Informatics in Biology, Health, and Food

Interdisciplinary Explorations in Life Science

Volume 2. Number 2. Dec. 2025 DOI: 10.22034/ibhf.2025.540076.1033

Original Article

XPA-Express: A Python-Based Tool for Transcriptomic Data Analysis with Application to DNA Repair Deficiency

Amir Mohammad Mazhari

Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran.

Received: 2025/08/08 **Accepted:** 2025/09/30 **Online Published:** 2025/09/30

Abstract

DNA repair mechanisms are essential for maintaining genome integrity and preventing genetic instability, which contributes to various diseases, including cancer, aging, and neurodegenerative disorders. Among these pathways, nucleotide excision repair (NER) plays a crucial role in damage recognition and the recruitment of repair components. Deficiencies in DNA repair pathways, such as those involving the xeroderma pigmentosum group A (XPA) protein, compromise repair capacity, leading to hypersensitivity to ultraviolet (UV) radiation and increased cancer susceptibility. Despite the established role of XPA in DNA repair, its broader impact on transcriptional regulation remains underexplored. To address this, we developed a Python-based tool that automates the retrieval, preprocessing, and differential expression (DE) analysis of RNA-seq data. In this study, we applied the tool to the GEO dataset GSE100855, which consists of 48 samples with different XPA statuses. The analysis identified differentially expressed genes, including STK26, PDE8B, and CYP4F3, which exhibited significant changes in expression. These findings illustrate how DNA repair deficiencies, specifically related to XPA, can alter transcriptional programs and demonstrate the utility of this tool for transcriptomic analysis in the context of DNA repair.

Keywords: RNA-seq, differential expression, DNA repair, transcriptional regulation, XPA, NER.

Introduction

DNA repair mechanisms are crucial for maintaining genome integrity and protecting cells from genetic instability, which can lead to various diseases such as cancer, aging, and neurodegenerative disorders. Among these pathways, nucleotide excision repair (NER) plays a vital role in repairing bulky DNA lesions caused by UV radiation and other genotoxic agents. The xeroderma pigmentosum group A (XPA) protein is a key factor in NER, recognizing damaged DNA and recruiting additional repair factors. A deficiency in XPA impairs DNA repair, resulting in heightened UV sensitivity and an increased risk of cancer. Despite extensive research on DNA repair, the transcriptional consequences of XPA dysfunction remain incompletely understood. While some studies have identified affected genes—particularly those related to mitochondrial function and steroid metabolism—systematic transcriptomic

Cite this article: Mazhari A. M. XPA-Express: A Python-Based Tool for Transcriptomic Data Analysis with Application to DNA Repair Deficiency. *Informatics in Biology, Health, and Food*. 2025;2(2):78-107.

Corresponding authors: Amir Mohammad Mazhari **Email**: am.mazhari@mail.sbu.ac.ir

analyses across different cell types and conditions are limited. To address this gap, we analyzed the GEO dataset GSE100855, which includes 48 RNA-seq samples with different XPA statuses (inactive, active, complemented, and partially active). This dataset was selected for its comprehensive design in comparing transcriptional profiles across these conditions.

For this purpose, we developed a Python-based tool that automates the preprocessing and differential expression (DE) analysis of publicly available RNA-seq data. Rather than focusing on computational details, the tool enables researchers to explore transcriptional changes associated with DNA repair deficiencies.

In this study, we describe the development of the tool, apply it to the GSE100855 dataset, and report the differentially expressed genes associated with XPA status. This approach provides insights into the molecular consequences of DNA repair dysfunction and offers a valuable resource for further investigation of repair-related disorders.

Several web-based platforms allow non-programmers to perform transcriptomic analyses on publicly available data (1, 2). For example, GEOexplorer provides a Shiny-based web interface for querying and retrieving GEO datasets, as well as performing exploratory and differential expression (DE) analyses without requiring coding skills (1). GREIN (GEO RNA-seq Experiments Interactive Navigator) offers a graphical user interface (GUI) backed by a pipeline that processes thousands of GEO RNA-seq experiments (2). GREIN supports data subsetting, quality control (QC), DE analysis, and enrichment using a large database of pre-processed datasets (2).

In contrast, XPA-Express is a Python-based tool for RNA-seq data analysis. It automates the extraction of raw count files and metadata from GEO datasets, performs differential expression (DE) analysis, and generates visual outputs, such as volcano plots. XPA-Express provides a streamlined workflow for RNA-seq analysis, particularly for DNA repair-related studies.

Beyond GEO-focused platforms, several R/Shiny applications offer comprehensive RNA-seq analysis pipelines. For instance, iDEP (integrated Differential Expression and Pathway analysis) integrates dozens of Bioconductor packages (63 in its original release) and extensive annotation databases to perform normalization, clustering, DE testing, and pathway enrichment for over 200 species (3). Tools such as PIVOT and VisRseq wrap many open-source transcriptomics packages into unified graphical interfaces. PIVOT supported over 40 analysis packages and included a graphical data management system for tracking intermediate datasets (4), emphasizing interactivity and workflow provenance. VisRseq similarly offered modular R "apps" that automated tasks such as normalization, PCA, and DE analysis through an accessible GUI (5). Other tools, including START and Shiny-Seq, provided guided analysis pipelines using R/Shiny. These tools typically proceeded through steps like QC, normalization, and enrichment. For example, START was designed to facilitate data upload and visualization of RNA-seq data (6). Shiny-Seq offered a broad set of features, including batch correction, WGCNA, enrichment, and report generation, within a single application (6).

XPA-Express implements a fixed workflow specifically designed for transcriptomic data analysis. It automatically extracts raw counts and metadata from GEO, aligns samples based on specific experimental conditions, and runs integrated DE analysis. This simplified approach makes transcriptomic analysis more accessible to researchers working with DNA repair-related data, without the need for additional computational infrastructure.

Several tools emphasize interactive differential expression analysis. Sleuth is an R/Bioconductor package (with a Shiny interface) that performs transcript- and gene-level DE analysis using Kallisto's bootstrap-based uncertainty estimates (7). DEBrowser is another R/Shiny tool that supports every stage of count data analysis, providing interactive visualizations, batch-effect correction, and iterative filtering through a web interface (8). DEApp specifically targeted researchers without bioinformatics expertise, offering GUI-driven DE analysis with parameter tuning and model cross-validation between edgeR, limma-voom, and DESeq2 (9).

