByExpr's 'results'. Between minimal and no filtering, there were 288 DEGs common and one DEG different between them, which was found in the minimal filtering 'results'. Between filterByExpr and minimal filtering, 222 DEGs were common, and 67 DEGs were different, which were found in the minimal filtering 'results'. Between filterByExpr and no filtering, 221 DEGs were common, and 68 DEGs were different. 67 of these were found in no filtering 'results', while 1 DEG was found in filterByExpr 'results'. Between count and minimal filtering, 208 DEGs were common, and 81 DEGs were different, and were all found in the minimal filtering 'results'. Between count and no filtering, 208 DEGs were common and 80 DEGs were different, and were all found in the no filtering 'results'. Table 3 shows information about the common and different DEGs between the different filtering strategies results'. Likewise, Figure 1 (A-N) compares the expression of the 14 DEGs between filterByExpr and count filtering. Figure 2 (A-I) compares the expression of some DEGs between filterByExpr and minimal filtering.

Neuroblastoma

Comparison of the DEG results from filterByExpr and count filtering identified 1579 DEGs as common to both

strategies and 83 DEGs as different between them. These 83 DEGs were found to be all in filterByExpr's 'results'. Between minimal and no filtering, there were 2058 DEGs common and 18 DEGs different between them, and 17 of these were found in minimal filtering 'results' while 1 DEG was found in no filtering 'results'. Between filterByExpr and minimal filtering, 1612 DEGs were common, while 513 DEGs were different between them. Between filterByExpr and no filtering, 1599 DEGs were common, and 523 DEGs were different. Between count and minimal filtering, 1538 DEGs were common, and 578 DEGs were different. Between count and no filtering, 1525 DEGs were common, and 588 DEGs were different. Table 3 shows information about the common and different DEGs between the different filtering strategies results'. Likewise, Figure 3 (A-J) compares the expression of some DEGs different between filterByExpr and count filtering.

Table 2: Results from the filtering strategies applied to the GSE150706 and neuroblastoma dataset.

	Filtering	Retained	Removed	Low counts	Up	Down	Total
GSE150706							
	FilterbyExpr	12248	12368	0	162	60	222
	None	24616	0	2975	192	96	288
	Minimal	16701	7915	324	193	96	289
	Count	12644	12572	0	156	52	208
Neuroblastoma	FilterbyExpr	26843	33640	0	544	1118	1662
	None	60483	0	16268	647	1412	2059
	Minimal	49055	11428	10468	649	1426	2075
	Count	26343	34140	0	525	1054	1579

Table 3: Comparison results of the DEGs from the different filtering strategies applied to the GSE150706 and neuroblastoma dataset.

Filtering strategy	dataset	diff	comm	comments
filterByExpr vs count	GSE150706	14	208	All differences found in filterByExpr
	Neuroblastoma	83	1579	
Minimal vs no filtering	GSE150706	1	288	All differences found in minimal
	Neuroblastoma	18	2058	17 in minimal, 1 in no filtering
filterByExpr vs minimal	GSE150706	67	222	All differences found in minimal
	Neuroblastoma	513	1612	463 in minimal, 50 in filterbyExpr
filterByExpr vs no filtering	GSE150706	68	221	67 in no filtering, 1 in filterByExpr
	Neuroblastoma	523	1599	460 in no filtering, 63 in filterbyExpr
Count vs minimal	GSE150706	81	208	All differences found in minimal
	Neuroblastoma	578	1538	41 in count, 537 in minimal
Count vs no filtering	GSE150706	80	208	All differences found in no filtering
	Neuroblastoma	588	1525	534 in no filtering, 54 in count

Table 4: Comparison of the enriched terms between the filterByExpr and minimal enrichment results: KEGG pathways.

Pathways
Epstein-Barr virus infection
Human papillomavirus infection
Pertussis
Necroptosis
Systemic lupus erythematosus*
Cytokine-cytokine receptor interaction
Staphylococcus aureus infection*
Transcriptional misregulation in cancer*
Viral protein interaction with cytokine and cytokine receptor*
Phagosome

***Unique to FilterByExpr**

Gene Enrichment Analysis

Common and different terms between the filterByExpr and minimal enrichment results for GSE150706 are presented in Table 4. The pathway results (KEGG and Reactome) revealed common and different pathways enriched (Table 4 and 5) between the filterByExpr and minimal DEG enrichment results. The significant Gene Ontology: Molecular Factors (GO:MF) terms were the same in both filterByExpr and minimal results. The enriched terms were cytokine receptor activity (GO:0004896), and, double-stranded RNA binding (GO:0003725). For Gene Ontology: Cellular Component (GO:CC), there were four enriched terms for filterByExpr (specific granule (GO:0042581), secretory granule lumen (GO:0034774), collagen-containing, extracellular matrix (GO:0062023) and mitochondrial envelope (GO:0005740), and no significant term for minimal.

