2026 Volume: 02 Issue: 08 www.ijrpa.com ISSN 2456-9995 Review Article

International Journal Research Publication Analysis

Page: 01-10

MACHINE LEARNING IN DIFFERENTIAL GENE EXPRESSION ANALYSIS PIPELINES AND BIOINFORMATICS TOOLS FOR CANCER BIOMARKER IDENTIFICATION.

Anushka Choudaha*[1], Dr. Elamathi Natarajan[2]

[1]Student, [2]Professor

Department of Bioinformatics Biotecnika Info Labs, Bengaluru, India.

Article Received: 23 June 2026
Article Revised: 13 July 2026
Published on: 01 August 2026

***Corresponding Author: Anushka Choudaha**
Student, Department of Bioinformatics Biotecnika Info Labs, Bengaluru, India.

DOI: https://doi-doi.org/101555/ijrpa.8989

ABSTRACT

In the modern era, machine learning has become an important computational tool and plays a crucial role in the analysis of RNA sequencing data. Next-Generation Sequencing (NGS) is a high-throughput technology that generates large-scale DNA and RNA datasets, enabling the identification of significant genetic profiles that may not be detectable using conventional methods. Differential Gene Expression (DGE) analysis is a key strategy in RNA-seq data analysis to identify differentially expressed genes across biological samples. Supervised machine learning methods have shown improved performance over conventional approaches for identifying survival-related genes and potential cancer biomarkers from RNA-seq datasets. Several studies have applied Random Forest classification algorithms and the Extreme Pseudo-Sample (EPS) approach, along with variational autoencoders (VAE) and regression models to classify genes and improve predictive accuracy. Various bioinformatics tools and R/Bioconductor packages are available for the statistical analysis of RNA-seq data, helping to identify genes with statistically significant differences between comparable biological samples. **Rosati et al .(2024)**

KEYWORDS: Machine Learning; Differential Gene Expression (DGE); RNA Sequencing (RNA-seq); Next-Generation Sequencing (NGS); Bioinformatics; Cancer Biomarkers.

INTRODUCTION

Differential Gene Expression (DGE) analysis is a key technique in bioinformatics that plays a vital role in understanding the complex molecular mechanisms underlying various

biological processes. By analyzing gene expression levels under different biological conditions, researchers can identify differentially expressed genes (DEGs), providing valuable insights into biological networks, normal physiology, and disease states. This review explores the principles, applications, and limitations of differential gene expression analysis, highlighting its significance in advancing our understanding of complex molecular mechanisms, biological processes, gene regulatory networks, and signaling pathways. Differential gene expression analysis aims to identify genes that exhibit significant changes in expression levels between two or more sample groups, such as different cell types, tumour tissues, disease states, or treatment conditions . **Libbrecht, M. W., & Noble, W. S. (2015)**

DNA contains the genetic information required to determine the structure, properties, and functions of every cell in a living organism. Although all cells contain the same genetic material, different cell types express distinct sets of genes according to their biological functions. Gene expression is the process by which specific genes are selectively activated or repressed to regulate cellular functions. The study of differential gene expression enables comparison of gene expression profiles across different cell types, tissues, and experimental conditions to identify genes that play important roles in determining cellular phenotypes and biological functions. RNA sequencing (RNA-seq), based on next-generation sequencing technology, has largely replaced microarray-based methods for gene expression studies because of its high sensitivity, accuracy, and ability to detect novel transcripts. During RNA-seq experiments, millions of short sequence fragments (reads) are generated from RNA molecules and analyzed to identify differentially expressed genes under different biological conditions. Several bioinformatics software tools have been developed for RNA-seq differential expression analysis. R/Bioconductor packages such as DESeq2, edgeR, and limma provide statistical methods for data preprocessing, normalization, differential expression analysis. **Frances, F. (2014).**

RNA-Seq Differential Expression Analysis: Methods and Bioinformatic Tools

RNA sequencing (RNA-Seq) is a high-throughput technology widely used to analyze gene expression levels under different biological conditions. This technique enables the identification of differentially expressed genes (DEGs), novel transcripts, splice variants, and gene regulatory mechanisms. The standard RNA-Seq workflow includes RNA extraction, library preparation, sequencing, read alignment, quantification, normalization, and differential gene expression analysis. Among these steps, normalization is particularly

important because it minimizes technical variations resulting from differences in sequencing depth and sample composition, thereby improving the accuracy of downstream analyses.

