scientific reports

Check for updates

OPEN

Confidence: a web app for cross-platform differential gene expression analysis, gene scoring, and enrichment analysis

Abhishek Shastry[1,7], Benjamin P. Ott[2,7], Tanvi Nandani[2,4,5], Alex Paterson[3], Matt Simpson[4], Kimberly J. Dunham-Snary[1,4,5,8]✉ & Charles C. T. Hindmarch[1,2,4,5,6,8]✉

RNA-seq quantifies the abundance of transcripts within a biological sample and performs differential analysis between different conditions to reveal regulated gene signatures. Three challenges exist: (1) different analytical packages can often report different expression patterns and false-discovery-rates and *P*-values; (2) the effective use of these analytical packages requires substantial knowledge of programming and bioinformatics; and (3) there are a lack of intuitive methods to prioritize target genes for further investigation. To address these challenges, we developed Confidence, a web-based application to perform simultaneous statistical analysis of RNA-seq count data. Confidence incorporates the Confidence Score (CS), ranging from 1 to 4 to aid in gene prioritization, where 1 represents low confidence and 4 represents high confidence. The Confidence web-based application was designed for rapid and intuitive analysis of standard experimental metadata and gene count inputs providing a web-based, 'wide-net' approach to RNA-seq analysis. Gene scoring allows for unbiased gene selection and identification of novel genes strongly associated with disease/treatment models across multiple species. Pathway analysis has been integrated so that highly confident genes can be placed into biological context. Confidence provides a new strategy for target prioritization in RNA-seq analysis and the generation of publication-quality figures, which we demonstrate here using a published database.

Keywords RNA sequencing, Gene expression omnibus, DESeq2, EdgeR, NOISEQ, Limma

The publication of the draft human genome in 2001 introduced a post-genomic era that promised global resolution of the molecular landscape in both preclinical and clinical research environments[1]. To capitalize on this new information, new technologies were developed; among them, the microarray, capable of profiling the simultaneous expression of tens of thousands of genes[2–6]. The endpoint of a microarray experiment is the identification of differentially expressed genes between two conditions[7,8]. The emergence of various RNA sequencing (RNA-seq) technologies substantially improved genome-wide transcriptomic analyses. Rather than relying on the hybridization of mRNA fragments to a library of prescribed probes, as with microarrays[3,4,9], RNA-seq turned the sample into a library, facilitating novel transcript discovery[8,10–15]. This high-throughput sequencing method provides several benefits compared to older transcriptomic analysis methods, including the ability to differentiate between isoforms of the same gene, identify coding and non-coding genes, and detect low-abundance transcripts[10,16–18]. The products of RNA-seq are often short, unaligned reads from fragments of the template material. These sequence data are stored in.FASTA or.FASTQ files. To characterize gene expression, reads must be aligned to a previously compiled, annotated, and species-specific reference genome. As the number of defined genomes increased after the assembly of the human genome, several genome alignment tools

[1]Department of Medicine, Queen's University, Kingston K7L 3N6, Canada. [2]Queen's CardioPulmonary Unit, Queen's University, Kingston K7L 3N6, Canada. [3]Dorothy Hodgkin Building, University of Bristol, Bristol, United Kingdom. [4]Queen's Health Sciences, Queen's University, Kingston K7L 3N6, Canada. [5]Department of Biomedical and Molecular Sciences, Queen's University, Kingston K7L 3N6, Canada. [6]Translational Institute of Medicine (TIME) Departments of Biomedical & Molecular Sciences and Medicine, Queen's Cardio Pulmonary Unit 1522, Queen's University, Kingston, ON K7L 3N6, Canada. [7]Abhishek Shastry and Benjamin P. Ott contributed equally to this work [8]Kimberly J. Dunham-Snary and Charles C. T. Hindmarch jointly supervised this work ✉email: kimberly.dunhamsnary@queensu.ca; c.hindmarch@queensu.ca

also emerged to assess the probability that sequenced reads match known genomic sequences[19–23]. After genome alignment, gene counts can be approximated using predictive models assessing the probability of alignment to exon sequences[24–28].

As with microarrays, differential expression analysis in RNA-seq studies compares gene counts between experimental samples through various statistical tests that were adopted from older experimental pipelines, which proved to be insufficient for this new data. In addition, variances in inter-sample library size, sequencing variability, and batch effects between biological replicates were often left unaccounted for in simple statistical analyses[29]. To address these issues, numerous bespoke analytical packages were developed that incorporated varying statistical assumptions about read count distributions, data normalization, and methods for hypothesis testing between samples[30]. In isolation, these analytical packages provide appropriate differential expression analysis based on experimental metadata and gene count input by the user. These packages can all be implemented using the R statistical programming language.

The use of next-generation RNA-seq technology to profile the transcriptome has become increasingly prominent in almost every domain of biological study. It is therefore important to provide non-bioinformaticians an avenue with which they can analyze large gene expression datasets without the requirement for advanced computational skillsets. Here, we built an intuitive web-based application using Shiny, called Confidence. Confidence is easy to use and answers each of the three challenges laid out above; 1. Confidence provides a way for users to model experimental metadata and visualize the clustering of samples used in RNA-seq experiments; 2. Confidence has a simple interface that allows count.csv files, and a metadata table to be uploaded; 3. Confidence provides sequential analysis of RNA-seq derived gene counts through multiple analytical packages, providing publication-ready figures and tables that can be directly downloaded to a local machine.

We demonstrate the utility of this software by mining a dataset from an independent group who have recently published an RNA sequencing dataset from pulmonary arterial smooth muscle cells (PASMC) from 5 patients with pulmonary arterial hypertension (PAH), and 4 healthy donors (GSE263226). The publication revealed 1,499 differentially regulated genes, and provided some pathway enrichment, identifying pathways involved in mitotic cell cycle and positive regulation of the cell cycle process. We reanalyze these data and demonstrate the utility of Confidence to reach the same conclusions.