XPA-Express eliminates the need for complex configurations by constructing the count matrix internally, extracting XPA status from GEO metadata, and executing DE analysis automatically. Users do not need to select statistical models or manually process data, making the tool particularly accessible for researchers without extensive computational expertise.

In summary, while existing tools like GEOexplorer (1), GREIN (2), iDEP (3), and others address broad transcriptomic analysis needs, XPA-Express provides a focused, lightweight solution for researchers studying DNA repair processes, particularly in the context of RNA-seq.

Methods

The development of XPA-Express, a Python-based interactive tool for transcriptomic exploration of XPA-related DNA repair deficiency, focused on providing an efficient and accessible framework for RNA-Seq data analysis. The tool was designed to process and analyze publicly available datasets, specifically targeting XPA-related gene expression data.

Software Framework

XPA-Express was built using the Python programming language with the Flask framework, enabling interactive web-based usage. The architecture integrated various Python libraries, including GEOparse for querying the Gene Expression Omnibus (GEO) database, pandas for data manipulation, scipy for statistical testing, and seaborn and matplotlib for data visualization. Flask provided the web interface that organized the RNA-Seq analysis workflow.

Data Retrieval

The tool facilitates the automatic download of metadata from publicly available GEO datasets, while raw RNA-Seq count files are manually provided by the user. In this study, the GSE100855 dataset was utilized, which contains data on human cell lines with disrupted XPA function.

The retrieval process included:

1. Metadata Extraction: GEOparse was used to extract metadata associated with the dataset. This metadata included sample characteristics such as cell line type (XPA-proficient vs. XPA-deficient) and treatment conditions (e.g., retinoic acid).
2. Count Matrix Extraction: The raw count data, stored in compressed .txt.gz files, were extracted from the provided TAR archive using Python's tarfile module. These files were parsed into a structured count matrix, where rows represented genes and columns represented samples.

Data Preprocessing

After data extraction, preprocessing steps were applied:

1. Metadata Parsing: Metadata was parsed to extract relevant sample information, focusing on the XPA status of each sample. This enabled the creation of sample groups for differential expression analysis.
2. Count Matrix Construction: The raw count data from the .counts.txt.gz files were merged into a single count matrix using pandas. The matrix contained gene expression values across multiple samples, with each gene's expression level represented in individual columns.
3. Sample Alignment: To ensure proper alignment, the tool verified that sample names in the count matrix matched those in the metadata file. This ensured accurate linking of XPA status to the corresponding samples.

Differential Expression Analysis

The core of the analysis was the differential expression (DE) test, performed using a t-test to compare gene expression between XPA-deficient and XPA-proficient cell lines.

Steps included:

1. Group Identification: Metadata defined two groups: XPA-proficient and XPA-deficient. Samples were grouped automatically.
2. Statistical Testing: For each gene, expression values in the two groups were compared using an independent two-sample t-test. The log fold change ($\log_2 FC$) was calculated, and a p-value was computed to assess statistical significance.
3. Multiple Testing Correction: Adjusted p-values were calculated using a modified version of the Benjamini-Hochberg procedure, where the p-values were adjusted based on their rank and the number of comparisons.

Data Visualization

XPA-Express generated several visualization outputs to aid interpretation of DE results:

1. Volcano Plot: A volcano plot visualized the relationship between $\log_2$FC and statistical significance. Genes with large fold changes and significant p-values were highlighted.
2. Table of Differentially Expressed Genes (DEGs): A table of the top DEGs displayed gene names, fold changes, and adjusted p-values. The table was sortable and downloadable.
3. Count Matrix Preview: A preview of the count matrix allowed inspection of raw gene expression data for individual samples.

User Interface

The interface was designed to be straightforward. It included:
1. Input Fields: GEO dataset ID and TAR archive path for raw count files.
2. Results Section: Interactive display of results, including volcano plot, DEG table, and count matrix preview.
3. Download Options: Results such as metadata, DEG lists, and the count matrix could be downloaded in standard formats.

Automation and Efficiency

A key feature of XPA-Express is its automation. While users must manually upload the raw RNA-Seq count files (e.g., in a TAR archive), the tool automatically processes the data and performs the analysis. This includes automatic retrieval of metadata and integration of the necessary files required for analysis. The automation allows researchers to focus on biological interpretation rather than computational procedures. This feature is especially useful for researchers without extensive bioinformatics expertise.

Code and Libraries

The implementation used Python and the following libraries:
• Flask: Interactive web-based interface.
• Pandas: Data manipulation and count matrix construction.
• GEOparse: GEO datasets and metadata retrieval.
• Matplotlib and Seaborn: Visualization, including volcano plots.
• SciPy: Statistical testing (e.g., t-tests).
• NumPy: Numerical operations such as log fold change calculation.
• gzip and tarfile: Extraction of compressed raw data files.
• os and pathlib: File and directory operations.
• base64 and io: Encoding volcano plots as images for web display.
• json and warnings: Data handling and warning suppression.

All libraries were distributed under permissive open-source licenses. Flask, GEOparse, NumPy, Pandas, Matplotlib, Seaborn, and SciPy were licensed under the BSD-3-Clause or BSD-compatible licenses. Standard Python modules (gzip, tarfile, os, pathlib, base64, io, json, and warnings) were covered by the PSF/BSD-compatible license.

Application to Dataset

As a proof of concept, XPA-Express was applied to the GSE100855 dataset, which contains transcriptomic data from XPA-proficient and XPA-deficient human cell lines. The analysis identified differential expression in a subset of genes involved in mitochondrial function and steroid hormone metabolism, providing insight into molecular consequences of XPA deficiency.

In conclusion, XPA-Express is a specialized, automated, interactive tool for the analysis of RNA-Seq data associated with XPA deficiency. It supports more accessible and efficient transcriptomic analyses for researchers studying DNA repair. The tool enables investigation of the molecular mechanisms underlying XPA dysfunction and their implications for human health.

Case Study: Application of XPA-Express to Transcriptomic Data

In this case study, we demonstrate the application of XPA-Express, a Python-based interactive tool, for analyzing RNA-seq data specifically related to XPA deficiency. The tool automates the retrieval, preprocessing, and analysis of RNA-seq datasets, focusing on identifying the transcriptional consequences of XPA disruption.