The primary objective of differential gene expression analysis is to identify genes that exhibit statistically significant changes in expression between experimental conditions. Methods used for RNA-Seq differential expression analysis are generally classified into two categories: parametric and non-parametric approaches. Parametric methods assume that count data follow a known statistical distribution, such as the Poisson or Negative Binomial distribution, whereas non-parametric methods do not rely on predefined distributions and are useful when data characteristics are highly variable or unknown.

Several bioinformatics tools have been developed for DEG analysis, each employing distinct statistical models and normalization strategies. DESeq2 utilizes a Negative Binomial model and median-of-ratios normalization, while edgeR applies empirical Bayes methods for dispersion estimation. Limma-voom transforms count data into log-counts and uses linear modeling for differential expression analysis. Other commonly used tools include baySeq , DEGseq , EBSeq, NOISeq, SAMseq, and Sleuth, which provide alternative approaches for identifying differentially expressed genes and transcripts. Following DEG identification, biological interpretation is performed through pathway enrichment analysis and biomarker discovery. Consequently, RNA-Seq has become an indispensable tool in cancer research, disease biomarker identification, and precision medicine**., Rosati, D., et al. (2024**)

COMMON SOFTWARE TOOLS USED FOR RNA – seq Different Gene Expression Aalysis

Software Tool	**Simple Description**	**Main Use**
DESeq2	Uses normalization and statistical modeling to identify differentially expressed genes from RNA-Seq data.	DEG identification
edgeR	Provides accurate differential expression analysis, particularly for datasets with small sample sizes.	Differential expression analysis
limma-voom	Transforms count data and applies linear modeling for gene expression studies.	Gene expression analysis
baySeq	Employs Bayesian statistical methods to estimate differential expression probabilities.	DEG detection
DEGseq	Detects genes with significant expression differences between experimental conditions.	Expression profiling
EBSeq	Uses an empirical Bayesian approach for gene- and transcript-level analysis.	Transcriptomics analysis
NOISeq	A non-parametric method suitable for noisy and highly	DEG identification

Software Tool	Simple Description	Main Use
	variable datasets.	
SAMseq	Uses resampling techniques to identify differentially expressed genes while controlling false discoveries.	Statistical gene analysis
Sleuth	Performs transcript-level differential expression analysis using pseudo-alignment data.	Transcript expression analysis
(Love et al., 2014; Robinson et al., 2010; Law et al., 2014).		

GE02R Web Application and tool :-

- GEO2R is a widely utilized web-based bioinformatics platform designed for differential gene expression analysis of publicly available datasets deposited in the Gene Expression Omnibus (GEO) repository. The tool integrates established Bioconductor packages, including GEOquery for data retrieval and limma for statistical modeling and differential expression analysis. GEO2R enables researchers to compare multiple experimental conditions, identify significantly differentially expressed genes (DEGs), and visualize expression patterns through user-friendly analytical workflows. Its robust statistical framework, incorporating p-value adjustment and data normalization procedures, facilitates the discovery of candidate biomarkers and provides valuable insights into the molecular mechanisms underlying various diseases, particularly cancer and other complex disorders. *GEO2R is an online analytical tool available through the GEO database that allows researchers to examine gene expression profiles and compare different sample groups. By incorporating Bioconductor resources such as GEOquery and limma, it supports data retrieval, normalization, and statistical evaluation of expression datasets. The platform helps identify differentially expressed genes (DEGs), visualize expression trends, and explore potential molecular biomarkers associated with disease development and progression. Owing to its accessibility and ease of use, GEO2R has become a commonly used resource in gene expression and biomarker discovery study.* ***Barrett et al. (2013)***

Enrichment Analysis TOOLS Functional:-

Functional enrichment analysis is widely employed to explore the biological relevance of differentially expressed genes (DEGs) identified through transcriptomic studies. By examining the overrepresentation of specific biological functions and pathways, this approach provides insights into the molecular mechanisms underlying disease initiation and progression (Ashburner et al., 2000). Following DEG identification, enrichment analysis

helps researchers determine whether particular biological processes, cellular components, or molecular functions are significantly associated with the gene set under investigation. Gene Ontology (GO) is one of the most frequently used resources for functional annotation, offering a structured framework for categorizing genes based on their biological roles (Ashburner et al., 2000). In addition, the Kyoto Encyclopedia of Genes and Genomes (KEGG) database facilitates the identification of enriched signaling and metabolic pathways, enabling a deeper understanding of gene interactions and disease-related mechanisms (Kanehisa & Goto, 2000). Collectively, functional enrichment analysis contributes to biomarker discovery and the identification of potential therapeutic targets by linking gene expression changes to biologically meaningful functions and pathways. ***Kanehisa, M., & Goto, S. (2000).***

Gene Ontology (GO)-

analysis is used to classify genes into three major categories: Biological Processes (BP), Molecular Functions (MF), and Cellular Components (CC). It helps researchers understand the functional roles of differentially expressed genes and their involvement in specific biological activities **(Ashburner et al., 2000**).