Methods

Application development

We built the Confidence web application using the Shiny package v1.10.0[31] for R v4.4.3 (http://www.r-project.org), bslib v0.9.0, bsicons v0.1.2, DESeq2 v1.44.0[32], NOISeq v2.48.0[33,34], limma v3.60.6[35], edgeR v4.2.2[36], shinycssloaders v1.1.0 shinyjs v2.1.0, plotly v4.10.4[37], thematic v0.1.6, ggvenn v0.1.10[38], dplyr v1.1.4, enrichplot v1.24.4[39], shinyWidgets v0.9.0, spsComps v0.3.3.0, gprofiler2 v0.2.3[40], tidyr v1.3.1, eply v0.1.2, DT v0.33, ggrepel v0.9.6, ggplot2 v3.5.2[41], Matrix v1.7-3, SummarizedExperiment v1.34.0, Biobase v2.64.0, MatrixGenerics v1.16.0, matrixStats v1.5.0, GenomicRanges v1.56.2, GenomeInfoDb v1.40.1, IRanges v2.38.1, S4Vectors v0.42.1, BiocGenerics v0.50.0. Updates of software will happen periodically, and depreciated packages removed and replaced as appropriate.

Data input

Confidence requires users to input two files: one containing gene counts and the other containing experimental metadata. Gene counts must be formatted from count estimation computational tools in a tabular format, with column values representing sample names and rows representing genes; experimental metadata can be constructed as a comma-separated values file in programs such as Microsoft Excel®. Currently, only a single species can be run at one time.

Data modelling

Confidence provides extensive options to model experiment-associated variables found in the metadata, including selection of the main effect (the primary factor being investigated; e.g., treatment), reference class (the reference level for comparison within the main effect; e.g., untreated), and comparison class (the level to be compared to the reference class within the main effect; e.g., penicillin-treated). This is typically sufficient for two-factor comparisons (e.g., untreated vs. penicillin-treated). As a feature to guide the design of experimental questions, Confidence also allows users to build complex experimental designs controlling for multiple continuous or discrete factors (e.g., sex, age, diet) to isolate the impact of the primary variable of interest on gene expression while controlling for other factors.

Analytical packages

Confidence also requires the user to define a *P*-adj threshold and provides options for low count filtration based on the smallest comparison group size before initiation. The Confidence pipeline was developed to perform gene count normalization, filtration, and hypothesis testing sequentially across DESeq2 v1.44.0[32], limma v3.60.64[42,43], edgeR v4.2.2[36], and NOISeq v2.48.0[33,34], which are the most popular analytical packages used in published transcriptomic research. edgeR counts are modelled using a generalized linear model (GLM) quasi-likelihood pipeline through glmQLFit and glmQLFTest. DESeq2 uses a negative binomial GLM with a Wald test between the two levels of your main factor of interest. Limma first transforms counts to log2 Counts Per Million using voom, then fits a linear model, followed by Empirical Bayes Statistics and contrasts. Noiseq uses the NoiseqBIO pipeline. All the above methods can be applied to multifactor designs in Confidence. The default normalization settings and hypothesis tests were maintained in each analytical package (see Table 1) as per their respective vignettes[32–36,42]. Annotation of Ensembl gene IDs was facilitated using the biomaRt database v2.60.1[44–46]. A dropdown list of Human, Rat, and Mouse genome annotations are provided to characterize ENSEMBL gene IDs.

Platform	Original use	Proportion of false discoveries		Design formula complexity	Normalization; transformation	Multiple test correction method	Market share
		Small sample sizes	Large sample sizes				
DESeq2	RNA-seq,CHIP-seq	Low	Low	Multi-factor	DESeq2 Size-factors; Variance stabilizing transformation regularized log transformation	Holm, Hochberg, Hommel, Bonferroni, Benjamini-Hochberg, Benjamini-Yekutieli	26.36%
Limma	Microassays	Low	Low	Multi-factor	TMM; voom	Holm, Hochberg, Hommel, Bonferroni, Benjamini-Hochberg, Benjamini-Yekutieli	3.31%
edgeR	RNA-seq	Moderate	Moderate	Multi-factor	TMM, upper quartil, DESeq2-like: counts per million	Benjamini-Hochberg, Benjamini-Yekutieli, Holm	20.84%
NOISeq	RNA-seq	Moderate-high	Low-moderate	Multifactor	RPKM, TMM, upper quartile,None	A posteriori probability-akin to Benjamini-Hochberg FDR	1.38%

Table 1. Summary of key features and characteristics of analytical packages implemented in Confidence. The original use of each package was retrieved from respective manuals and vignettes. Proportion of false discoveries were generalized and run times were estimated and adapted from Seyednasrollah *et al.* (2015). Design formula complexity was observed through practical use of each package and consultation with package vignettes. Gene count normalization, transformation, and multiple test correction methods are summarized from the manuals and vignettes of each package; when multiple options are available, default options are bolded. Market share refers to the estimated percentage of total citations of an analytical package, adapted from survey of package citations by Costa-Silva *et al.* (2023). Abbreviations: RNA-seq - RNA sequencing; CHIP-seq - chromatin immunoprecipitation sequencing; TMM - trimmed mean of M values; voom - variance modeling at the observational level; RPKM - reads per kilobase per million mapped reads; FDR - false discovery rate.

This list will be expanded in the future to include more species. Due to the presence of uncharacterized mRNA reads in RNA-seq datasets, unannotated results that do not have gene names are also included in the Confidence results table, relying on their ENSEMBL gene ID annotation instead. An overview of the Confidence application methodology is presented in Figure 1.

Differential gene expression analysis

After contrasting samples based on selected main effects, each analytical package outputs a table of DEGs along with each gene's ID, $\log_2$FC, *P*-value, and *P*-adj, across conditions. NOISeq is an exception, however, as it produces an *a posteriori* probability *p* that a gene is differentially expressed between two conditions[33,34]. Here, we specify that the *P*-adj for NOISeq is equivalent to 1-*p*, as instructed in the NOISeq vignette.