We applied XPA-Express to the publicly available GSE100855 dataset, which includes RNA-seq data from human cell lines with varying XPA statuses. This dataset contains 48

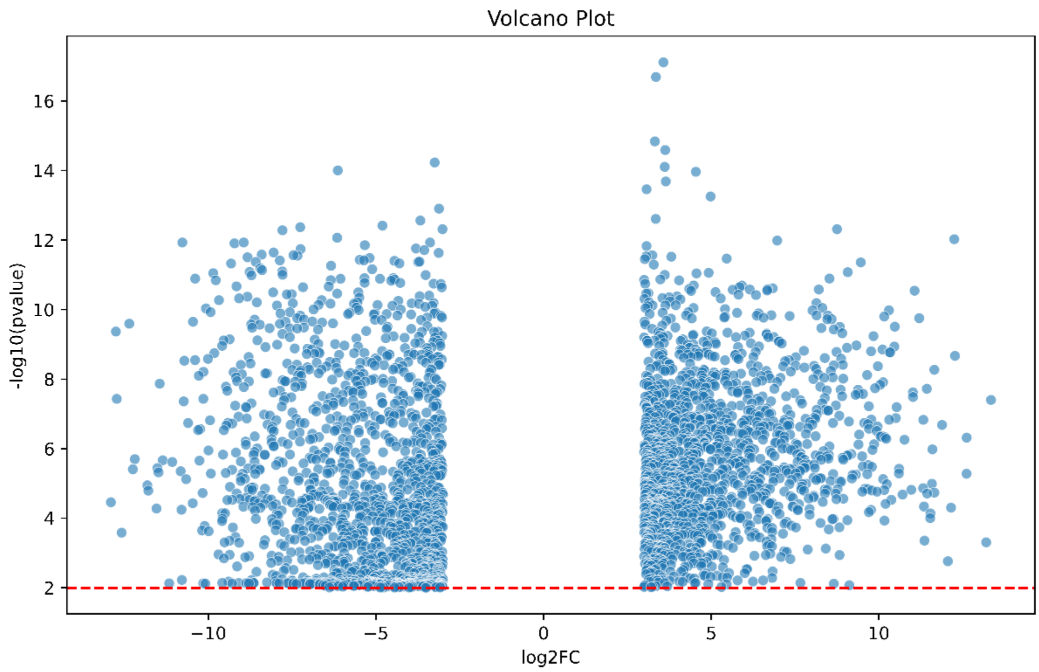

Figure 1. Volcano plot depicting the differential expression results for XPA-deficient vs. XPA-proficient cell lines. The x-axis represents the $\log_2$ fold change ($\log_2$FC), while the y-axis shows the $-\log_{10}$ p-values. The red dashed line indicates the threshold for significance ($p < 0.01$).

samples, representing different cell lines, experimental conditions, and treatment groups.

The primary goal of this analysis was to identify differentially expressed genes (DEGs) that are associated with XPA dysfunction using the automated features of XPA-Express, including data preprocessing, and differential expression analysis. While users must manually upload the raw RNA-Seq count files (e.g., in a TAR archive), the tool automatically handles the rest of the data processing, including metadata retrieval, quality checks, and analysis.

Upon running the analysis, XPA-Express automatically performed differential expression (DE) analysis, identifying genes that were significantly affected by XPA deficiency. These results were visualized through volcano plots and summarized in tables, providing a clear representation of the genes influenced by the XPA status.

For example, the volcano plot in Figure 1 illustrates the distribution of $\log_2$ fold changes ($\log_2$FC) and $-\log_{10}$ p-values for the DEGs. This visualization clearly highlights genes with significant expression changes, particularly those involved in mitochondrial function and steroid hormone metabolism.

The software also generated a comprehensive output that included a list of the top DEGs along with their statistical metrics, such as $\log_2$FC and adjusted p-values. A screenshot of the XPA-Express output interface, displaying the DE results and the volcano plot, is shown in Figure 2. This interface allows researchers to easily interpret the analysis results without requiring advanced programming or computational skills.

The results revealed that XPA deficiency leads to the differential expression of a subset of genes involved in essential cellular functions, particularly those related to mitochondrial activity and steroid hormone metabolism. These findings suggest that XPA may play a role in regulating genes associated with cellular processes and metabolic pathways.

XPA-Express simplifies transcriptomic analysis by providing an accessible and automated platform for exploring the biological consequences of XPA deficiency.

Figure 2. Screenshots of the XPA-Express output interface displaying the volcano plot and the top differentially expressed genes. These outputs offer a visual representation of the analysis results, including a table of significant genes and their statistical metrics.

Data Extraction from GSE100855 Dataset

The first step in the analysis involved extracting the raw count data from the GSE100855_RAW.tar archive. XPA-Express successfully retrieved RNA-seq count data for 48 samples, each containing expression values for 26,364 genes—the total number of unique genes analyzed across all samples. After processing, 3,143 genes were found to be significantly differentially expressed in the dataset.

Metadata Annotation and Sample Grouping

Metadata for each sample was parsed, including critical experimental details such as XPA status. The samples were grouped as follows:

- Inactive (18 samples)
- Active, complemented with XPA gene (12 samples)
- Active (12 samples)
- Partially active (6 aa deletion) (6 samples)

Construction of Gene Expression Count Matrix

After extracting the raw data, the tool merged all sample files into a unified count matrix with dimensions 26,364 × 48, representing gene expression values across all samples. This matrix served as the basis for differential expression analysis.

Differential Expression (DE) Analysis

To assess the impact of XPA deficiency on gene expression, we performed differential expression analysis using a two-sample t-test between each pair of XPA status groups. The analysis was fully automated by XPA-Express, which compared each group to every other group, resulting in multiple pairwise comparisons:

- Inactive vs Active, complemented with XPA gene
- Inactive vs Active
- Inactive vs Partially active
- Active, complemented with XPA gene vs Active
- Active, complemented with XPA gene vs Partially active
- Active vs Partially active

The tool calculated the $\log_2$ fold change ($\log_2 FC$) for each gene in each pairwise comparison and adjusted the p-values using a modified Benjamini-Hochberg procedure, where the p-values were adjusted based on their rank and the number of comparisons. A total of 3,143 significant genes were identified.

Visualization of DE Results: Volcano Plots

Results were visualized using volcano plots to highlight genes with both statistically significant p-values and meaningful expression changes. These visualizations helped identify key genes involved in pathways related to mitochondrial function and steroid hormone metabolism.