KEGG Pathway Analysis-

KEGG pathway analysis identifies significantly enriched signaling and metabolic pathways associated with the selected gene set. This approach provides insights into molecular interactions and biological pathways that may contribute to disease development and progression (**Kanehisa & Goto, 2000**).

Reactome Pathway Analysis-

Reactome pathway analysis is a curated pathway database that enables the exploration of biological reactions and molecular pathways. It helps identify key cellular processes, signaling cascades, and functional networks associated with differentially expressed genes (**Jassal et al., 2020**).

Biomarker Validation and Visualization Tools

Heatmap Anaylsis

Heatmap analysis is a commonly used visualization approach for displaying gene expression patterns across multiple samples and experimental conditions. It enables the identification of expression clusters, co-expressed genes, and potential biomarkers by representing expression levels through color gradients (Eisen et al., 1998).

Volcano Plot Analysis

Volcano plots are widely applied to visualize differentially expressed genes by integrating fold-change values with statistical significance. This graphical approach facilitates the rapid identification of significantly upregulated and downregulated genes that may serve as potential diagnostic or prognostic biomarkers (Cui & Churchill, 2003).

Receiver Operating Characteristic (ROC) Curve Analysis

ROC curve analysis is frequently employed to assess the diagnostic performance of candidate biomarkers. The area under the ROC curve (**AUC**) provides a quantitative measure of the sensitivity and specificity of a biomarker, thereby evaluating its ability to discriminate between disease and control groups (Robin et al., 2011).

SURVIVAL ANALYSIS

Survival analysis is used to determine the association between gene expression levels and clinical outcomes. Kaplan–Meier survival curves and Cox proportional hazards models are commonly utilized to evaluate the prognostic significance of candidate biomarkers and their impact on patient survival (Kaplan & Meier, 1958). **Challenges and Limitations of Bioinformatics-Based Biomarker Discovery**

Despite significant advances in bioinformatics technologies, several challenges remain in the identification and validation of reliable biomarkers. One major limitation is the **heterogeneity of biological datasets**, which may arise from differences in sample size, experimental design, sequencing platforms, and data processing methods. Such variations can affect the reproducibility and consistency of research findings (**Ritchie et al., 2015).**

Another challenge is the occurrence of **false-positive and false-negative results** during differential gene expression and enrichment analyses. Although statistical correction methods are commonly applied, biologically irrelevant genes may still be identified as significant, potentially influencing downstream analyses (**Benjamini & Hochberg, 1995**).

The interpretation of high-throughput omics data also remains complex due to the intricate interactions among genes, proteins, and signaling pathways. Consequently, computational predictions alone may not accurately reflect biological functions and therefore require experimental validation through laboratory-based studies (**Hasin et al., 2017).**

Furthermore, many publicly available datasets contain incomplete clinical information, which can limit the assessment of biomarker performance across diverse patient populations. The lack of standardized analytical pipelines and independent validation cohorts further restricts the clinical translation of candidate biomarkers (**Subramanian et al., 2005).**

Finally, while bioinformatics approaches can efficiently identify potential biomarkers and therapeutic targets, their diagnostic and prognostic utility must be confirmed through **in vitro**, **in vivo**, and clinical investigations before implementation in precision medicine and healthcare application. **(Hasin et al., 2017).**

Future Perspectives

The rapid advancement of **bioinformatics, high-throughput sequencing technologies, and multi-omics approaches** is expected to significantly enhance biomarker discovery and precision medicine in the coming years. The integration of genomic, transcriptomic, proteomic, and metabolomic data will provide a more comprehensive understanding of disease mechanisms and facilitate the identification of robust biomarkers with improved diagnostic and prognostic value **Hasin (et al., 2017**).