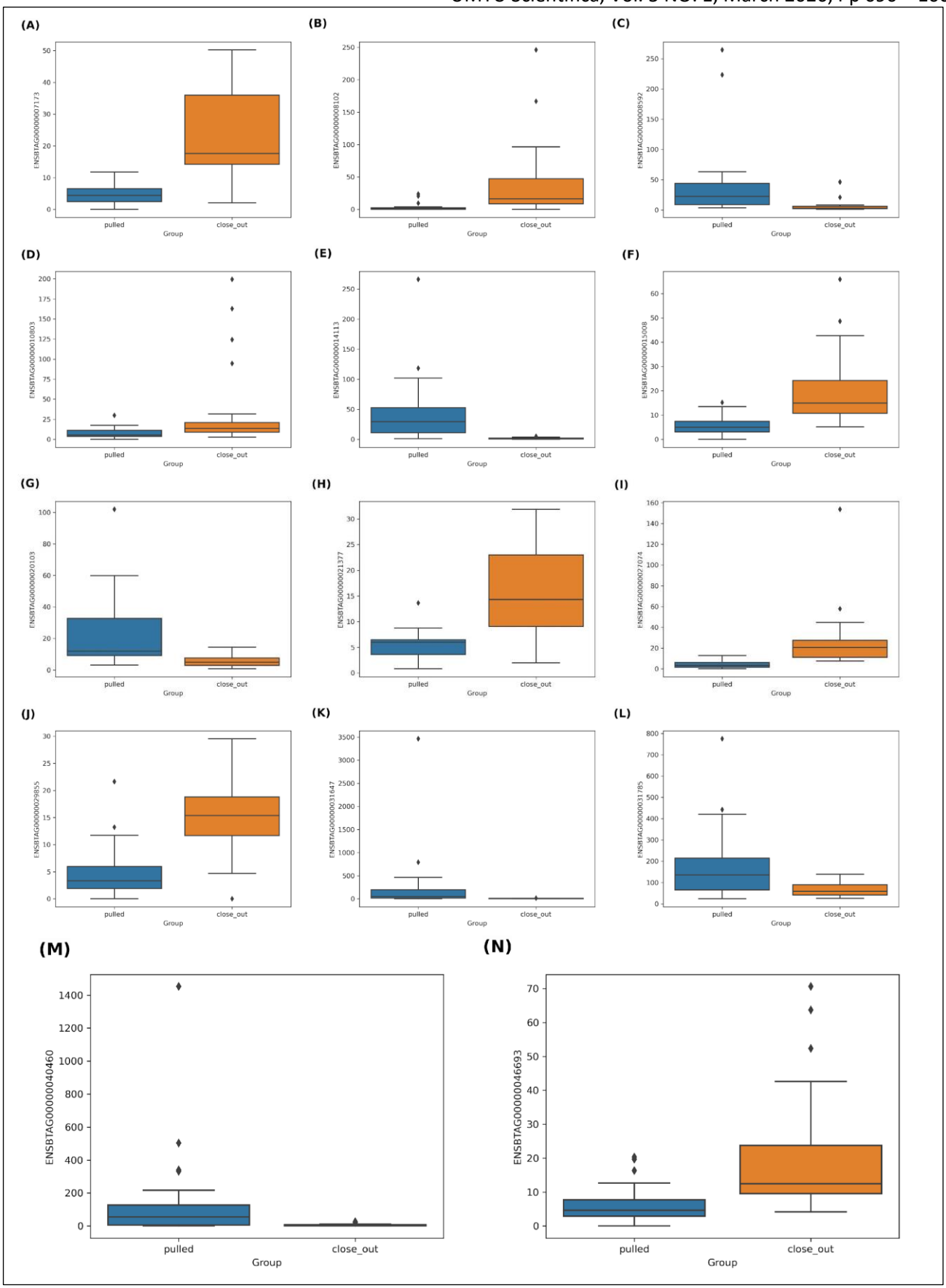

Figure 1: Comparison of the expression of the 14 different DEGs between filterByExpr & count results from the GSE150706 dataset. (A) ENSBTAG00000007173 (B) ENSBTAG00000008102 (C) ENSBTAG00000008592 (D) ENSBTAG00000010803 (E) ENSBTAG00000014113 (F) ENSBTAG00000015008 (G) ENSBTAG00000020103 (H) ENSBTAG00000021377 (I) ENSBTAG00000027074 (J) ENSBTAG00000029855 (K) ENSBTAG00000031647 (L) ENSBTAG00000031785 (M) ENSBTAG00000040460 (N) ENSBTAG000000466993

For Gene Ontology: Biological Process (GO:BP), there were 54 common enriched terms between filterByExpr and minimal, 11 enriched terms unique to minimal, and 35 enriched terms unique to filterByExpr.

