Biological Forum – An International Journal 15(11): 499-504(2023)

ISSN No. (Print): 0975-1130
ISSN No. (Online): 2249-3239

DEXA: A Python-based Tool for the Advanced Deciphering of Differential Gene Expression Patterns

***Shbana Begam*[1]*, Samarth Godara*[2*]*, Ramcharan Bhattacharya*[1]*, Rajender Parsad*[2] *and Sudeep Marwaha*[2]**

[1]*ICAR-National Institute for Plant Biotechnology, New Delhi, India.*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.*

(Corresponding author: Samarth Godara)*

(Received: 08 September 2023; Revised: 10 October 2023; Accepted: 20 October 2023; Published: 15 November 2023)
(Published by Research Trend)

ABSTRACT: In today's world of genomics research, where vast RNA-seq datasets are generated, there is a need for bioinformatics tools that provide a user-friendly and potent solution. These tools should enable researchers to swiftly and accurately identify differentially expressed genes, contributing to a deeper understanding of biological mechanisms and pathways of transcription regulators. In this direction, the present study introduces DEXA (Differential Gene Expression Analysis), a user-friendly Python tool for robust RNA-seq data analysis. Utilizing sophisticated statistical methods, DEXA identifies quantitative changes in gene expression by comparing normalized read count data between two conditions. The pipeline is developed from scratch, exhibiting autonomy from any pre-existing bioinformatics package or software. This independence enhances its capability to identify differentially expressed genes at the genetic level. DEXA takes normalized gene counts with replications from two different conditions as input and calculates log2 fold change values based on replicated normalized counts. By identifying genes acting as activators (exclusively expressed in treatment) and deactivators (exclusively expressed in control), DEXA offers valuable insights into the dynamics of gene regulation. DEXA contributes to the advancement of RNA-seq analysis by offering a comprehensive instant solution for researchers in genomics and molecular biology.

Keywords: Activator, deactivators, genomics, differentially expressed gene, normalization, RNA-seq data.

Availability and implementation: This program, accessory utilities and their documentation are available at: https://github.com/ICAR-BIOINFORMATICS/DEXA

INTRODUCTION

The advent of high-throughput sequencing technologies has ushered in an era of unprecedented exploration into the intricate landscape of genomics. Scientists globally now possess potent instruments that facilitate the examination of RNA sequences called RNA-sequencing (RNA-seq), it is one of the most widely used technologies in transcriptomics and has dramatically increased our understanding of the functions and dynamics of complex biological processes (Conesa *et al.*, 2016). Identifying diverged transcripts under different situations, also known as differentially expressed genes (DEGs), is a key step in RNA-seq analysis (Clark *et al.*, 2014). This analysis is crucial in advancing plant research by providing insights into the molecular mechanisms underlying various biological processes. Therefore, differential analysis has been regarded as a valuable approach that involves comparing the expression levels of genes between different experimental conditions, such as different plant tissues, developmental stages, or responses to different treatments. The analysis helps unravel the intricate regulatory networks that govern plant growth, adaptation, and stress tolerance. DEG analysis is particularly valuable in the context of crop improvement, as it aids in identifying genes associated with desirable traits such as increased yield, disease resistance, and environmental resilience. By delving into the differential expression of genes, plant researchers can deepen their understanding of plant biology and pave the way for innovative strategies in crop breeding and sustainable agriculture. DEGs are identified from the count data matrix of FPKM/RPKM/TPM. Currently, more open-source R/Bioconductor packages have been developed for RNA-seq differential expression analysis, particularly DESeq2 (Love *et al.*, 2014) and EdgeR (Robinson *et al.*, 2010). These are popular tools frequently used to get the DEGs with multiple other functionality. Despite multiple functionalities, there are some challenges users face while dealing with these tools such as (Liu *et al.*, 2021):

— Expertise in R Programming Language: they rely on the R programming language for implementation. This dependency may pose a limitation for users who are not proficient in R, requiring a learning curve for effective utilization.

— Dependencies on Multiple Sources and Packages: these tools have dependencies on various external packages, potentially leading to challenges in managing and ensuring compatibility with different versions and updates. This complexity might hinder the seamless integration of these packages into certain computational environments.