Results

Volcano Plot Analysis

The volcano plot generated for the GSE100855 dataset represents the relationship between $\log_2$ fold change ($\log_2$FC) and the $-\log_{10}$(p-value) of each gene in the differential expression analysis. Each point on the plot represents an individual gene, with the x-axis showing the $\log_2$FC and the y-axis representing the $-\log_{10}$(p-value). This format allows the identification of genes with significant expression changes between the groups (XPA-deficient vs. XPA-proficient).

Key Features of the Plot

1. Horizontal Red Line: The red dashed line represents the significance threshold for the p-value, set at 0.01 (or $-\log_{10}(0.01) \approx 2$). Genes above this line have a p-value less than 0.01, indicating that their differential expression is statistically significant.
2. Log_2 Fold Change ($\log_2$FC): The $\log_2$FC on the x-axis quantifies the magnitude of change in gene expression between the groups. Genes with high $\log_2$FC values are positioned far to the left or right, indicating significant upregulation or downregulation. Genes close to the center show minimal differential expression.
3. Gene Distribution: The majority of genes with low expression levels were filtered out and are not visible in the plot. As a result, only genes with significant expression changes, either upregulated or downregulated, are represented in the volcano plot. This explains why there are no points in the center of the plot, as only the differentially expressed genes are shown.
4. Significant Genes: Genes located far from the center, in either the left or right tails of the plot, are considered differentially expressed and have been filtered based on both their $\log_2$FC and p-value thresholds.

Interpretation of Plot

• Upregulated Genes: Genes with a positive $\log_2$FC and p-value less than 0.01 are significantly upregulated in XPA-deficient samples. These genes may be involved in compensatory responses to DNA damage or stress pathways activated by the loss of XPA function.

• Downregulated Genes: Genes with a negative $\log_2$FC and p-value below 0.01 are significantly downregulated. These genes may be involved in processes disrupted due to impaired DNA repair mechanisms linked to XPA deficiency.

• Non-significant Genes: Genes near the center of the plot (low absolute $\log_2$FC and p-values above the significance threshold) have been filtered out due to low expression levels and are not displayed, as they do not show significant differences between the XPA-deficient and proficient groups.

This volcano plot provides a clear method for identifying key differentially expressed genes related to XPA deficiency, supporting further targeted analyses. It highlights both the magnitude and statistical significance of expression changes, offering insights into the molecular pathways affected by XPA dysfunction.

Functional Interpretation of Top Differentially Expressed Genes

To demonstrate a complete analysis workflow using XPA-Express, we conducted a thorough examination of the top differentially expressed genes from the GSE100855 dataset. After identifying these genes through the automated differential expression pipeline, we further enriched the results by extracting detailed functional annotations from the NCBI Gene database. This approach ensures a comprehensive understanding of the biological significance of the identified genes. Table 1 presents these annotations, including gene symbols, full names, identifiers, RefSeq status, tissue-specific expression profiles, and functional summaries, providing valuable context for interpreting the transcriptional consequences of XPA deficiency.

Table 1: Differentially expressed genes identified in the analysis of the GSE100855 dataset. The table provides detailed annotations of genes that exhibit significant expression changes due to XPA dysfunction. Data were retrieved from the NCBI Gene database, offering insights into the genes' biological roles, tissue-specific expression patterns, and potential involvement in various cellular processes. The annotations serve to contextualize the functional implications of these genes within the framework of XPA-related DNA repair deficiencies.

Gene Symbol	Official Full Name	Gene ID	Summary	Expression
RYR2	ryanodine receptor 2	6262	This gene encodes a ryanodine receptor found in cardiac muscle sarcoplasmic reticulum. The encoded protein is one of the components of a calcium channel, composed of a tetramer of the ryanodine receptor proteins and a tetramer of FK506 binding protein 1B proteins, that supplies calcium to cardiac muscle. Mutations in this gene are associated with stress-induced polymorphic ventricular tachycardia and arrhythmogenic right ventricular dysplasia. [provided by RefSeq, Jul 2008]	Biased expression in heart (RPKM 47.0) and brain (RPKM 4.2)
VWDE	von Willebrand factor D and EGF domains	221806	Predicted to enable signaling receptor binding activity. Predicted to be active in cell surface and extracellular region. [provided by Alliance of Genome Resources, Jul 2025]	Biased expression in thyroid (RPKM 1.2), heart (RPKM 0.6) and 13 other tissues
INSL4	insulin like 4	3641	INSL4 encodes the insulin-like 4 protein, a member of the insulin superfamily. INSL4 encodes a precursor that undergoes post-translational cleavage to produce 3 polypeptide chains, A-C, that form tertiary structures composed of either all three chains, or just the A and B chains. Expression of INSL4 products occurs within the early placental cytotrophoblast and syncytiotrophoblast. [provided by RefSeq, Jul 2008]	Restricted expression toward placenta (RPKM 30.0)
STK26	serine/threonine kinase 26	51765	The product of this gene is a member of the GCK group III family of kinases, which are a subset of the Ste20-like kinases. The encoded protein contains an amino-terminal kinase domain, and a carboxy-terminal regulatory domain that mediates homodimerization. The protein kinase localizes to the Golgi apparatus and is specifically activated by binding to the Golgi matrix protein GM130. It is also cleaved by caspase-3	Broad expression in endometrium (RPKM 22.0), bone marrow (RPKM 21.7) and 22 other tissues