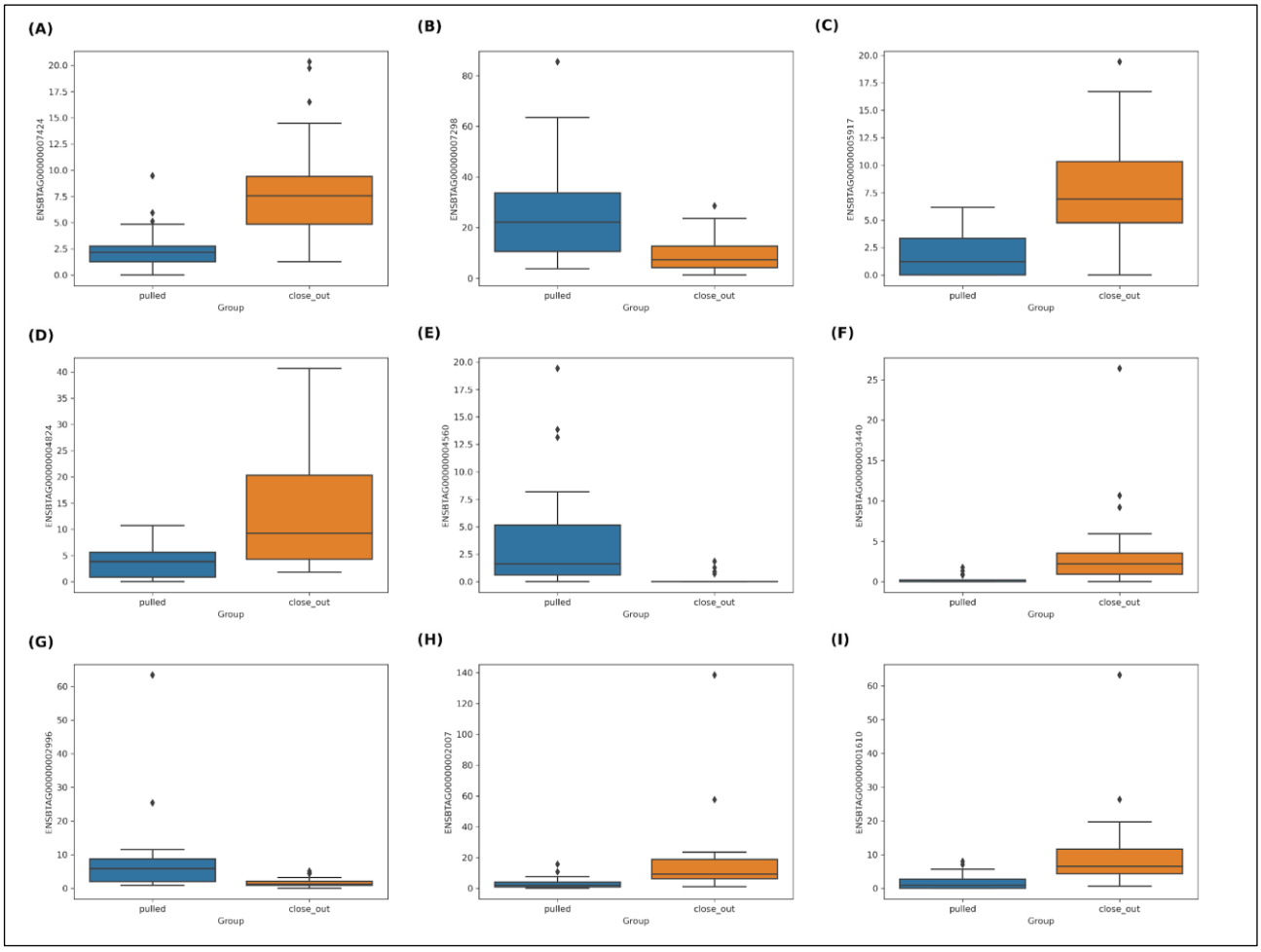

Figure 2: Comparison of the expression of different DEGs between filterByExpr & minimal results from the GSE150706 dataset. (A) ENSBTAG00000007424 (B) ENSBTAG00000007298 (C) ENSBTAG00000005917 (D) ENSBTAG00000004824 (E) ENSBTAG00000004560 (F) ENSBTAG00000003440 (G) ENSBTAG00000002996 (H) ENSBTAG00000002007 (I) ENSBTAG00000001610

DISCUSSION

This study aimed to assess the effect of different filtering strategies on DGE analysis. DGE analysis can provide biological insight into the genetics underlying a condition of interest. The sensitivity of RNA-Seq allows the analysis of transcript levels of all expressed genes, including low abundance transcripts which makes determining which expression changes are biologically significant (Manthey *et al.*, 2014). The presence of noisy, low-expression genes can decrease the sensitivity of detecting DEGs. Thus, identification and filtering of these low-expression genes may improve DEG detection and overall DEG reliability.