Confidence score

After sequential tests of differential expression are performed by each analytical package, Confidence Scores are calculated for each DEG. The Confidence Score is a tally of the number of packages where a significant DEG is demonstrated to be up- or down-regulated. This 'consensus' among analytical packages allows for the filtering of genes that are likely to be both biologically and statistically significant. As an example, if three of the four packages implemented in Confidence report a DEG, the gene's associated Confidence Score is 3. Herein, we qualitatively describe genes with a Confidence Score of 0 as 'not confident', Confidence Score of 1 as 'least confident', a Confidence Score of 2 as 'moderately confident', a Confidence Score of 3 as 'highly confident', and a Confidence Score of 4 as 'most confident'. To elucidate overlaps between DEGs reported by each analytical package, a four-way Venn diagram is automatically generated by Confidence on the named list using the ggvenn package 0.1.10[47].

Outputs and visualizations

Confidence allows for the generation of standard visualizations found in transcriptomic literature. Principal component analysis (PCA) was integrated into Confidence using the DESeq2 plotPCA function and formatted using ggplot2[48]. PCA was performed using the *prcomp* function on variance stabilizing transformed gene counts; a PCA biplot is immediately available when experimental metadata and gene count files are submitted, so that the effect of including/excluding experimental variables and metadata factors (e.g., sex, diet, outliers, and low gene counts) can be visualized through sample clustering. To investigate the spread of DEGs according to their $\log_2$FC and *P*-adj values, volcano plots are generated on DEGs from each analytical package using ggplot2 v3.5.2 and plotly v4.10.4. Additionally, Box plots can be generated for each DEG to visualize normalized gene counts between experimental conditions; for box plots only, normalization is performed using $\log_2$-transformed, DESeq size-factors. These box plots are embedded into the final Confidence results table and are immediately accessible through a clickable button.

Pathway analysis

Pathway enrichment analysis (PEA) tools were additionally programmed into Confidence using the *GOSt* function from the gprofiler2 package v0.2.3[49]. *GOSt* analysis is provided to perform PEA using numerous databases including gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome (REAC), WikiPathways (WP), TRANScription FACtor database (TF), MicroRNA-Target database (miRTarBase),