— Assumption of Negative Binomial Distribution: DESeq2 assumes a negative binomial distribution for count data. While this assumption is suitable for many RNA-seq datasets, it may not accurately capture the distribution of counts in certain experimental conditions, potentially affecting the accuracy of the analysis.

— Resulted Data Size Exceeding Excel Row Limit: generated output data may become excessively large, exceeding the row limit of standard Excel sheets. Handling and visualizing such extensive datasets can be challenging without appropriate data preprocessing or the use of specialized tools.

To streamline and enhance the extraction of DEGs from count matrices, the present study introduces DEXA, a comprehensive Python pipeline designed for differential expression analysis in RNA-seq datasets, offering a user-friendly and efficient approach. DEXA operates as a command-line tool integrated with Python, offering user-friendly functionality and circumventing the necessity for proficiency in a specific programming language. DEXA utilizes statistical methods to identify quantitative changes in gene expression, comparing normalized read count data between experimental conditions. The pipeline implements advanced RNA-seq analysis techniques, which is built from scratch to have no dependency on any bioinformatics tool for the identification of (DEGs at the genetic level. Moreover, DEXA calculates log2 fold change values based on replicated normalized count data, providing a nuanced understanding of gene expression changes. It performs t-tests between samples for each gene, determining statistical differences in expression. Additionally, DEXA identifies activators and deactivators based on expression patterns.

This tool is primarily designed for scenarios characterized by two distinct situations, each involving any number of replications, wherein a count matrix serves as the requisite input file. Finally, the pipeline outputs a list of genes showing significant expression changes, aiding researchers in identifying key players in biological processes or disease pathways.

METHODOLOGY

The proposed command-line application is built from scratch and integrated with Python. The workflow used to develop the tool is illustrated in Figure 1, and step-by-step details are provided as follows:

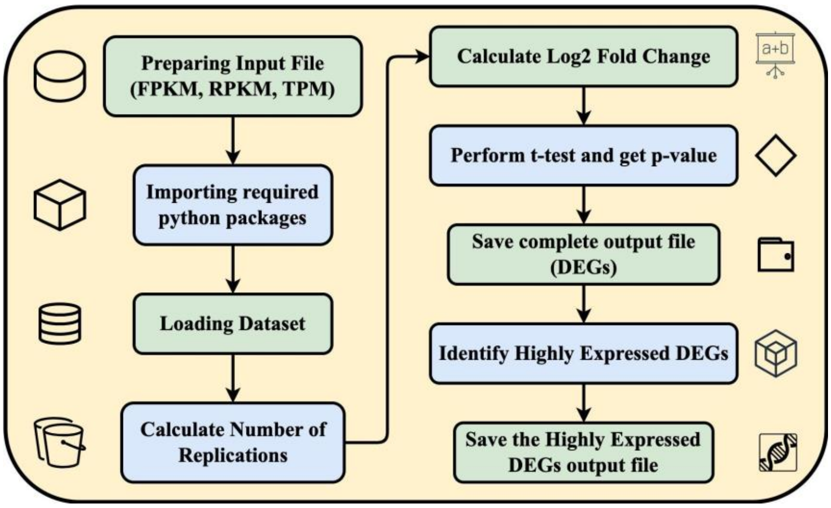
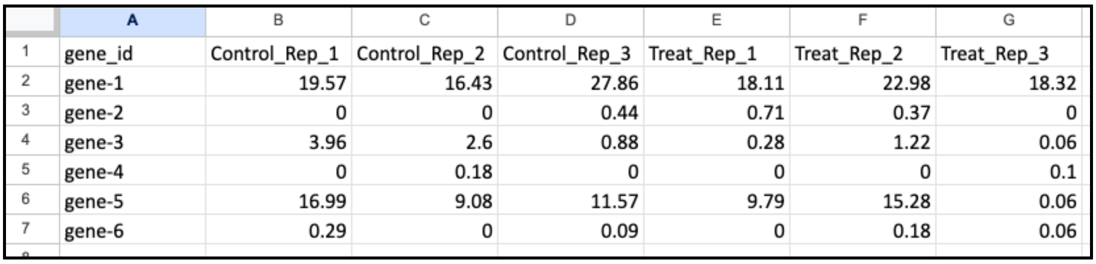

Preparing Input File (FPKM, RPKM, TPM)
Importing required python packages
Loading Dataset
Calculate Number of Replications
Calculate Log2 Fold Change
a+b
Perform t-test and get p-value
Save complete output file (DEGs)
Identify Highly Expressed DEGs
Save the Highly Expressed DEGs output file

Fig. 1. Diagrammatic representation illustrating the methodology of DEXA.