A key statistical implication of low-expressed gene is its effect on dispersion modeling in DESeq2. Genes with very low counts tend to exhibit high variance relative to their mean, which can distort shrinkage estimation and affect hypothesis testing. Although DESeq2 performs independent filtering during the results() step, our findings demonstrate that upstream filtering still meaningfully alters DEG output. This suggests that pre-filtering influences dispersion estimation prior to independent filtering, thereby affecting statistical power and false discovery rate (FDR) control. These findings align with observations by Sha *et al.* (2015), who reported improved detection sensitivity following removal of low-expression genes.

Benchmark studies have repeatedly demonstrated that statistical modeling choices substantially influence DEG reproducibility. For instance, Law *et al.* (2014) showed that precision-weighted linear modeling improves variance estimation, while Zhou *et al.* (2014) introduced observation-level weighting to enhance robustness in RNA-Seq data.

The results of the varied filtering strategies/methods highlighted important differences in the DGE analysis outputs and downstream applications (Table 2, 3, 4, 5 & 6). The higher number of DEGs detected under minimal and no filtering strategies may reflect increased sensitivity but also raises concerns regarding false discoveries. Recent large-scale reproducibility analyses have

highlighted the prevalence of unstable differential expression findings when statistical control is insufficient (Squair *et al.*, 2021). The high overlap between filterByExpr and count filtering indicates greater stability and reproducibility of identified DEGs under stricter criteria. In contrast, the large divergence observed when comparing minimal or no filtering suggests increased susceptibility to noise-driven detections.