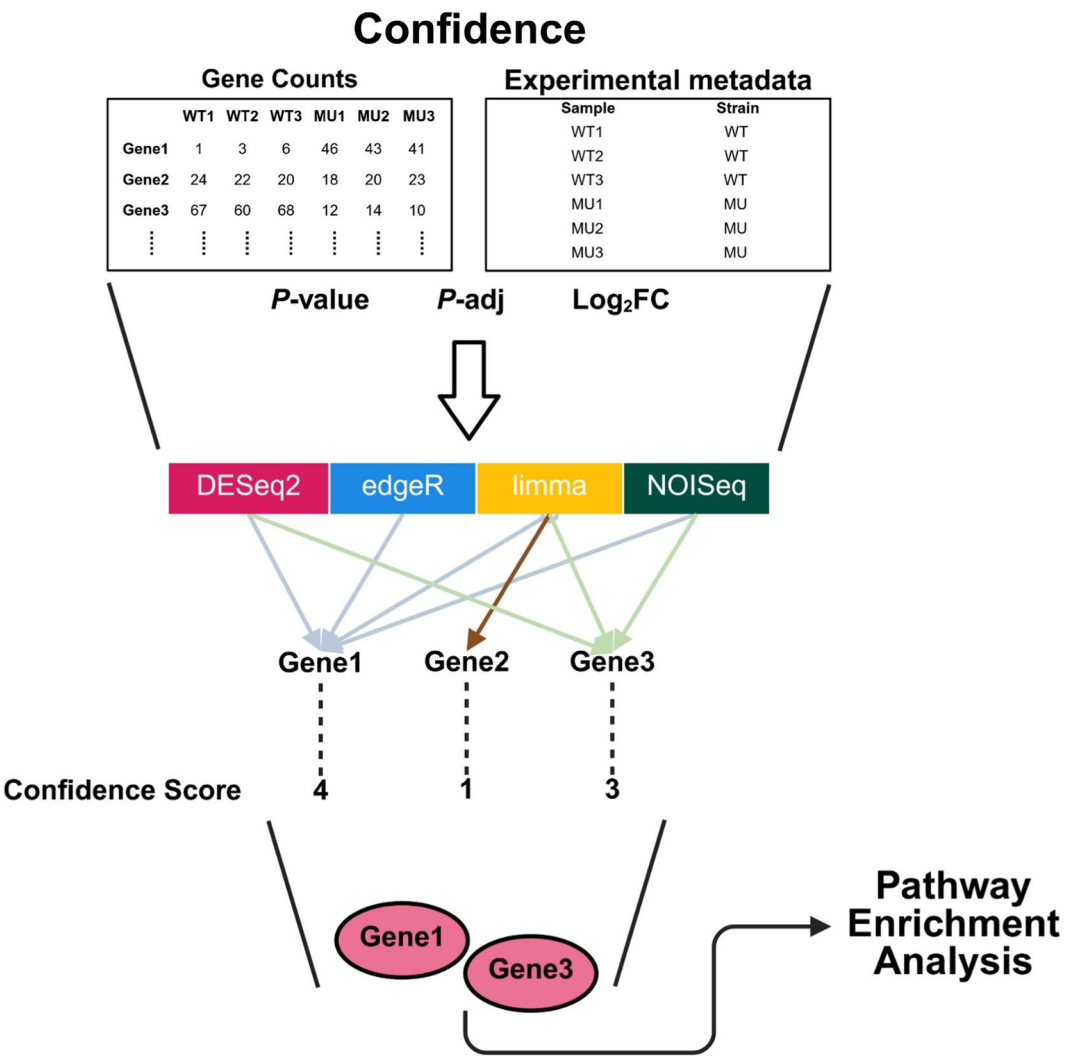

Fig. 1. Overview of the Confidence application. RNA sequencing technology allows for the quantification of mRNA transcripts, which are further aligned to a reference genome. This produces a gene counts table describing read abundance for each sample. Differential gene expression analysis requires a gene counts table, with samples as columns and gene IDs as rows, and experimental metadata describing sample groups. Confidence incorporates DESeq2, edgeR, limma, and NOISeq analytical packages in a single pipeline, and uses standard inputs such as *P*-value, adjusted *P*-value (*P*-adj), and $\log_2$(fold-change) ($\log_2$FC) thresholds. Differential expression analysis is performed sequentially across all packages. Confidence scores for each gene represent a tally of packages reporting that gene. Further, weighted scoring was implemented as a means to unify a gene's statistical significance and expression level into one score. Highly confident genes with high weighted scores can then be prioritized for further validation, manipulation, or characterization. Confidence also provides tools for visualization and pathway enrichment analysis using Kyoto Encyclopedia of Genes and Genomes and Gene Ontology libraries.

Human Protein Atlas (HPA), Comprehensive Resource of Mammalian protein complexes (CORUM), and Human Phenotype Ontology (HP), simultaneously.

Validation dataset preparation

To showcase Confidence, we wanted to analyze a dataset that was published by an independent group on a disease that we were familiar with. Pulmonary arterial hypertension is a fatal vasculopathy that prompts transcriptional changes in multiple tissues[13,50,51], and has an abundance of sequencing data available. We mined public datasets using search terms "pulmonary arterial hypertension," "RNA-seq," "PASMC," and "human" on NCBI's Gene Expression Omnibus (GEO). We selected GEO dataset GSE26366 from the retrieved results as it contained human untreated pulmonary artery smooth muscle cells (PASMC) from normal and PAH patient lungs samples. The raw FASTQ files were downloaded from Sequence Read Archive (SRA) using SRA Toolkit (sra-toolkit/3.0.9). Quality of the raw FASTQ files was assessed using FastQC (fastqc/0.12.1), and adapters and contaminants were then trimmed using BBDuk tool from BBMap (bbmap/39.06); reads with low quality (Phred

score < Q15) and reads shorter than 30 bp were removed. Post trimming quality was assessed using MultiQC (v1.33). The trimmed FASTQ files were then aligned to GRCh38.p14 (Genome Reference Consortium Human Build 38, patch release 14) reference genome using the STAR aligner software (star/2.7.11b). After aligning the reads, gene-level read counts were generated from the aligned BAM files using HTSeq software (HTSeq/2.0.5), producing a feature-by-sample count matrix with Ensembl gene IDs as rows and sample IDs as columns.

The individual count files were then combined using the dplyr package (v.1.2.0) in R to create the final count matrix with all samples in one CSV file. Additionally, metadata for this dataset was retrieved using the GEOquery (v.2.78.0) and curated to construct a structured metadata table containing sample IDs matching the count matrix, as well as genotype and diagnosis information for each sample.

Finally, differential gene expression analysis was performed within the Confidence application using the processed count matrix (Supplementary file 1) and curated metadata (Supplementary file 2) CSV files. Both these files can also be found on the help page (Supplementary file 3) of the Confidence graphic user interface. The resulting findings are presented in the Results section and were compared with the original publication to evaluate and validate the results from the application (Figures 2 and 3).

Results of validation data analysis

Confidence can be found, here: https://confidence.apps.meds.queensu.ca/

To showcase Confidence, we included figures generated within the application (Figure 2A–F and Figure 3A-D) using data from the publicly available validation dataset (GSE263226) reported by Lemay et al. (2025)[52]. The Lemay et al. study investigated the transcriptome of pulmonary artery smooth muscle cells (PASMCs), identifying AURKB as a potential therapeutic target. They performed gene expression profiling on PASMCs from patients with pulmonary arterial hypertension (PAH) (n = 5) compared to PASMCs from healthy control donors (n = 4). We used Confidence to model these data using principal component analysis (Figure 2A/B). Confidence identified 257 genes with CS=4 but also demonstrated that each tool called genes that other tools do not; for example, DESeq2 calls 4,062 genes uniquely, which the other tools do not identify as significant (Figure 2C). Differentially regulated genes are presented in a Volcano plot which allows user discrimination of genes identified by each statistical test (Figure 2D).

Consistent with the original publication, Confidence identified AURKB, FOXM1, BIRC5, and PLK1 as significantly upregulated, and SOD2 as downregulated. We have used Confidence to select AURKB and SOD2, which we have visualized using a box-and-whisker plot (Figure 2E-F).

Confidence also visualizes functional enrichment results as an interactive dot plot of enriched terms organized by domain (Figure 3A), as well as user-defined dot plots displaying the top functions, processes, or regulatory targets within each domain; here we display the output from KEGG (Figure 3B) and GO:BP (Figure 3C). Pathway enrichment analysis highlighting the functional analysis of the 257 robust genes that Confidence highlighted are involved in the mitotic cell cycle and positive regulation of cell cycle processes, consistent with the findings reported in the Lemay paper[52].