(a) Preparing input file: The proposed tool accepts data formatted as a count matrix (Fig. 2), composed of either Fragments Per Kilobase of transcript per Million mapped reads (FPKM) or Reads Per Kilobase of transcript per Million mapped reads (RPKM), or Transcripts Per Million (TPM), as obtained through abundance estimate methods. It inherently manages variations in the number of replications for individual samples, with the stipulation that each sample must possess an equivalent number of replications.



	A	B	C	D	E	F	G
1	gene_id	Control_Rep_1	Control_Rep_2	Control_Rep_3	Treat_Rep_1	Treat_Rep_2	Treat_Rep_3
2	gene-1	19.57	16.43	27.86	18.11	22.98	18.32
3	gene-2	0	0	0.44	0.71	0.37	0
4	gene-3	3.96	2.6	0.88	0.28	1.22	0.06
5	gene-4	0	0.18	0	0	0	0.1
6	gene-5	16.99	9.08	11.57	9.79	15.28	0.06
7	gene-6	0.29	0	0.09	0	0.18	0.06

Fig. 2. Sample input data file with three replications.

(b) Importing Essential Python Packages: As the pipeline script is authored in Python, its execution necessitates the installation of a few indispensable Python packages, including pandas, numpy, scipy (McKinney & Team 2015; Van Der Walt *et al*., 2011; Virtanen *et al*., 2020).

Loading the Dataset: DEXA's flexibility shines as it effortlessly reads datasets from command-line arguments.

Calculating the Number of Replications: RNA-seq analysis involves comparing gene expression between two conditions, each comprising multiple replications, typically three but sometimes two or four. The pipeline automatically calculates the number of replications from the input data. This feature has been incorporated to enhance user convenience.

Log2Fold Change Calculation: DEXA employs advanced mathematical algorithms to calculate log2 fold change values for each row in the replicated normalized count data, shedding light on the magnitude of gene expression variations. Here, log2fold change values are calculated using the standard equation 1 (Love *et al.,* 2014):

$$L2FC = Mean\left(log_2(condition1)\right) - Mean\left(log_2(condition2)\right)$$

(1)

Where L2FC represents the log2 fold change value obtained by taking the difference between the averages of the log2 values for all replicates in the two conditions.

P-Value Calculation using t-test: Taking the analysis a step further, DEXA conducts t-tests to calculate p-values for each row, providing statistical significance to the identified differentially expressed genes. Equation 2 illustrates the formula for the t-test statistic (Kim, 2015).

$$t = \frac{x-\mu}{s/\sqrt{n}} \qquad \ldots(2)$$

where x is the sample mean, μ is the group mean, s is the sample standard deviation, and n is the sample size. This statistical test is performed to get the probability value (P-value) of a DEG which indicates that there may be a significant difference in gene expression between the conditions being compared.

Saving Results in Output Files: DEXA ensures the traceability of results by diligently saving all the DEGs in an output file. This comprehensive record includes gene values, log2fold change values, and corresponding p-values.

Filter out highly expressed DEGs: Beyond standard analyses, DEXA classifies genes into activators, deactivators, and those with significantly high or low expressions, adding nuanced layers to the understanding of gene regulation. This classification is made up using the log2fold and P-value.