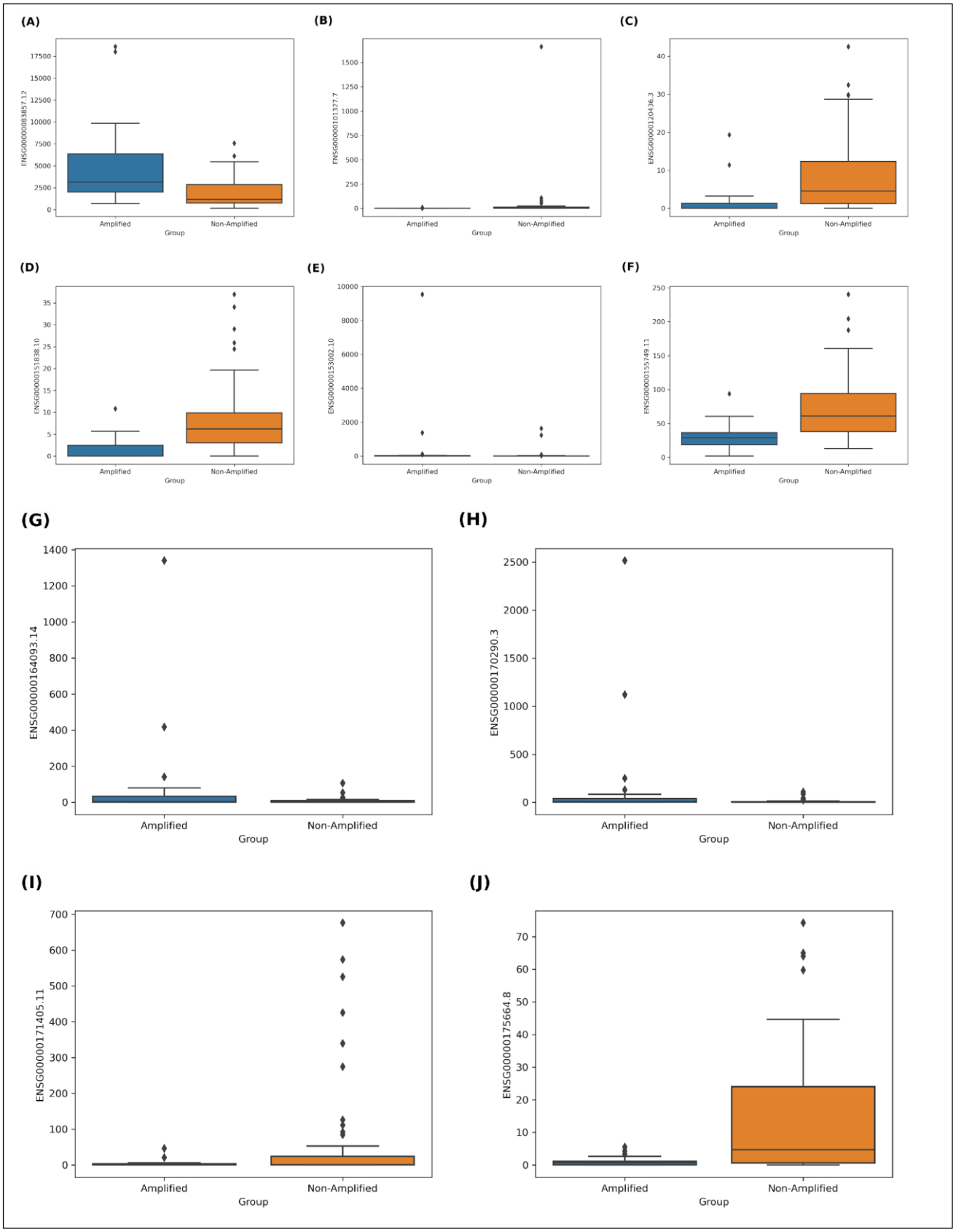

Figure 3: Comparison of the expression of different DEGs between filterByExpr & count results from the TARGET neuroblastoma dataset. (A) ENSG00000083857.12 (B) ENSG00000101327.7 (C) ENSG00000120436.3 (D) ENSG00000151838.10 (E) ENSG00000153002.10 (F) ENSG00000155749.11 (G) ENSG00000164093.14 (H) ENSG00000170290.3 (I) ENSG00000171405.11 (J) ENSG00000175664.8

Table 5: Comparison of the enriched terms between the filterByExpr and minimal enrichment results: Reactome pathways. (common terms *different in filterByExpr ⁺different in minimal)**

Reactome
Interleukin-2 Family Signaling R-HSA-451927**
SARS-CoV-2 Infection R-HSA-9694516
Translesion Synthesis By Y Family DNA Polymerases Bypasses Lesions On DNA Template R-HSA-110313**
SARS-CoV Infections R-HSA-9679506**
STING Mediated Induction Of Host Immune Responses R-HSA-1834941**
Biosynthesis Of DHA-derived SPMs R-HSA-9018677**
Biosynthesis Of Specialized Proresolving Mediators (SPMs) R-HSA-9018678**
IRF3 Mediated Activation Of Type 1 IFN R-HSA-1606341**
Infectious Disease R-HSA-5663205**
Regulation Of Innate Immune Responses To Cytosolic DNA R-HSA-3134975**
Cytosolic Sensors Of Pathogen-Associated DNA R-HSA-1834949**
Synthesis Of Lipoxins (LX) R-HSA-2142700**
Termination Of Translesion DNA Synthesis R-HSA-5656169**
TRAF6 Mediated IRF7 Activation R-HSA-933541**
SARS-CoV-2 Activates/Modulates Innate And Adaptive Immune Responses R-HSA-9705671**
DNA Damage Bypass R-HSA-73893*
Deubiquitination R-HSA-5688426*
ECM Proteoglycans R-HSA-3000178*
Pyroptosis R-HSA-5620971*
Interleukin-18 Signaling R-HSA-9012546*
SARS-CoV-2-host Interactions R-HSA-9705683*
Class I MHC Mediated Antigen Processing And Presentation R-HSA-983169*
Biosynthesis Of E-series 18(S)-resolvins R-HSA-9018896⁺
TRAF3-dependent IRF Activation Pathway R-HSA-918233⁺
Innate Immune System R-HSA-168249⁺
Biosynthesis Of EPA-derived SPMs R-HSA-9018679⁺
Interleukin-4 And Interleukin-13 Signaling R-HSA-6785807⁺
Non-integrin membrane-ECM Interactions R-HSA-3000171⁺
Syndecan Interactions R-HSA-3000170⁺

Biologically, inclusion of low-expression genes may introduce transcripts expressed sporadically or near detection limits (Figure 3 (B, E, G, H)), which may not represent robust biological regulation. This phenomenon explains the differences observed in downstream enrichment analyses.