Discussion

Differential expression analysis is often the penultimate step of transcriptomic studies, which statistically evaluates the differences in gene counts between experimental groups. Broadly, there are three key challenges to the use of analytical packages for differential expression analysis: *1. Modelling variability:* Transcriptomic studies using deep RNA-seq often incorporate multiple biological replicates to power the study. Therefore, there may be variability associated with sample replicates (e.g., heterogeneous disease expression between replicates), library sizes, and read quality, among other factors. Additionally, the presence of genes with abnormally skewed count distributions may cause true differentially expressed genes (DEGs) to be misclassified or overlooked. Most analytical packages incorporate preprocessing that includes batch correction, normalization, multiple test correction, and transformation to account for these factors[53], however hypothesis testing procedures differ between discrete analysis packages, and so does the expression level (e.g. log2FC, log-ratio), and the significance value reported for each gene[30,54,55]. *2. Expert analysis:* Appropriate use of analytical packages for transcriptomics requires a substantial bioinformatics and programming background. Indeed, qualitative studies have demonstrated that the complexity of software, lack of formal training, and decreased perceived ease-of-use of bioinformatics tools reduce non-bioinformaticians' ability to perform bioinformatics analyses such as functional enrichment and/or pathways analysis[56–58]. *3. Target prioritization:* A primary goal of transcriptomic analyses is the identification of genes whose differential regulation contributes to the experimental condition under investigation. However, the list of DEGs can be substantial. Genes of interest from these lists have traditionally been chosen arbitrarily through *P*-value, *P*-adj, and/or expression level parameters; alternatively, genes are often selected through the familiarity of the gene or its related protein to the investigator. Without a systematic, unbiased method to rank and delineate genes of interest, genes or gene sets that are reflective of real biological changes and play novel roles in the disease pathology, treatment intervention, or biological model may be overlooked.

Since the publication of the first RNA-seq study in 2006[59], several analytical packages and methodologies have been developed to normalize, model, and test differential expression effectively[60]. However, large variations in the number of discovered genes, false discovery rate, and expression changes across these analytical packages may confound the interpretation of results[30,54,55,61–64]. Among these transcriptomics analytical packages, DESeq2, limma, edgeR, and NOISeq are commonly used in deep RNA-seq studies[63–65]. Additionally, DESeq2 accounted for 26.36%, edgeR for 20.84%, limma-voom for 3.13%, and NOISeq for 1.38% of all citations for RNA-seq studies[60]. NOISeq, the most popular non-parametric method, was implemented into Confidence as it

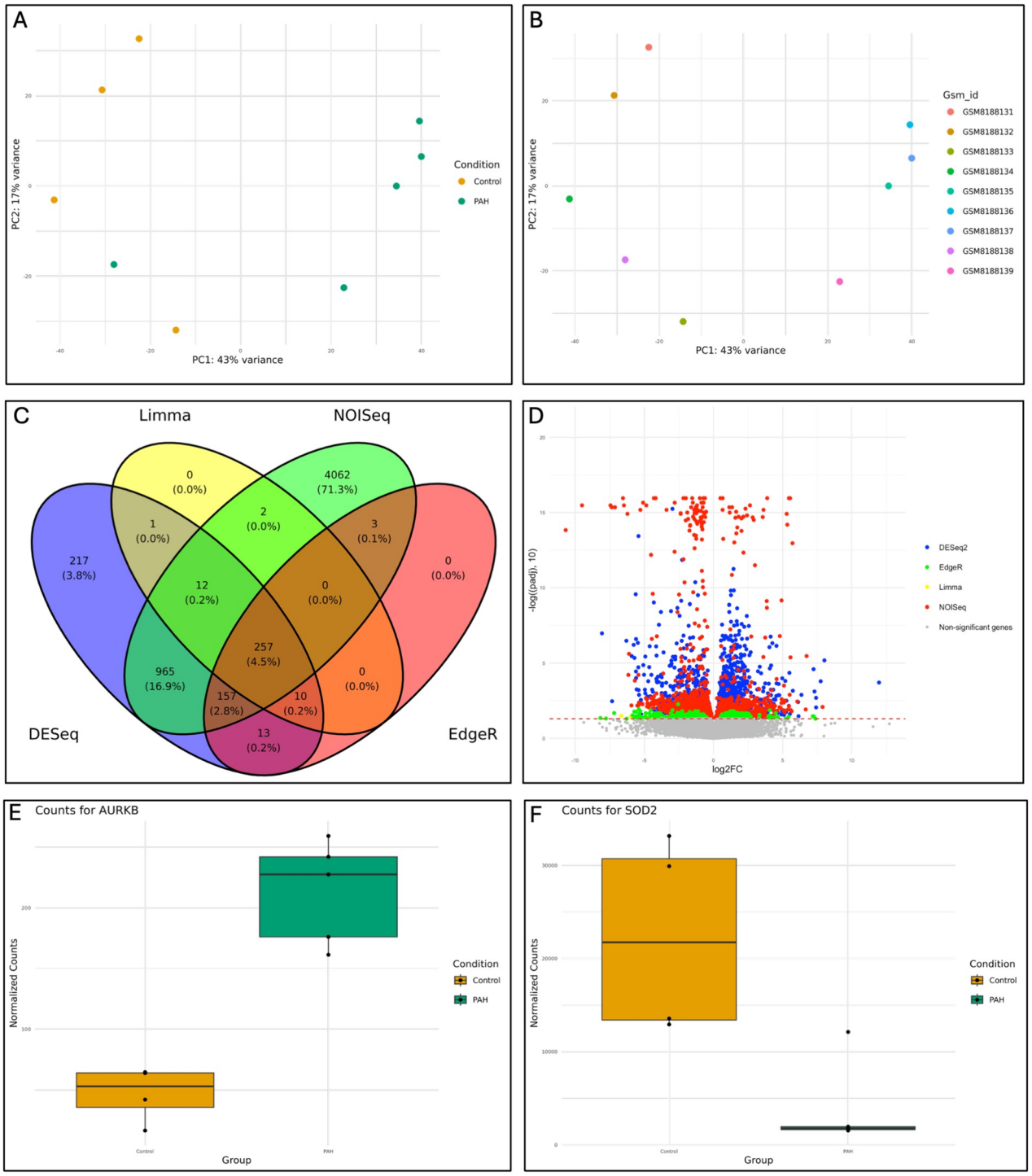

Fig. 2. Confidence allows the generation and download of high-resolution, publication-ready images. A publicly available dataset from the Gene Expression Omnibus (GSE26366), provided by Lemay et al. (2025), was used as a validation dataset, and original results from this dataset and publication are displayed in (**A**) Analyzing the raw counts and metadata through Confidence produced similar results, beginning with the generation of principal component analysis (PCA), which allows users to better model their data in line with the dependent and independent variables stated in the metadata they have uploaded. For example, the data can be coloured according to (**B**) condition or (**C**) sample ID (GSM ID). Visualizing data in this manner allows researchers to identify systemic problems with their experiment that might prompt improved modelling or more careful experimental design. Confidence also generates a (**D**) Venn diagram to show the number of genes predicted by each test and the intersections between tests, as well as an (**E**) volcano plot coloured according to the statistical test that predicted each gene. G/H. Users can select genes from the list and generate box plots based on the normalized count data.