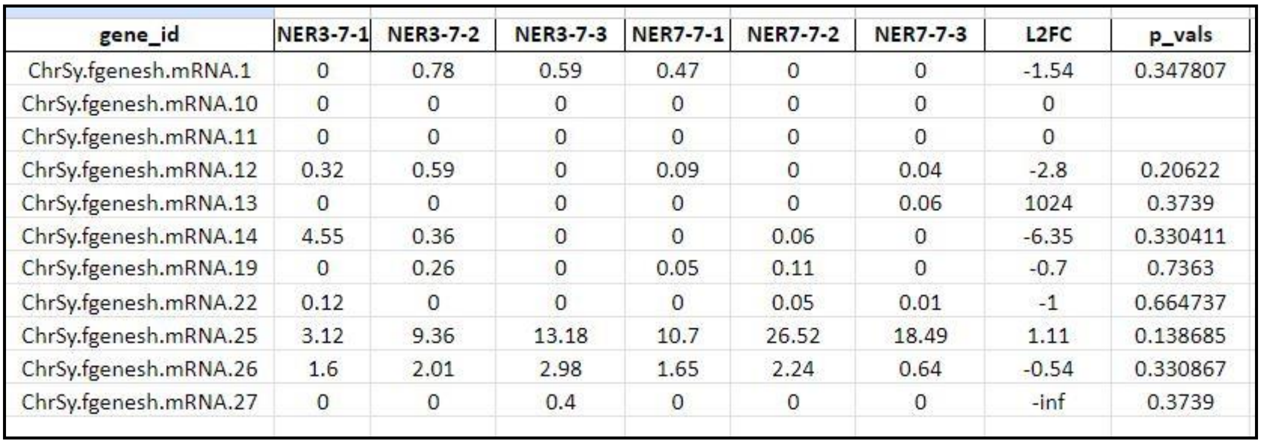
Saving Highly expressed DEGs list: The list of DEGs obtained from the previous step is saved as a quick reference for further in-depth analyses.

RESULTS

Experimental data. The developed DEXA pipeline underwent rigorous testing using 5 samples of different treatments from the rice (Oryza sativa) transcriptome datasets retrieved from Project ID PRJNA681071 (Singh *et al.*, 2022). In this experimental investigation, we are examining the DEGs derived from a treatment involving black and white rice samples. Specifically, we are analyzing five distinct black-versus-white sample pairs in this study.

Output files. DEXA, a novel tool for differential gene expression analysis provides results in two carefully curated files, each pivotal in unravelling the complexities of gene expression dynamics.

All DEGs Output File: This file presents a comprehensive overview of all DEGs values, including detailed log2 fold change values and their corresponding P-values. Further users can filter it as per their selection criteria (Fig. 3). The inclusion of such information establishes a robust foundation for researchers, facilitating an in-depth exploration of the nuanced landscape of gene expression alterations.



gene_id	NER3-7-1	NER3-7-2	NER3-7-3	NER7-7-1	NER7-7-2	NER7-7-3	L2FC	p_vals
ChrSy.fgenesh.mRNA.1	0	0.78	0.59	0.47	0	0	-1.54	0.347807
ChrSy.fgenesh.mRNA.10	0	0	0	0	0	0	0	
ChrSy.fgenesh.mRNA.11	0	0	0	0	0	0	0	
ChrSy.fgenesh.mRNA.12	0.32	0.59	0	0.09	0	0.04	-2.8	0.20622
ChrSy.fgenesh.mRNA.13	0	0	0	0	0	0.06	1024	0.3739
ChrSy.fgenesh.mRNA.14	4.55	0.36	0	0	0.06	0	-6.35	0.330411
ChrSy.fgenesh.mRNA.19	0	0.26	0	0.05	0.11	0	-0.7	0.7363
ChrSy.fgenesh.mRNA.22	0.12	0	0	0	0.05	0.01	-1	0.664737
ChrSy.fgenesh.mRNA.25	3.12	9.36	13.18	10.7	26.52	18.49	1.11	0.138685
ChrSy.fgenesh.mRNA.26	1.6	2.01	2.98	1.65	2.24	0.64	-0.54	0.330867
ChrSy.fgenesh.mRNA.27	0	0	0.4	0	0	0	-inf	0.3739

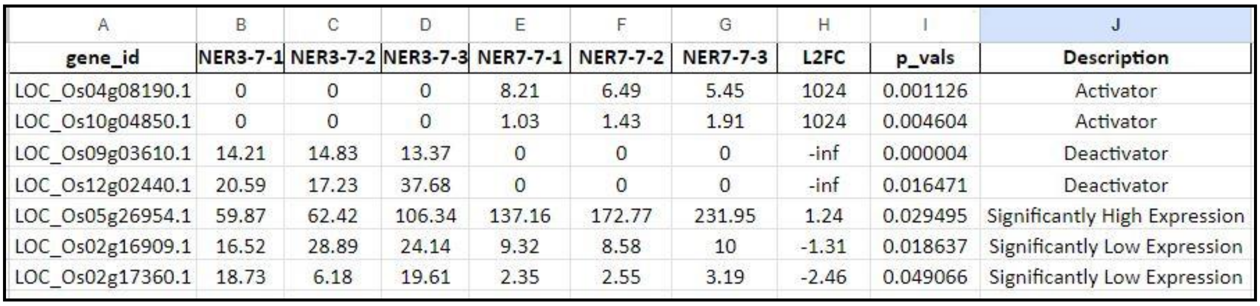
Fig. 3. Sample output data file with all the DEGs using DEXA.