Importantly, similar patterns were observed in both balanced (GSE150706) and unbalanced (TARGET neuroblastoma) group comparisons. These datasets were selected due to the sample sizes of the contrasting groups, in order to assess if sample sizes would influence filtering output. The GSE150706 had equal group sizes (24 pulled samples versus 24 close-out samples) while the TARGET neuroblastoma dataset had unequal group sizes (25 *MYCN* amplified samples versus 59 *MYCN* non-amplified samples). The pattern of results observed were the same in both datasets demonstrating that sample sizes of the contrasting groups do not influence filtering output and DGE analysis, at least for DESeq2 analysis package.

The boxplots of the normalized expression values gave indication of possible identification of incorrect DEGs based on the filtering method employed. The expression of the DEGs plotted in Figure 3 (B, E, G, H) suggested that they were possibly incorrectly identified as differentially expressed between the studied groups because there was little difference in raw expression between the contrasting groups. Although DGE analysis involves application of specific statistical methods and visualizing the raw expression is not an accurate indication of differential expression of a gene between groups. The enrichment analysis results revealed the downstream consequences of the filtering strategies. It showed that the choice of filtering strategy affects downstream enrichment analysis results.

A number of studies have thus attempted to develop filtering criteria for RNA-Seq data including Manthey *et al.* (2014) which chose genes with a mean RPKM (Reads Per Kilobase of transcript, per Million mapped reads) value greater than two for at least one experimental condition for further analysis. Sha *et al.* (2015) also observed that filtering low-expression genes improved DEG detection sensitivity. The filtering of low-expressed genes is therefore a common practice because it increases confidence in identified DEGs. The authors of DESeq2 in their online user guide suggest that while it is not necessary to pre-filter low count genes before running the DESeq2 functions, removing rows in which there are very few reads reduces the memory size of the data object and increases the speed of the transformation and testing functions within DESeq2. The differences in the DEGs identified were despite performing downstream independent filtering with the "results()" function. In this study, we demonstrate a need for filtering low expressed

genes before independent filtering. However, they also suggested minimal or the stricter general count filtering. Overall, this study highlights that filtering is not merely a preprocessing step for computational convenience but a determinant of DEG reliability, statistical robustness, and biological interpretation. Careful selection of filtering strategy is therefore essential to ensure reproducible and biologically meaningful transcriptomic conclusions.

Table 6: Comparison of the enriched terms between the filterByExpr and minimal enrichment results: GO:BP

GO: BP
response to cytokine (GO:0034097)+
cellular response to cytokine stimulus (GO:0071345)+
positive regulation of NIK/NF-kappaB signaling (GO:1901224)+
defense response to bacterium (GO:0042742)+
positive regulation of inflammatory response (GO:0050729)+
positive regulation of Wnt signaling pathway, planar cell polarity pathway (GO:2000096)+
positive regulation of MAPK cascade (GO:0043410)+
positive regulation of activated T cell proliferation (GO:0042104)+
positive regulation of peptidyl-tyrosine phosphorylation (GO:0050731)+
adenylate cyclase-modulating G protein-coupled receptor signaling pathway (GO:0007188)+
protein autoprocessing (GO:0016540)+
receptor signaling pathway via JAK-STAT (GO:0007259)*
negative regulation of response to biotic stimulus (GO:0002832)*
negative regulation of alpha-beta T cell proliferation (GO:0046642)*
positive regulation of viral entry into host cell (GO:0046598)*
positive regulation of tissue remodeling (GO:0034105)*
positive regulation of T cell apoptotic process (GO:0070234)*
positive regulation by symbiont of entry into host (GO:0075294)*
negative regulation of purine nucleotide metabolic process (GO:1900543)*
negative regulation of glycolytic process (GO:0045820)*
protein poly-ADP-ribosylation (GO:0070212)*
apoptotic mitochondrial changes (GO:0008637)*
positive regulation of NF-kappaB transcription factor activity (GO:0051092)*
positive regulation of immune effector process (GO:0002699)*
negative regulation of cell population proliferation (GO:0008285)*
response to type I interferon (GO:0034340)*
hepoxilin metabolic process (GO:0051121)*
hepoxilin biosynthetic process (GO:0051122)*
negative regulation of systemic arterial blood pressure (GO:0003085)*
positive regulation of mononuclear cell migration (GO:0071677)*
negative regulation of cellular process (GO:0048523)*
long-chain fatty acid biosynthetic process (GO:0042759)*
protein ADP-ribosylation (GO:0006471)*
synaptic vesicle budding (GO:0070142)*
response to interferon-beta (GO:0035456)*
dynamin family protein polymerization involved in membrane fission (GO:0003373)*
dynamin family protein polymerization involved in mitochondrial fission (GO:0003374)*
synaptic vesicle budding from presynaptic endocytic zone membrane (GO:0016185)*
cytoplasmic pattern recognition receptor signaling pathway (GO:0002753)*
negative regulation of CD4-positive, alpha-beta T cell proliferation (GO:2000562)*
cellular response to interleukin-18 (GO:0071351)*
interleukin-18-mediated signaling pathway (GO:0035655)*
RIG-I signaling pathway (GO:0039529)*
leukotriene metabolic process (GO:0006691)*
positive regulation of tumor necrosis factor production (GO:0032760)*
positive regulation of tumor necrosis factor superfamily cytokine production (GO:1903557)*