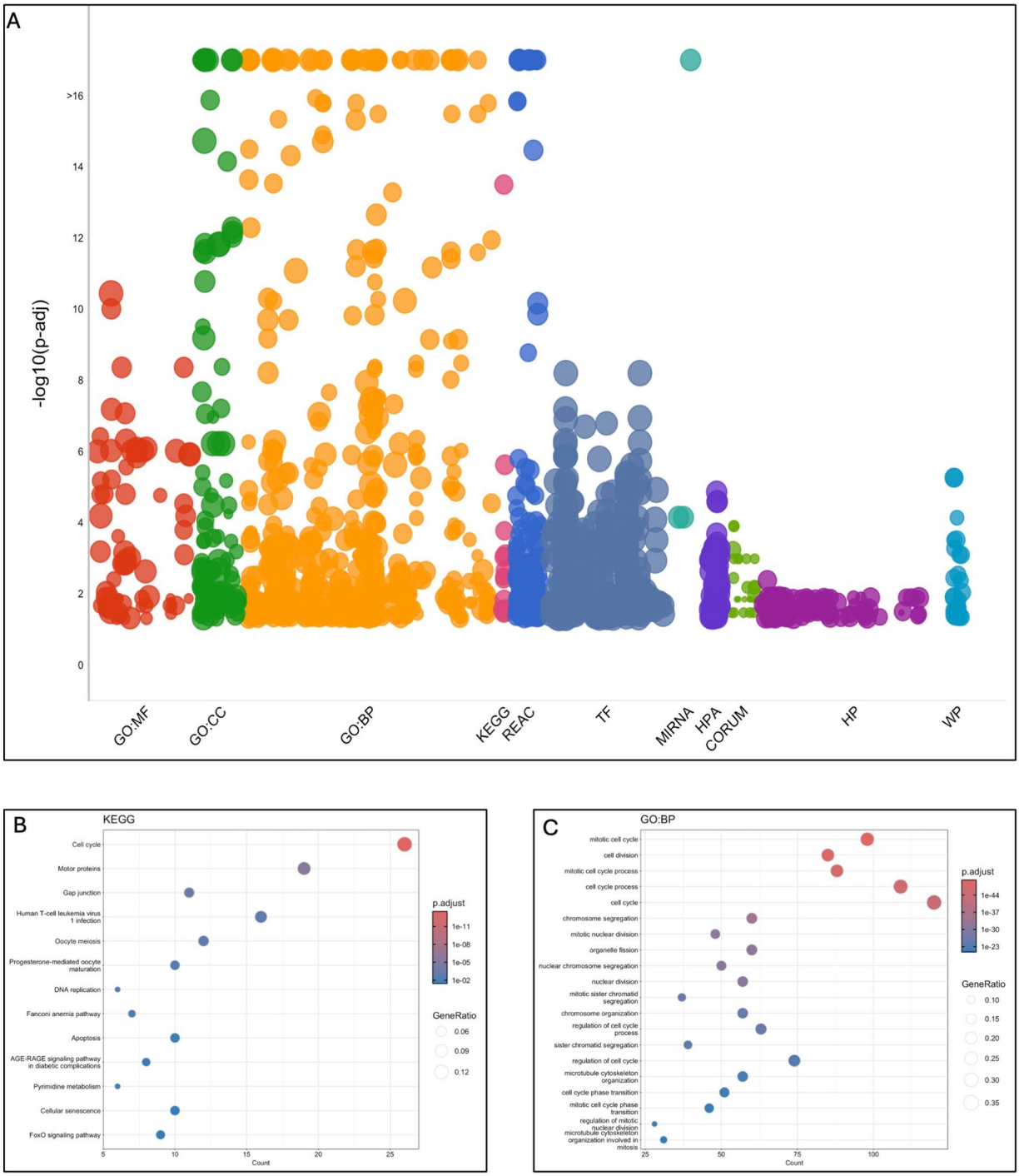

Fig. 3. Confidence takes significantly regulated genes and performs functional enrichment analysis on these data. Confidence identified overlapping enriched GO terms, as shown in (**A**) which displays an interactive plot of enriched Gene Ontology (GO) nodes across the molecular function (MF), cellular component (CC), and biological process (BP) domains. Enriched terms from the Kyoto Encyclopedia of Genes and Genomes (KEGG), Reactome (REACT), transcription factors (TF), microRNAs (miRNA), the Comprehensive Resource of Mammalian Protein Complexes (CORUM), the Human Protein Atlas (HP), and WikiPathways (WP) are also displayed. Each enrichment category can be interrogated and visualized as a dot plot of the most significantly enriched terms within that database. Here, (**B**) KEGG and (**C**) GO:BP enrichment plots are shown.

offers a unique computational and theoretical approach relative to parametric methods such as DESeq2, edgeR, and limma.