			in vitro, and may function in the apoptotic pathway. Several alternatively spliced transcript variants of this gene have been described, but the full-length nature of some of these variants has not been determined. [provided by RefSeq, Jul 2008]	
IFT74	intraflagellar transport 74	80173	This gene encodes a core intraflagellar transport (IFT) protein which belongs to a multi-protein complex involved in the transport of ciliary proteins along axonemal microtubules. IFT proteins are found at the base of the cilium as well as inside the cilium, where they assemble into long arrays between the ciliary base and tip. This protein, together with intraflagellar transport protein 81, binds and transports tubulin within cilia and is required for ciliogenesis. Naturally occurring mutations in this gene are associated with amyotrophic lateral sclerosis--frontotemporal dementia and Bardet-Biedl Syndrome. [provided by RefSeq, Mar 2017]	Broad expression in testis (RPKM 6.8), thyroid (RPKM 3.2) and 24 other tissues
HENMT1	HEN methyltransferase 1	113802	Enables small RNA 2'-O-methyltransferase activity. Involved in RNA methylation. Predicted to be located in P granule. Predicted to be active in cytoplasm and nucleus. [provided by Alliance of Genome Resources, Jul 2025]	Biased expression in testis (RPKM 38.3), lymph node (RPKM 5.6) and 13 other tissues
SEMA6D	semaphorin 6D	80031	Semaphorins are a large family, including both secreted and membrane associated proteins, many of which have been implicated as inhibitors or chemorepellents in axon pathfinding, fasciculation and branching, and target selection. All semaphorins possess a semaphorin (Sema) domain and a PSI domain (found in plexins, semaphorins and integrins) in the N-terminal extracellular portion. Additional sequence motifs C-terminal to the semaphorin domain allow classification into distinct subfamilies. Results demonstrate that transmembrane semaphorins, like the secreted ones, can act as repulsive axon guidance cues. This gene encodes a class 6 vertebrate transmembrane semaphorin that demonstrates alternative splicing. Several transcript variants have been	Broad expression in small intestine (RPKM 13.9), placenta (RPKM 10.4) and 19 other tissues

			identified and expression of the distinct encoded isoforms is thought to be regulated in a tissue- and development-dependent manner. [provided by RefSeq, Nov 2010]	
ANK2	ankyrin 2	287	This gene encodes a member of the ankyrin family of proteins that link the integral membrane proteins to the underlying spectrin-actin cytoskeleton. Ankyrins play key roles in activities such as cell motility, activation, proliferation, contact and the maintenance of specialized membrane domains. Most ankyrins are typically composed of three structural domains: an amino-terminal domain containing multiple ankyrin repeats; a central region with a highly conserved spectrin binding domain; and a carboxy-terminal regulatory domain which is the least conserved and subject to variation. The protein encoded by this gene is required for targeting and stability of Na/Ca exchanger 1 in cardiomyocytes. Mutations in this gene cause long QT syndrome 4 and cardiac arrhythmia syndrome. Multiple transcript variants encoding different isoforms have been described. [provided by RefSeq, Dec 2011]	Biased expression in brain (RPKM 29.6), kidney (RPKM 21.6) and 13 other tissues
CYP4F3	cytochrome P450 family 4 subfamily F member 3	4051	This gene, CYP4F3, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This protein localizes to the endoplasmic reticulum. The enzyme starts the process of inactivating and degrading leukotriene B4, a potent mediator of inflammation. This gene is part of a cluster of cytochrome P450 genes on chromosome 19. Another member of this family, CYP4F8, is approximately 18 kb away. Several transcript variants encoding two different isoforms have been found for this gene. [provided by RefSeq, Apr 2019]	Biased expression in liver (RPKM 51.6), bone marrow (RPKM 29.9) and 6 other tissues

FAM162A	family with sequence similarity 162 member A	26355	Involved in cellular response to hypoxia; positive regulation of apoptotic process; and positive regulation of release of cytochrome c from mitochondria. Located in cytosol and mitochondrion. [provided by Alliance of Genome Resources, Jul 2025]	Ubiquitous expression in colon (RPKM 86.7), esophagus (RPKM 60.9) and 25 other tissues
PLA2G4A	phospholipase A2 group IVA	5321	This gene encodes a member of the cytosolic phospholipase A2 group IV family. The enzyme catalyzes the hydrolysis of membrane phospholipids to release arachidonic acid which is subsequently metabolized into eicosanoids. Eicosanoids, including prostaglandins and leukotrienes, are lipid-based cellular hormones that regulate hemodynamics, inflammatory responses, and other intracellular pathways. The hydrolysis reaction also produces lysophospholipids that are converted into platelet-activating factor. The enzyme is activated by increased intracellular Ca(2+) levels and phosphorylation, resulting in its translocation from the cytosol and nucleus to perinuclear membrane vesicles. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Jul 2015]	Ubiquitous expression in urinary bladder (RPKM 6.0), adrenal (RPKM 5.6) and 24 other tissues
TUBB2A	tubulin beta 2A class IIa	7280	Microtubules, key participants in processes such as mitosis and intracellular transport, are composed of heterodimers of alpha- and beta-tubulins. The protein encoded by this gene is a beta-tubulin. Defects in this gene are associated with complex cortical dysplasia with other brain malformations-5. Two transcript variants encoding distinct isoforms have been found for this gene. [provided by RefSeq, Jul 2015]	Broad expression in brain (RPKM 301.1), bone marrow (RPKM 37.9) and 16 other tissues
PLXNA2	plexin A2	5362	This gene encodes a member of the plexin-A family of semaphorin co-receptors. Semaphorins are a large family of secreted or membrane-bound proteins that mediate repulsive effects on axon pathfinding during nervous system development. A subset of semaphorins are recognized by plexin-A/neuropilin transmembrane receptor complexes, triggering a cellular signal transduction cascade that leads to axon repulsion. This plexin-	Ubiquitous expression in ovary (RPKM 8.4), lung (RPKM 6.9) and 23 other tissues

			A family member is thought to transduce signals from semaphorin-3A and -3C. [provided by RefSeq, Jul 2008]	
KIF3C	kinesin family member 3C	3797	Predicted to enable ATP hydrolysis activity; microtubule binding activity; and microtubule motor activity. Predicted to be involved in microtubule-based movement. Predicted to be located in microtubule cytoskeleton; neuronal cell body; and neuronal ribonucleoprotein granule. Predicted to be part of kinesin complex. Predicted to be active in cytoplasm and microtubule. [provided by Alliance of Genome Resources, Jul 2025]	Biased expression in brain (RPKM 57.3), adrenal (RPKM 4.3) and 1 other tissue
EDIL3	EGF like and discoidin domains 3	10085	The protein encoded by this gene is an integrin ligand. It plays an important role in mediating angiogenesis and may be important in vessel wall remodeling and development. It also influences endothelial cell behavior. [provided by RefSeq, Jul 2008]	Broad expression in brain (RPKM 53.2), gall bladder (RPKM 22.9) and 17 other tissues
ALPK2	alpha kinase 2	115701	Predicted to enable ATP binding activity; protein serine kinase activity; and protein serine/threonine kinase activity. Involved in several processes, including epicardium morphogenesis; heart development; and negative regulation of Wnt signaling pathway involved in heart development. Acts upstream of or within regulation of gene expression. Located in basolateral plasma membrane. [provided by Alliance of Genome Resources, Jul 2025]	Biased expression in heart (RPKM 9.8), lymph node (RPKM 2.0) and 4 other tissues
MEST	mesoderm specific transcript	4232	This gene encodes a member of the alpha/beta hydrolase superfamily. It is imprinted, exhibiting preferential expression from the paternal allele in fetal tissues, and isoform-specific imprinting in lymphocytes. The loss of imprinting of this gene has been linked to certain types of cancer and may be due to promotor switching. The encoded protein may play a role in development. Alternatively spliced transcript variants encoding multiple isoforms have been identified for this gene. Pseudogenes of this gene are	Biased expression in placenta (RPKM 292.2), fat (RPKM 48.4) and 6 other tissues