***Unique to FilterByExpr +Unique to Minimal**

CONCLUSION

In conclusion, we assessed the effect of different filtering strategies on differential gene expression analysis. We find that filterByExpr and count filtering are more conservative, accurate and recommended. Filtering strategy must be taken into consideration and selected with the goal to prevent the identification of false DEGs which will impact downstream analysis and lead to wrong conclusions.

ACKNOWLEDGMENT

We acknowledge the Lagos State University for providing the facilities that facilitated conduction of this study.

REFERENCES

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology, 11*, Article R106. [Crossref]

Bray, N., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near optimal probabilistic RNA-seq quantification. *Nature Biotechnology, 34*, 525–527. [Crossref]

Chen, E., Tan, C., Kou, Y., Duan, Q., Wang, Z., Meirelles, G., Clark, N., & Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics, 14*, Article 128. [Crossref]

Chen, Y., Lun, A., & Smyth, G. (2016). From reads to genes to pathways: Differential gene expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; referees 5; Approved]. *F1000Research, 5*, 1438. [Crossref]

Conesa, A., Madrigal, P., & Tarazona, S. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17*, 1–19. [Crossref]

Costa-Silva, J., Domingues, D., & Lopes, F. M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE, 12*(12), Article e0190152. [Crossref]

Crow, M., Lim, N., Ballouz, S., Pavlidis, P., & Gillis, J. (2019). Predictability of human differential gene expression. *Proceedings of the National Academy of Sciences of the United States of America, 116*, 6491–6500. [Crossref]

Dong, Z., & Chen, Y. (2013). Transcriptomics: advances and approaches. *Science China Life Sciences, 56*, 960–967. [Crossref]

Eshibona, N., Giwa, A., Rossouw, S., Gamieldien, J., Christoffels, A., & Bendou, H. (2022). Upregulation of FHL1, SPNS3, and MPZL2 predicts poor prognosis in pediatric acute myeloid leukemia patients with FLT3-ITD mutation. *Leukemia & Lymphoma, 63*, 1897–1906. [Crossref]

Giwa, A., & Giwa, R. (2022). A 20-Gene expression diagnostic signature of bovine respiratory disease in cattle. *Journal of Scientific Research, 14*, 593–599. [Crossref]

Giwa, A., Fatai, A., Gamieldien, J., Christoffels, A., & Bendou, H. (2020). Identification of novel prognostic markers of survival time in high-risk neuroblastoma using gene expression profiles. *Oncotarget, 11*, 4293–4305. [Crossref]

Hayden, H., Savin, K., Wadeson, J., Gupta, V., & Mele, P. (2018). Comparative metatranscriptomics of wheat rhizosphere microbiomes in disease suppressive and non-suppressive soils for rhizoctonia solani AG8. *Frontiers in Microbiology, 9*, Article 859. [Crossref]

Ismail, R., Baldwin, R., Fang, J., Browning, D., Karlan, B., Gasson, J., & Chang, D. (2000). Differential gene expression between normal and tumor-derived ovarian epithelial cells. *Cancer Research, 60*, 6744–6749. [Link]

Kuleshov, M., Jones, M., & Rouillard, A. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research, 44*, W90–W97. [Crossref]

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology, 15*, Article R29. [Crossref]

Law, C., Alhamdoosh, M., Su, S., Smyth, G., & Ritchie, M. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research, 5*, 1408. [Crossref]

Levin, L., Ekau, W., Gooday, A., Jorissen, F., Middelburg, J., Naqvi, S., Neira, C., Rabalais, N., & Zhang, J. (2009). Effects of natural and human-induced hypoxia on coastal benthos. *Biogeosciences, 6*, 2063–2098. [Crossref]

Love, M., Anders, S., Kim, V., & Huber, W. (2016). RNA-Seq workflow: gene-level exploratory analysis and differential expression. *F1000Research, 4*, 1070. [Crossref]

Love, M., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*, Article 550. [Crossref]

Love, M., Soneson, C., & Patro, R. (2018). Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. *F1000Research, 7*, 952. [Crossref]