Significantly Expressed DEGs File:

This file contains filtered DEGs that are highly expressed by removing all DEGs having zero and less than zero log2 fold values. DEXA elevates the analysis further by categorizing highly expressed significant genes into four distinct categories (Fig. 4):

— *Highly Significant Positive Regulators*: genes in this category are highly expressed positive DEGs

— *Highly Significant Negative Regulators*: genes in this category are highly expressed negative DEGs
— *Activators*: genes in this category are exclusively expressed positive DEGs in the treated or second condition.
— *Deactivators*: genes in this category are exclusively expressed negative DEGs in the control or first condition.



A	B	C	D	E	F	G	H	I	J
gene_id	NER3-7-1	NER3-7-2	NER3-7-3	NER7-7-1	NER7-7-2	NER7-7-3	L2FC	p_vals	Description
LOC_Os04g08190.1	0	0	0	8.21	6.49	5.45	1024	0.001126	Activator
LOC_Os10g04850.1	0	0	0	1.03	1.43	1.91	1024	0.004604	Activator
LOC_Os09g03610.1	14.21	14.83	13.37	0	0	0	-inf	0.000004	Deactivator
LOC_Os12g02440.1	20.59	17.23	37.68	0	0	0	-inf	0.016471	Deactivator
LOC_Os05g26954.1	59.87	62.42	106.34	137.16	172.77	231.95	1.24	0.029495	Significantly High Expression
LOC_Os02g16909.1	16.52	28.89	24.14	9.32	8.58	10	-1.31	0.018637	Significantly Low Expression
LOC_Os02g17360.1	18.73	6.18	19.61	2.35	2.55	3.19	-2.46	0.049066	Significantly Low Expression

Fig. 4. Sample output data file with filtered highly expressed DEGs using DEXA.

This strategic categorization simplifies data interpretation, enabling researchers to focus on genes with specific roles in different expression scenarios. In essence, DEXA not only provides raw data but also intelligently organizes it, enabling a more targeted and insightful exploration of the genetic landscape. These output files collectively empower researchers to gain a comprehensive understanding of the altered gene expression patterns and the potential regulatory roles of specific genes in diverse biological contexts.

Performance Evaluation: For the evaluation of the developed DEXA tool we checked the log2 fold values calculated by DEXA for each gene expression against the log2 fold values obtained by the DESeq2 tool for this analysis we undertook five distinct black-versus-white rice sample pairs. The evaluation matrices undertaken in the study are Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2) denoted by the equations 3, 4 & 5 respectively (Chai *et al.*, 2014).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad \text{...(3)}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}| y_i - \hat{y}_i| \quad \text{...(4)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \quad \text{... (5)}$$

Where N is the number of observations, y_i represents the actual values, $\hat{y}_i$ represents the predicted values and $\hat{y}$ is the mean of the actual values.