			located on the short arm of chromosomes 3 and 4, and the long arm of chromosomes 6 and 15. [provided by RefSeq, Dec 2011]	
PDE8B	phosphodiesterase 8B	8622	The protein encoded by this gene is a cyclic nucleotide phosphodiesterase (PDE) that catalyzes the hydrolysis of the second messenger cAMP. The encoded protein, which does not hydrolyze cGMP, is resistant to several PDE inhibitors. Defects in this gene are a cause of autosomal dominant striatal degeneration (ADSD). Several transcript variants encoding different isoforms have been found for this gene.[provided by RefSeq, Jul 2010]	Biased expression in thyroid (RPKM 82.4), brain (RPKM 15.1) and 5 other tissues
ZNF577	zinc finger protein 577	84765	Predicted to enable DNA-binding transcription factor activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity. Predicted to be involved in regulation of transcription by RNA polymerase II. Predicted to be active in nucleus. [provided by Alliance of Genome Resources, Jul 2025]	Ubiquitous expression in skin (RPKM 6.4), ovary (RPKM 5.9) and 25 other tissues
GDA	guanine deaminase	9615	This gene encodes an enzyme responsible for the hydrolytic deamination of guanine. Studies in rat ortholog suggest this gene plays a role in microtubule assembly. Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Nov 2011]	Biased expression in small intestine (RPKM 33.5), duodenum (RPKM 32.2) and 6 other tissues
IFT57	intraflagellar transport 57	55081	Predicted to enable DNA binding activity. Acts upstream of or within apoptotic process and regulation of apoptotic process. Located in ciliary base and photoreceptor connecting cilium. Part of intraciliary transport particle B. Implicated in orofaciodigital syndrome XVIII. [provided by Alliance of Genome Resources, Jul 2025]	Broad expression in testis (RPKM 34.1), lung (RPKM 27.3) and 24 other tissues
TSHZ3	teashirt zinc finger homeobox 3	57616	This gene encodes a zinc-finger transcription factor that regulates smooth muscle cell differentiation in the developing urinary tract. Consistent with this role, mice in which this gene has been inactivated exhibit abnormal gene expression in urinary tract smooth muscle cell precursors and kidney defects including hydronephrosis. The encoded	Broad expression in ovary (RPKM 16.6), endometrium (RPKM 15.6) and 19 other tissues

			transcription factor comprises a gene silencing complex that inhibits caspase expression. Reduced expression of this gene and consequent caspase upregulation may be correlated with progression of Alzheimer's disease in human patients. [provided by RefSeq, Jul 2016]	
PFKFB2	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2	5208	The protein encoded by this gene is involved in both the synthesis and degradation of fructose-2,6-bisphosphate, a regulatory molecule that controls glycolysis in eukaryotes. The encoded protein has a 6-phosphofructo-2-kinase activity that catalyzes the synthesis of fructose-2,6-bisphosphate, and a fructose-2,6-biphosphatase activity that catalyzes the degradation of fructose-2,6-bisphosphate. This protein regulates fructose-2,6-bisphosphate levels in the heart, while a related enzyme encoded by a different gene regulates fructose-2,6-bisphosphate levels in the liver and muscle. This enzyme functions as a homodimer. Two transcript variants encoding two different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]	Broad expression in thyroid (RPKM 25.6), duodenum (RPKM 11.7) and 19 other tissues
RAI14	retinoic acid induced 14	26064	Predicted to enable actin binding activity. Predicted to be involved in cell differentiation and spermatogenesis. Located in cytosol; fibrillar center; and nucleoplasm. [provided by Alliance of Genome Resources, Jul 2025]	Broad expression in endometrium (RPKM 27.5), placenta (RPKM 14.1) and 23 other tissues
ATRN	attractin	8455	This gene encodes both membrane-bound and secreted protein isoforms. A membrane-bound isoform exhibits sequence similarity with the mouse mahogany protein, a receptor involved in controlling obesity. A secreted isoform is involved in the initial immune cell clustering during inflammatory responses that may regulate the chemotactic activity of chemokines. [provided by RefSeq, Apr 2016]	Ubiquitous expression in duodenum (RPKM 20.8), thyroid (RPKM 17.2) and 25 other tissues

ASPHD1	aspartate beta-hydroxylase domain containing 1	253982	Predicted to enable dioxygenase activity. Predicted to be located in membrane. [provided by Alliance of Genome Resources, Jul 2025]	Biased expression in brain (RPKM 13.0), stomach (RPKM 3.5) and 7 other tissues
CMTM3	CKLF like MARVEL transmembrane domain containing 3	123920	This gene belongs to the chemokine-like factor gene superfamily, a novel family that is similar to the chemokine and the transmembrane 4 superfamilies of signaling molecules. This gene is one of several chemokine-like factor genes located in a cluster on chromosome 16. Alternatively spliced transcript variants containing different 5' UTRs, but encoding the same protein, have been identified. [provided by RefSeq, Jul 2008]	Broad expression in testis (RPKM 24.7), placenta (RPKM 16.7) and 24 other tissues
MCRIP1	MAPK regulated corepressor interacting protein 1	348262	Involved in regulation of epithelial to mesenchymal transition. Located in cytoplasmic stress granule and nucleus. [provided by Alliance of Genome Resources, Jul 2025]	Ubiquitous expression in testis (RPKM 31.8), fat (RPKM 25.2) and 25 other tissues
MYBL1	MYB proto-oncogene like 1	4603	Enables DNA-binding transcription activator activity, RNA polymerase II-specific and RNA polymerase II cis-regulatory region sequence-specific DNA binding activity. Involved in positive regulation of transcription by RNA polymerase II. Located in nucleoplasm. [provided by Alliance of Genome Resources, Jul 2025]	Broad expression in lymph node (RPKM 9.0), testis (RPKM 6.1) and 15 other tissues
MNS1	meiosis specific nuclear structural 1	55329	This gene encodes a protein highly similar to the mouse meiosis-specific nuclear structural 1 protein. The mouse protein was shown to be expressed at the pachytene stage during spermatogenesis and may function as a nuclear skeletal protein to regulate nuclear morphology during meiosis. [provided by RefSeq, Oct 2008]	Biased expression in testis (RPKM 28.7), kidney (RPKM 5.6) and 13 other tissues
COL25A1	collagen type XXV alpha 1 chain	84570	This gene encodes a brain-specific membrane associated collagen. A product of proteolytic processing of the encoded protein, CLAC (collagenous Alzheimer amyloid plaque component), binds to amyloid beta-peptides found in Alzheimer amyloid plaques	Biased expression in testis (RPKM 1.9), fat (RPKM 1.8) and 13 other tissues