Manthey, A., Terrell, A., Lachke, S., Polson, S., & Duncan, M. (2014). Development of novel filtering criteria to analyze RNA-sequencing data obtained from the murine ocular lens during embryogenesis. *Genomics Data, 2*, 369–374. [Crossref]

Nearing, J., Douglas, G., Hayes, M., MacDonald, J., Desai, D., Allward, N., Jones, C., Wright, R., Dhanani, A., Comeau, A., & Langille, M. (2022). Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications, 13*, Article 342. [Crossref]

Niu, S., Yang, J., McDermaid, A., Zhao, J., Kang, Y., & Ma, Q. (2018). Bioinformatics tools for quantitative and functional metagenome and metatranscriptome data analysis in microbes. *Briefings in Bioinformatics, 19*, 1415–1429. [Crossref]

Patro, R., Duggal, G., Love, M., Irizarry, R., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods, 14*, 417–419. [Crossref]

Rai, M., Tycksen, E., Sandell, L., & Brophy, R. (2018). Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears. *Journal of Orthopaedic Research, 36*, 484–497. [Crossref]

Rao, M., Van Vleet, T., Ciurlionis, R., Buck, W., Mittelstadt, S., Blomme, E., & Liguori, M. (2019). Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity

studies. *Frontiers in Genetics, 9*, Article 636. [Crossref]

Robinson, M., McCarthy, D., & Smyth, G. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics, 26*, 139–140. [Crossref]

Saliani, M., Jalal, R., & Javadmanesh, A. (2022). Differential expression analysis of genes and long non-coding RNAs associated with KRAS mutation in colorectal cancer cells. *Scientific Reports, 12*, Article 7965. [Crossref]

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., & Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA, 22*(6), 839–851. [Crossref]

Sha, Y., Phan, J., & Wang, M. (2015). Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. *Conference Proceedings of the Annual International Conference of the IEEE Engineering Medicine and Biology Society, 2015*, 6461–6464. [Crossref]

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics, 14*, Article 91. [Crossref]

Squair, J. W., Gautier, M., Kathe, C., Anderson, M. A., James, N. D., Hutson, T. H., Hudelle, R., Qaiser, T., Matson, K. J. E., Barraud, Q., Levine, A. J., Manno, G. L., Skinnider, M. A., & Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications, 12*, Article 5692. [Crossref]

Stelpflug, S., Sekhon, R., Vaillancourt, B., Hirsch, C., & Buell, C. (2016). An expanded maize gene expression atlas based on RNA sequencing and its use to explore root development. *The Plant Genome, 9*, 1–16. [Crossref]

Sun, H., Srithayakumar, V., Jiminez, J., Jin, W., Hosseini, A., Raszek, M., Orsel, K., Guan, L., & Plastow, G. (2020). Longitudinal blood transcriptomic analysis to identify molecular regulatory patterns of bovine respiratory disease in beef cattle. *Genomics, 112*, 3968–3977. [Crossref]

Tello-Ruiz, M., Stein, J., & Wei, S. (2016). Comparative plant genomics and pathway resources. *Nucleic Acids Research, 44*, D1133–D1140. [Crossref]

van der Kloet, F., Buurmans, J., Jonker, M., Smilde, A., & Westerhuis, J. (2020). Increased comparability between RNA-Seq and microarray data by utilization of gene sets. *PLoS Computational Biology, 16*, Article e1008295. [Crossref]

Van Verk, M., Hickman, R., Pieterse, C., & Van Wees, S. (2013). RNA-Seq: revelation of the messengers. *Trends in Plant Science, 18*, 175–179. [Crossref]

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics, 10*, 57–63. [Crossref]

Westwood, J. (2018). Using transcriptomics to study behavior. In R. T. Gerlai (Ed.), *Molecular-genetic and statistical techniques for behavioral and neural research* (pp. 267–288). Academic Press. [Crossref]

Xue, J., Liu, Y., Wan, L., & Zhu, Y. (2020). Comprehensive analysis of differential gene expression to identify common gene signatures in multiple cancers. *Medical Science Monitor, 26*, Article e919953. [Crossref]

Yang, J., Liu, D., Wang, X., Ji, C., & Cheng, F. (2016). The genome sequence of allopolyploid Brassica juncea and analysis of differential homolog gene expression influencing selection. *Nature Genetics, 48*(10), 1225–1232. [Crossref]

Zhao, S., Fung-Leung, W., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One, 9*, Article e78644. [Crossref]

Zhou, X., Lindsay, H., & Robinson, M. D. (2014). Robustly detecting differential expression in RNA sequencing data using observation weights. *Nucleic Acids Research, 42*(11), Article e91. [Crossref]