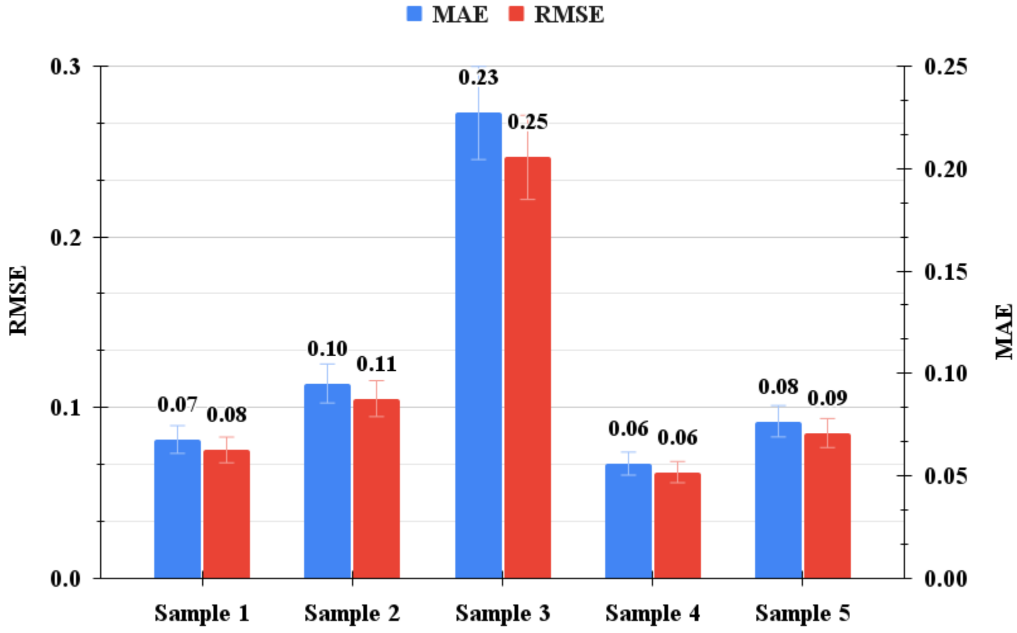


Fig. 4. MAE and RMSE comparisons of the DEXA and DESeq2.

Fig. 4 gives the RMSE and MAE values derived from the output of the DEXA and DESeq2 tools across all analyzed samples. The data reveals that the log2 fold values for RMSE average at 0.11 (ranging from 0.06 to 0.25), while the corresponding MAE values average at 0.10 (ranging from 0.06 to 0.23). These results suggest a close proximity between the outcomes of the implemented tool and those of DESeq2. The observed small deviations in the error metrics underscore the alignment of DEXA with the established DESeq2 methodology. The consistently low RMSE and MAE values indicate the efficacy of the designed tool in producing results closely resembling those generated by DESeq2 across diverse experimental samples.

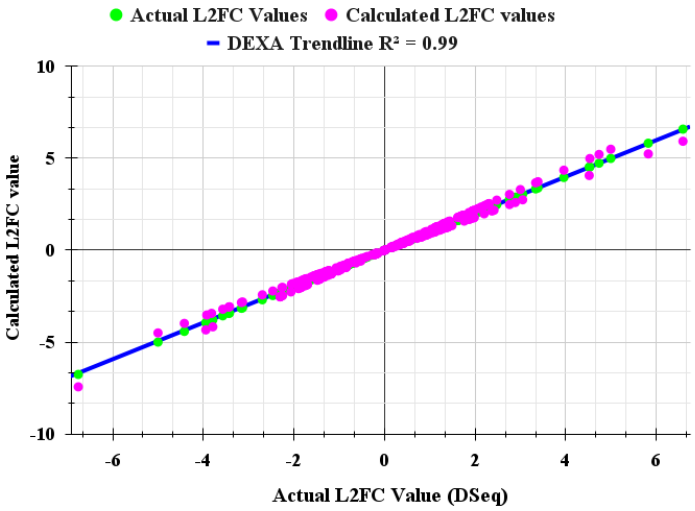


Fig. 5. Coefficient of determination (R^2) of log2 Fold change values of DEXA and DESeq2.

Observing Fig. 5, it is evident that the obtained values exhibit a high degree of goodness of fit, aligning closely with the values produced by the DESeq tool (R^2 = 0.99). The graphical representation highlights the precision of the calculated values, particularly as they approach zero. However, as the values deviate from zero, a discernible increase in the discrepancy between the calculated and actual values is noted. This trend underscores the accuracy of the calculations near zero but indicates a widening disparity as values move away from this central point. The high R^2 value signifies a strong correlation between the calculated and reference values, affirming the reliability of the implemented methodology in capturing the underlying patterns within the dataset.