			but CLAC inhibits rather than facilitates amyloid fibril elongation (PMID: 16300410). A study of over-expression of this collagen in mice, however, found changes in pathology and behavior suggesting that the encoded protein may promote amyloid plaque formation (PMID: 19548013). Multiple transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Dec 2011]	
NOVA1	NOVA alternative splicing regulator 1	4857	This gene encodes a neuron-specific RNA-binding protein, a member of the Nova family of paraneoplastic disease antigens, that is recognized and inhibited by paraneoplastic antibodies. These antibodies are found in the sera of patients with paraneoplastic opsoclonus-ataxia, breast cancer, and small cell lung cancer. Alternatively spliced transcripts encoding distinct isoforms have been described. [provided by RefSeq, Jul 2008]	Biased expression in brain (RPKM 10.5), fat (RPKM 4.3) and 12 other tissues
CYP4F11	cytochrome P450 family 4 subfamily F member 11	57834	This gene, CYP4F11, encodes a member of the cytochrome P450 superfamily of enzymes. The cytochrome P450 proteins are monooxygenases which catalyze many reactions involved in drug metabolism and synthesis of cholesterol, steroids and other lipids. This gene is part of a cluster of cytochrome P450 genes on chromosome 19. Another member of this family, CYP4F2, is approximately 16 kb away. Alternatively spliced transcript variants encoding the same protein have been found for this gene. [provided by RefSeq, Jul 2008]	Biased expression in liver (RPKM 19.6), gall bladder (RPKM 11.3) and 11 other tissues
DCHS1	dachsous cadherin-related 1	8642	This gene is a member of the cadherin superfamily whose members encode calcium-dependent cell-cell adhesion molecules. The encoded protein has a signal peptide, 27 cadherin repeat domains and a unique cytoplasmic region. This particular cadherin family member is expressed in fibroblasts but not in melanocytes or keratinocytes. The cell-cell adhesion of fibroblasts is thought to be necessary for wound healing. [provided by RefSeq, Jul 2008]	Ubiquitous expression in endometrium (RPKM 15.0), placenta (RPKM 6.8) and 22 other tissues

OLFM1	olfactomedin 1	10439	This gene product shares extensive sequence similarity with the rat neuronal olfactomedin-related ER localized protein. While the exact function of the encoded protein is not known, its abundant expression in brain suggests that it may have an essential role in nerve tissue. Several alternatively spliced transcripts encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]	Biased expression in brain (RPKM 118.4), adrenal (RPKM 7.6) and 1 other tissue
BNC1	basonuclin zinc finger protein 1	646	This gene encodes a zinc finger protein present in the basal cell layer of the epidermis and in hair follicles. It is also found in abundance in the germ cells of testis and ovary. This protein is thought to play a regulatory role in keratinocyte proliferation and it may also be a regulator for rRNA transcription. Disruption of this gene has been implicated in premature ovarian failure as well as testicular premature aging. [provided by RefSeq, Sep 2020]	Biased expression in testis (RPKM 16.5), esophagus (RPKM 5.8) and 2 other tissues
C14orf132	chromosome 14 open reading frame 132	56967	Predicted to be located in membrane. [provided by Alliance of Genome Resources, Jul 2025]	Broad expression in brain (RPKM 28.5), endometrium (RPKM 10.8) and 14 other tissues
MYL9	myosin light chain 9	10398	Myosin, a structural component of muscle, consists of two heavy chains and four light chains. The protein encoded by this gene is a myosin light chain that may regulate muscle contraction by modulating the ATPase activity of myosin heads. The encoded protein binds calcium and is activated by myosin light chain kinase. Two transcript variants encoding different isoforms have been found for this gene. [provided by RefSeq, Jul 2008]	Broad expression in prostate (RPKM 922.2), urinary bladder (RPKM 715.8) and 17 other tissues
GGACT	gamma-glutamylamine cyclotransferase	87769	The protein encoded by this gene aids in the proteolytic degradation of crosslinked fibrin by breaking down isodipeptide L-gamma-glutamyl-L-epsilon-lysine, a byproduct of fibrin degradation. The reaction catalyzed by the encoded gamma-	Biased expression in kidney (RPKM 12.0), fat (RPKM 1.8) and 13 other tissues

			glutamylaminecyclotransferase produces 5-oxo-L-proline and a free alkylamine. Two transcript variants encoding the same protein have been found for this gene.[provided by RefSeq, Aug 2010]	
FBLN1	fibulin 1	2192	Fibulin 1 is a secreted glycoprotein that becomes incorporated into a fibrillar extracellular matrix. Calcium-binding is apparently required to mediate its binding to laminin and nidogen. It mediates platelet adhesion via binding fibrinogen. Four splice variants which differ in the 3' end have been identified. Each variant encodes a different isoform, but no functional distinctions have been identified among the four variants. [provided by RefSeq, Jul 2008]	Broad expression in gall bladder (RPKM 239.4), placenta (RPKM 204.3) and 16 other tissues
DOCK2	dedicator of cytokinesis 2	1794	The protein encoded by this gene belongs to the CDM protein family. It is specifically expressed in hematopoietic cells and is predominantly expressed in peripheral blood leukocytes. The protein is involved in remodeling of the actin cytoskeleton required for lymphocyte migration in response to chemokine signaling. It activates members of the Rho family of GTPases, for example RAC1 and RAC2, by acting as a guanine nucleotide exchange factor (GEF) to exchange bound GDP for free GTP. Mutations in this gene result in immunodeficiency 40 (IMD40), a combined form of immunodeficiency that affects T cell number and function, also with variable defects in B cell and NK cell function. [provided by RefSeq, May 2018]	Broad expression in lymph node (RPKM 13.3), bone marrow (RPKM 11.2) and 15 other tissues
CAMK4	calcium/calmodulin dependent protein kinase IV	814	The product of this gene belongs to the serine/threonine protein kinase family, and to the Ca(2+)/calmodulin-dependent protein kinase subfamily. This enzyme is a multifunctional serine/threonine protein kinase with limited tissue distribution, that has been implicated in transcriptional regulation in lymphocytes, neurons and male germ cells. [provided by RefSeq, Jul 2008]	Broad expression in brain (RPKM 7.0), lymph node (RPKM 3.7) and 21 other tissues