Emerging applications of **artificial intelligence (AI) and machine learning (ML)** are likely to improve the analysis of large-scale biological datasets by enabling the detection of complex molecular patterns that may not be identifiable through conventional statistical methods. These computational approaches have the potential to accelerate biomarker validation and support personalized treatment strategies **(Hasin et al., 2017).**

In addition, the continuous expansion of public repositories such as **Gene Expression Omnibus (GEO)** and the development of advanced analytical platforms will provide researchers with access to increasingly diverse and high-quality datasets. This will enhance the reproducibility of bioinformatics studies and facilitate cross-validation of candidate biomarkers across different populations and disease conditions (**Barrett et al., 2013).**

Future studies are also expected to benefit from improved **functional enrichment and pathway-based analyses**, including Gene Ontology (GO), KEGG, and Reactome databases. These resources will continue to support the interpretation of biological functions and molecular pathways associated with disease progression, thereby improving the identification of clinically relevant therapeutic targets **(Ashburner et al., 2000; Kanehisa & Goto, 2000; Jassal et al., 2020).**

Furthermore, the integration of bioinformatics findings with experimental and clinical validation studies will be essential for translating candidate biomarkers into routine clinical practice. Such multidisciplinary approaches may contribute to earlier disease detection, more accurate prognosis, and the development of targeted therapeutic interventions in precision et al medicines **(Hasin et al., 2v017).**

CONCLUSION

Bioinformatics has emerged as a powerful and cost-effective approach for the identification of potential biomarkers and therapeutic targets from large-scale biological datasets. The availability of public repositories such as **Gene Expression Omnibus (GEO)**, together with analytical platforms like **GEO2R**, has enabled researchers to efficiently identify differentially expressed genes (DEGs) associated with various diseases Despite considerable progress, challenges related to dataset heterogeneity, statistical bias, and limited clinical validation remain important obstacles in biomarker research. Therefore, computational findings should be complemented by experimental and clinical investigations to ensure their biological relevance and clinical applicability **(Hasin et al., 2017).**

Overall, the integration of bioinformatics tools, functional enrichment strategies, and biomarker validation approaches provides a comprehensive framework for understanding disease mechanisms and advancing precision medicine. Continued developments in multi-omics technologies, artificial intelligence, and large-scale data integration are expected to further improve the accuracy, reproducibility, and clinical utility of biomarker discovery in the future **(Hasin et al., 2017).**

The integration of **functional enrichment analyses**, including **Gene Ontology (GO)**, **KEGG**, and **Reactome** pathway databases, has significantly improved the biological interpretation of gene expression data by revealing key molecular functions, biological processes, and signaling pathways involved in disease progression **(Ashburner et al., 2000; Kanehisa & Goto, 2000; Jassal et al., 2020).** These approaches provide valuable insights into the complex molecular mechanisms underlying human diseases and facilitate the identification of clinically relevant biomarkers.

Furthermore, visualization and validation techniques such as **heatmaps**, **volcano plots**, **ROC curve analysis**, and **survival analysis** play a crucial role in evaluating the diagnostic and prognostic significance of candidate biomarkers. These methods enhance the reliability of bioinformatics findings and support their potential translation into clinical applications **(Robin et al., 2011; Kaplan & Merier,1958)**

REFERENCE

1. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000). Gene ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25–29. https://doi.org/10.1038/75556
2. Barrett, T., Wilhite, S. E., Ledoux, P., et al. (2013). NCBI GEO: Archive for functional genomics data sets—Update. Nucleic Acids Research, 41(D1), D991–D995. https://doi.org/10.1093/nar/gks1193
3. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300. https://www.jstor.org/stable/2346101
4. Cui, X., & Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. Genome Biology, 4(4), 210. https://doi.org/10.1186/gb-2003-4-4-210
5. Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863
6. Francesconi, M., Remondini, D., Neretti, N., et al. (2014). RNA-Seq analysis and applications: An overview. https://doi.org/10.2390/biecoll-jib-2014-235
7. Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. Genome Biology, 18(1), 83. https://doi.org/10.1186/s13059-017-1215-1
8. Jassal, B., Matthews, L., Viteri, G., et al. (2020). The Reactome pathway knowledgebase. Nucleic Acids Research, 48(D1), D498–D503. https://doi.org/10.1093/nar/gkz1031
9. Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research, 28(1), 27–30. https://doi.org/10.1093/nar/28.1.27
10. Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

11. 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(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29
12. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321–332. https://doi.org/10.1038/nrg3920
13. 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), 550. https://doi.org/10.1186/s13059-014-0550-8
14. Robin, X., Turck, N., Hainard, A., et al. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77. https://doi.org/10.1186/1471-2105-12-77
15. 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. https://doi.org/10.1093/bioinformatics/btp616
16. Rosati, D., et al. (2024). Differential Gene Expression Analysis Pipelines and Bioinformatic Tools for the Identification of Specific Biomarkers. https://doi.org/10.3390/biotech13020020
17. Ritchie, M. E., Phipson, B., Wu, D., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research, 43(7), e47. https://doi.org/10.1093/nar/gkv007
18. Subramanian, A., Tamayo, P., Mootha, V. K., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102