DISCUSSION

DEXA reveals a robust and user-friendly tool that advances the landscape of differential gene expression analysis. Notably, its integration with Python provides accessibility to a broad user base, streamlining the analysis process for researchers, irrespective of programming language expertise. DEXA's automated handling of replication numbers and efficient execution of t-tests mark a significant departure in gene expression analysis, making it approachable and efficient. The tool's output organization, with the Gene Values File detailing comprehensive gene information and the Significantly Expressed DEGs File categorizing genes strategically, enhances interpretability. The categorization into subgroups like Positive and Negative Regulators, Activators, and Deactivators facilitates nuanced data interpretation, empowering researchers to focus on genes with specific roles in diverse expression scenarios. DEXA's versatility in accommodating different replication scenarios showcases its applicability across diverse experimental setups. Furthermore, its proficiency in handling large-scale datasets positions it as a valuable resource in the era of big data genomics. While it exhibits notable strengths, there is room for continual improvement, suggesting a promising trajectory for further enhancements and broader adoption in genomics research. To get more confidence in the results of DEXA the evaluation matrices RMSE, MAE and R^2 were calculated concerning results of DESeq2. The RMSE and MAE results indicate a close agreement between the DEXA and DESeq2 tools across all samples, with small deviations affirming the efficacy of the designed tool. Additionally, the high R-squared values in Fig. 5 underscore the robust goodness of fit, revealing a strong correlation between the calculated and reference values. These collective findings validate the accuracy and reliability of the DEXA tool in elucidating differential gene expression patterns in diverse experimental scenarios.

CONCLUSIONS

The development and implementation of DEXA, a pipeline for differential gene expression analysis, signify a noteworthy advancement in RNA-seq data analysis. DEXA, a user-friendly Python pipeline, effectively leverages statistical methods and cutting-edge RNA-seq analysis techniques, to identify differentially expressed genes (DEGs) at the genetic level. Robust features within the pipeline, such as log2 fold change calculation, t-test execution, and DEGs classification analysis, contribute to a nuanced understanding of changes in gene expression.

DEXA's ability to generate a curated list of DEGs with significant expression alterations enhances its utility for researchers exploring the complexities of biological processes or investigating specific biological pathways. By addressing limitations present in existing methods (DESeq2 and EdgeR), DEXA emerges as a valuable tool for the scientific community, offering a versatile and comprehensive approach to unravelling the complexities of differential gene expression. The successful implementation of DEXA represents a crucial step forward in the field of RNA-seq analysis, providing researchers with an efficient and reliable means to gain insights into the molecular landscape underlying various experimental conditions.

FUTURE SCOPE

The future development and evolution of DEXA will likely involve a combination of technical refinements, interdisciplinary collaborations, and a keen responsiveness to the evolving needs of researchers in genomics and molecular biology.

Acknowledgement. We would like to thanks ICAR-NIPB, New Delhi and ICAR-IASRI, New Delhi for providing the recourses to conduct the study.
Conflict of Interest. None.

REFERENCES

Chai, Tianfeng, and Roland R. Draxler (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions, 7.1*, 1525-1534.

Clark, N. R., Hu, K. S., Feldmann, A. S., Kou, Y., Chen, E. Y., Duan, Q., & Ma'ayan, A. (2014). The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics*, *15*(1), 1-16.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome biology*, *17*(1), 1-19.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, *68*(6), 540-546.

Liu, S., Wang, Z., Zhu, R., Wang, F., Cheng, Y., & Liu, Y. (2021). Three differential expression analysis methods for RNA sequencing: limma, EdgeR, DESeq2. *JoVE (Journal of Visualized Experiments)*, (175), e62528.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 1-21.

Love, M., Anders, S., & Huber, W. (2014). Differential analysis of count data–the DESeq2 package. *Genome Biol*, *15*(550), 10-1186.

McKinney, W., & Team, P. D. (2015). Pandas-Powerful python data analysis toolkit. *Pandas—Powerful Python Data Analysis Toolkit*, *1625*.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics*, *26*(1), 139-140.

Singh, P. K., Rawal, H. C., Panda, A. K., Roy, J., Mondal, T. K., & Sharma, T. R. (2022). Pan-genomic, transcriptomic, and miRNA analyses to decipher genetic diversity and anthocyanin pathway genes among the traditional rice landraces. *Genomics*, *114*(5), 110436.

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering*, *13*(2), 22-30.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., & Van Mulbregt, P. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

How to cite this article: Shbana Begam, Samarth Godara, Ramcharan Bhattacharya, Rajender Parsad and Sudeep Marwaha (2023). DEXA: A Python-based Tool for the Advanced Deciphering of Differential Gene Expression Patterns. *Biological Forum – An International Journal, 15*(11): 499-504.