SYNE1	spectrin repeat containing nuclear envelope protein 1	23345	This gene encodes a spectrin repeat containing protein expressed in skeletal and smooth muscle, and peripheral blood lymphocytes, that localizes to the nuclear membrane. Mutations in this gene have been associated with autosomal recessive spinocerebellar ataxia 8, also referred to as autosomal recessive cerebellar ataxia type 1 or recessive ataxia of Beauce. Alternatively spliced transcript variants encoding different isoforms have been described. [provided by RefSeq, Jul 2008]	Ubiquitous expression in ovary (RPKM 11.9), bone marrow (RPKM 8.2) and 24 other tissues
PELI2	pellino E3 ubiquitin protein ligase family member 2	57161	Predicted to enable protein-macromolecule adaptor activity and ubiquitin protein ligase activity. Acts upstream of or within positive regulation of MAPK cascade and positive regulation of protein phosphorylation. Predicted to be located in cytosol. [provided by Alliance of Genome Resources, Jul 2025]	Low expression observed in reference dataset
FGF1	fibroblast growth factor 1	2246	The protein encoded by this gene is a member of the fibroblast growth factor (FGF) family. FGF family members possess broad mitogenic and cell survival activities, and are involved in a variety of biological processes, including embryonic development, cell growth, morphogenesis, tissue repair, tumor growth and invasion. This protein functions as a modifier of endothelial cell migration and proliferation, as well as an angiogenic factor. It acts as a mitogen for a variety of mesoderm- and neuroectoderm-derived cells in vitro, thus is thought to be involved in organogenesis. Multiple alternatively spliced variants encoding different isoforms have been described. [provided by RefSeq, Jan 2009]	Biased expression in brain (RPKM 36.0), kidney (RPKM 21.0) and 4 other tissues
PLEKHO1	pleckstrin homology domain containing O1	51177	Predicted to be involved in regulation of myoblast fusion. Predicted to act upstream of or within several processes, including lamellipodium morphogenesis; myoblast fusion; and myoblast migration. Predicted to be located in cytoplasm; nucleus; and plasma membrane. Predicted to be active in muscle cell projection membrane and ruffle membrane. [provided by Alliance of Genome Resources, Jul 2025]	Broad expression in testis (RPKM 30.7), appendix (RPKM 29.1) and 24 other tissues

TDRD9	tudor domain containing 9	122402	Predicted to enable ATP hydrolysis activity; RNA binding activity; and helicase activity. Involved in spermatogenesis. Located in cytoplasm and nucleus. Implicated in spermatogenic failure 30. [provided by Alliance of Genome Resources, Jul 2025]	Biased expression in testis (RPKM 26.4) and thyroid (RPKM 5.7)

Discussion

This study demonstrates the utility of XPA-Express, a Python-based tool designed for the automated transcriptomic analysis of RNA-seq data. XPA-Express offers an efficient platform for researchers to explore transcriptional changes associated with various biological conditions and dysfunctions through automated preprocessing and differential expression analysis.

The application of XPA-Express to the GSE100855 dataset showcases the tool's capability to identify differentially expressed genes and present the results through intuitive visualizations, such as volcano plots. These features empower researchers to interpret complex RNA-seq data effectively, without requiring specialized computational expertise. The tool's fully automated pipeline streamlines the RNA-seq analysis process, making it accessible for researchers across diverse fields.

Conclusion

Looking ahead, XPA-Express holds promise for use with a wide variety of datasets, offering valuable insights into gene expression across different biological contexts. Its user-friendly interface and efficient workflow make it a valuable resource for transcriptomic research, enabling the detailed exploration of gene expression patterns and contributing to further understanding of the molecular mechanisms underlying diseases and biological processes.

References

1. Hunt GP, Grassi L, Henkin R, Smeraldi F, Spargo TP, Kabiljo R, et al. GEOexplorer: a webserver for gene expression analysis and visualisation. Nucleic Acids Research. 2022;50(W1):W367-W74. doi:10.1093/nar/gkac364

2. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. Scientific Reports. 2019;9(1):7580. doi:10.1038/s41598-019-43935-8

3. Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics. 2018;19(1):534. doi:10.1186/s12859-018-2486-6

4. Zhu Q, Fisher SA, Dueck H, Middleton S, Khaladkar M, Kim J. PIVOT: platform for interactive analysis and visualization of transcriptomics data. BMC Bioinformatics. 2018;19(1):6. doi:10.1186/s12859-017-1994-0

5. Younesy H, Möller T, Lorincz MC, Karimi MM, Jones SJM. VisRseq: R-based visual framework for analysis of sequencing data. BMC Bioinformatics. 2015;16(11):S2. doi:10.1186/1471-2105-16-S11-S2.

6. Sundararajan Z, Knoll R, Hombach P, Becker M, Schultze JL, Ulas T. Shiny-Seq: advanced guided transcriptome analysis. BMC Research Notes. 2019;12(1):432. doi:10.1186/s13104-019-4471-1

7. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nature Methods. 2017;14(7):687-90. doi:10.1038/nmeth.4324

8. Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. DEBrowser: interactive differential expression analysis and visualization tool for count data. BMC Genomics. 2019;20(1):6. doi:10.1186/s12864-018-5362-x.

9. Li Y, Andrade J. DEApp: an interactive web interface for differential expression analysis of next generation sequence data. Source Code for Biology and Medicine. 2017;12(1):2. doi:10.1186/s13029-017-0063-4