In untargeted omics analyses more broadly, the process to select features to characterize, validate, or manipulate must take several considerations into account, including the cost of reagents and the considerable amount of time needed to investigate findings. We propose that Confidence can provide an intuitive means to filter large gene lists and streamline gene selection. Additionally, although many analytical package-consensus tools have been developed[66,67], they are implemented as computational packages within coding languages such as Python or R. In a survey conducted by Alomair *et al.* (2023) of scientists, only one-third used programming languages to address biological problems, and half reported hiring a bioinformatician to conduct the analysis[56]. To address this issue, Confidence was developed as a web-based application, which included options to produce the visualizations presented in this study, as well as clear descriptions of analytical methods, plot interpretations, and plot customization tools. One limitation of Confidence that we feel necessary to raise, is that the Confidence Score reflects inter-method agreement rather than biological relevance; the analysis by default equally weights methods.

Confidence addresses three major challenges to identify and prioritize DEGs from transcriptomic studies: Firstly, Confidence allows the easy modelling of data using different sample attributes reported in the user-defined metadata; PCA's will allow for the rapid review of sample variability using condition, as well as other information such as sex, batch, animal ID, and time. Secondly, Confidence integrates standard pipelines for differential expression analysis into a simple web-based platform, which may ensure that differential expression analysis is more accessible for investigators without programming experience. Lastly, using the Confidence Score, Confidence provides a consensus tool to identify common DEGs across analytical packages employing diverse statistical methodologies, enhancing target prioritization.

Conclusions

We have developed the Confidence application to perform multi-package differential expression analysis, provide intuitive scoring systems to streamline gene prioritization and selection, and provide a web-based tool for non-bioinformaticians to analyze transcriptomics data. Further development of Confidence will facilitate the integrate the vast array of analytical packages available in other omics fields, such as single cell transcriptomics, proteomics, methylomics, lipidomics, and metabolomics (to name a few). Finally, we foresee the potential that the ability for Confidence to streamline marker selection without the influence of knowledge bias may reduce costs associated with unproductive validation and experimentation of these data.

Data availability

The Confidence application is available at https://confidence.apps.meds.queensu.ca/. We have included the meta data and count data for the Lemay paper (*52*) on the website, and in the supplementary data of this paper. Any questions regarding code availability and the conclusions of this paper should be directed to the corresponding author.

Received: 18 December 2025; Accepted: 21 April 2026
Published online: 28 April 2026

References

1. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Pozhitkov, A. E., Tautz, D. & Noble, P. A. Oligonucleotide microarrays: Widely applied--Poorly understood. *Brief. Funct. Genomic. Proteomic.* **6**, 141–148 (2007).
3. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
4. Ghorbel, M. T. et al. Microarray screening of suppression subtractive hybridization-PCR cDNA libraries identifies novel RNAs regulated by dehydration in the rat supraoptic nucleus. *Physiol. Genomics* **24**, 163–172 (2006).
5. Hindmarch, C. C. et al. The transcriptome of the medullary area postrema: The thirsty rat, the hungry rat and the hypertensive rat. *Exp. Physiol.* **96**, 495–504 (2011).
6. Stewart, L. et al. Hypothalamic transcriptome plasticity in two rodent species reveals divergent differential gene expression but conserved pathways. *J. Neuroendocrinol.* **23**, 177–185 (2011).
7. Hindmarch, C. C. T. & Murphy, D. The transcriptome and the hypothalamo-neurohypophyseal system. *Endocr. Dev.* **17**, 1–10 (2010).
8. Pauža, A. G. et al. Osmoregulation of the transcriptome of the hypothalamic supraoptic nucleus: A resource for the community. *J Neuroendocrinol* **33**, e13007 (2021).
9. Liu, H., Bebu, I. & Li, X. Microarray probes and probe sets. *Front Biosci (Elite Ed)* **2**, 325–338 (2010).
10. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
11. Bentley, R. E. T., Hindmarch, C. C. T. & Archer, S. L. Using omics to breathe new life into our understanding of the ductus arteriosus oxygen response. *Semin. Perinatol.* **47**, 151715 (2023).
12. Kloosterman, R. et al. A transcriptome analysis of basal and stimulated VWF release from endothelial cells derived from patients with type 1 VWD. *Blood Adv.* **7**, 1477–1487 (2023).
13. Potus, F., Hindmarch, C. C. T., Dunham-Snary, K. J., Stafford, J. & Archer, S. L. Transcriptomic signature of right ventricular failure in experimental pulmonary arterial hypertension: Deep sequencing demonstrates mitochondrial, fibrotic, inflammatory and angiogenic abnormalities. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms19092730 (2018).
14. Jiang, L. et al. RNA sequencing analysis of human podocytes reveals glucocorticoid regulated gene networks targeting non-immune pathways. *Sci. Rep.* **6**, 35671 (2016).
15. Johnson, K. R. et al. A RNA-Seq analysis of the rat supraoptic nucleus transcriptome: Effects of salt loading on gene expression. *PLoS One* **10**, e0124523 (2015).
16. Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* **9**, e78644 (2014).

17. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).
18. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
19. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
20. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
21. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**, R36 (2013).
22. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
23. Dobin, A. et al. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
24. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
25. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
26. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
27. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
28. Anders, S., Pyl, P. T. & Huber, W. HTSeq--A Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
29. Jeanmougin, M. et al. Should we abandon the t-test in the analysis of gene expression microarray data: A comparison of variance modeling strategies. *PLoS ONE* **5**, e12336 (2010).
30. Seyednasrollah, F., Laiho, A. & Elo, L. L. Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief. Bioinform.* **16**, 59–70 (2015).
31. W. Chang *et al.*, in *R package version 1.9.1.9000*. (2024).
32. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**550. (2014).
33. Tarazona, S. et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res.* **43**, e140 (2015).
34. Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: A matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
35. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
36. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. EdgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
37. C. Sievert *et al.*, plotly: Create Interactive Web Graphics via plotly.js. (2024).
38. L. Yan, ggvenn: Draw Venn Diagram by ggplot2. (2023).
39. G. Yu, enrichplot: Visualization of Functional Enrichment Result. (2024).
40. L. Kolberg, U. Raudvere, gprofiler2: Interface to the g:Profiler Toolset. (2024).
41. H. Wickham *et al.*, ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. (2024).
42. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
43. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
44. R. J. Kinsella *et al.*, Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)* **2011**, bar030 (2011).
45. Smedley, D. et al. BioMart--Biological queries made easy. *BMC Genomics* **10**, 22 (2009).
46. J. Zhang *et al.*, BioMart: a data federation framework for large collaborative projects. *Database (Oxford)* **2011**, bar038 (2011).
47. Gao, C.-H., Yu, G. & Cai, P. ggVennDiagram: An intuitive, easy-to-use, and highly customizable R package to generate Venn diagram.. *Front. Genet.* https://doi.org/10.3389/fgene.2021.706907 (2021).
48. K. Blighe, A. Lun, in *R package version 2.16.0*. (2024).
49. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. Gprofiler2 -- An R package for gene list functional enrichment analysis and namespace conversion toolset G:Profiler. *F1000Res.* https://doi.org/10.12688/f1000research.24956.1 (2020).
50. Hindmarch, C. C. T. et al. Tet methylcytosine dioxygenase 2 (TET2) mutation drives a global hypermethylation signature in patients with pulmonary arterial hypertension (PAH): Correlation with altered gene expression relevant to a common T cell phenotype. *Compr. Physiol.* **15**, e70011 (2025).
51. Hindmarch, C. C. T. et al. An integrated proteomic and transcriptomic signature of the failing right ventricle in monocrotaline induced pulmonary arterial hypertension in male rats. *Front. Physiol.* **13**, 966454 (2022).
52. Lemay, S. E. et al. Unraveling AURKB as a potential therapeutic target in pulmonary hypertension using integrated transcriptomic analysis and pre-clinical studies. *Cell Rep. Med.* **6**, 101964 (2025).
53. Abrams, Z. B., Johnson, T. S., Huang, K., Payne, P. R. O. & Coombes, K. A protocol to evaluate RNA sequencing normalization methods. *BMC Bioinformatics* **20**, 679 (2019).
54. Soneson, C. & Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* **14**, 91 (2013).
55. Kvam, V. M., Liu, P. & Si, Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* **99**, 248–256 (2012).
56. Alomair, L. & Abolfotouh, M. A. Awareness and predictors of the use of bioinformatics in genome research in Saudi Arabia. *Int. J. Gen. Med.* **16**, 3413–3425 (2023).
57. Shachak, A., Shuval, K. & Fine, S. Barriers and enablers to the acceptance of bioinformatics tools: A qualitative study. *J. Med. Libr. Assoc.* **95**, 454–458 (2007).
58. Wijesooriya, K., Jadaan, S. A., Perera, K. L., Kaur, T. & Ziemann, M. Urgent need for consistent standards in functional enrichment analysis. *PLoS Comput. Biol.* **18**, e1009935 (2022).
59. Bainbridge, M. N. et al. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* **7**, 246 (2006).
60. Costa-Silva, J., Domingues, D. S., Menotti, D., Hungria, M. & Lopes, F. M. Temporal progress of gene expression analysis with RNA-Seq data: A review on the relationship between computational methods. *Comput. Struct. Biotechnol. J.* **21**, 86–98 (2023).
61. Corchete, L. A. et al. Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Sci. Rep.* **10**, 19737 (2020).
62. Nookaew, I. et al. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: A case study in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **40**, 10084–10097 (2012).

63. Li, D. et al. An evaluation of RNA-seq differential analysis methods. *PLoS One* https://doi.org/10.1371/journal.pone.0264246 (2022).
64. Das, S., Rai, A., Merchant, M. L., Cave, M. C. & Rai, S. N. A comprehensive survey of statistical approaches for differential expression analysis in single-cell RNA sequencing studies. *Genes* https://doi.org/10.3390/genes12121947 (2021).
65. Huang, H. C., Niu, Y. & Qin, L. X. Differential expression analysis for RNA-Seq: An overview of statistical methods and computational software. *Cancer Inform.* **14**, 57–67 (2015).
66. Waardenberg, A. J. & Field, M. A. consensusDE: An R package for assessing consensus of multiple RNA-seq algorithms with RUV correction. *PeerJ* **7**, e8206 (2019).
67. Costa-Silva, J., Domingues, D. & Lopes, F. M. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS One* **12**, e0190152 (2017).

Acknowledgments
We would like to thank the Environment Genome Interface, and the MitoMetabLab labs at Queen's University for their suggestions and beta testing of the Confidence application. We would also like to thank the Queen's CardioPulmonary Unit (QCPU), and the Translational Institute of Medicine (TIME) at Queen's University for hosting in perpetuity this software.

Author contributions
Conceptualization: A.S., B.P.O, K.DS., AP., C.C.T.H. Methodology: A.S., B.P.O., K.DS., C.C.T.H. Investigation: A.S., B.O., C.C.T.H. Visualization: B.P.O., A.S. TN. Formal Analysis: A.S., B.P.O, TN. Project administration: A.S., B.P.O., M.S., K.DS., C.C.T.H. Resources: B.P.O., M.S., K.DS., C.C.T.H. Software: B.P.O., A.S., TN., C.C.T.H. Supervision: K.DS., C.C.T.H. Writing – original draft: A.S., K.DS., C.C.T.H. Writing – review & editing: A.S., M.S., B.P.O, K.DS., C.C.T.H.

Declarations

Competing interests
Author AP is the owner of a bioinformatics consultancy in the United Kingdom; AP's work on this paper predates this company and so the consultancy was not involved in the development, commercialization, or application of the software described in this manuscript. The authors therefore declare no competing financial interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-50527-w.

Correspondence and requests for materials should be addressed to K.J.D.-S. or C.C.T.H.