THESIS

AN INTEGRATIVE RNA-SEQ PIPELINE FOR LINKING MICRORNA AND MRNA DIFFERENTIAL EXPRESSION

Submitted by

Paige A. Lillibridge

Department of Biology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Summer 2026

Master's Committee:

Advisor: Taiowa Montgomery

Ann Hess
Kim Hoke

ABSTRACT

AN INTEGRATIVE RNA-SEQ PIPELINE FOR LINKING MICRORNA AND MRNA DIFFERENTIAL EXPRESSION

High-throughput sequencing technologies enable parallel measurement of small RNA and messenger RNA (mRNA) expression; however, most RNA-seq analysis workflows evaluate these data types independently. While differential expression analyses identify transcripts that change between conditions, they often do not directly address the regulatory relationships between microRNAs (miRNAs) and their predicted mRNA targets. This separation can limit biological interpretation and make it harder to identify meaningful relationships between miRNAs and their target genes.

To address this challenge, RNA-integrate was developed as a reproducible and flexible RNA-seq workflow that combines DESeq2-based differential expression analysis of small RNA and mRNA expression data. The workflow is fully parameterized through external YAML configuration files and supports multiple RNA-seq quantification formats, allowing application across diverse datasets and experimental designs. Following differential expression analysis, miRNAs are linked to predicted target mRNAs using a user-defined gene-target table, generating paired datasets for integrative analysis.

RNA-integrate produces a diverse set of outputs, including differential expression tables, quality-control visualizations, integrative results tables, interactive HTML reports, and publication-quality figures. Integrative analyses include cosmic plots and slope plots, which enable visualization of expression relationships between miRNAs and their predicted targets. In

addition, the workflow implements a one-sided binomial sign test to evaluate whether inverse miRNA–mRNA relationships occur more frequently than expected by chance.

To demonstrate the utility of the workflow, RNA-integrate was applied to a *pash-1* mutant versus wild-type dataset in *Caenorhabditis elegans*. Because *pash-1* is required for miRNA biogenesis, disruption of *pash-1* provides a biologically relevant system for evaluating miRNA-mediated regulation. Analysis revealed a strong enrichment of inverse relationships among significant miRNA–mRNA target pairs, consistent with established models of post-transcriptional repression by miRNAs.

By combining reproducible workflow design, differential expression analysis, integrative visualization, and statistical evaluation within a single framework, RNA-integrate provides an accessible platform for investigating RNA regulatory networks. The workflow facilitates cross-dataset analyses, improves reproducibility, and enables a flexible framework of miRNA–mRNA interactions across organisms and experimental systems.

‘

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Dr. Taiowa Montgomery, for his mentorship, guidance, and support throughout this project and my graduate studies. His encouragement and confidence in my abilities helped me grow as a researcher and were invaluable to the completion of this work.

I am also grateful to my committee members, Dr. Ann Hess and Dr. Kim Hoke, for their expertise, thoughtful feedback, and support throughout this project. Their guidance helped strengthen both this work and my development as a scientist.

I would like to thank the members of the Montgomery Laboratory for their support throughout the development of RNA-Integrate. Special thanks to Spencer Kuhn, who developed the original version of this pipeline and provided the foundation upon which this work was built.

Finally, I would like to thank my family and friends for their unwavering encouragement and support throughout my graduate studies. I am especially grateful to my mom, dad, and stepmom, whose careers in science inspired my own interest in research from an early age. This project would not have been possible without the support, guidance, and opportunities provided by so many people along the way.

TABLE OF CONTENTS

ABSTRACT ii
ACKNOWLEDGEMENTS iv
INTRODUCTION 1
METHODS 5
RNA Sequencing 5
Software Design and Implementation 7
Quality Control and Diagnostic Visualizations 9
Principal Component Analysis (PCA) 10
Intra-Condition Scatterplots 10
Mean Expression Plots 10
MA Plots 10
Heatmaps 10
Target Pair Filtering 10
Slope Plots 12
Reproducibility and Output Generation 12
Overview of Differential Expression Analysis 13
Integrative Analysis of Small RNA and mRNA Relationships 18
DISCUSSION 23
CONCLUSION 26
DATA AVAILABILITY 27
CODE AVAILABILITY 28
REFERENCES 29

INTRODUCTION

MicroRNAs (miRNAs) are one of the most extensively studied classes of small RNAs and function primarily at the post-transcriptional level (Bartel, 2009). miRNAs play critical roles in regulating gene expression across eukaryotic organisms (Bartel, 2004; Bartel, 2009). Mature miRNAs are typically 18–22 nucleotides in length and regulate gene expression through sequence-specific interactions with target messenger RNAs (mRNAs) (Bartel, 2004). These interactions can result in translational repression, mRNA destabilization, or transcript degradation depending on the degree of sequence complementarity and cellular context (Bartel, 2004; Bartel, 2009). Through these mechanisms, miRNAs influence diverse biological processes including development, differentiation, metabolism, and stress responses.

miRNA biogenesis is a multi-step process that begins with transcription of primary miRNA transcripts (pri-miRNAs) within the nucleus. These transcripts are processed by the Microprocessor complex, which contains the endoribonuclease Drosha and the RNA-binding protein DGCR8 (Pasha in invertebrates) and cleaves pri-miRNAs to generate precursor miRNAs (pre-miRNAs) (Lee et al., 2003; Gregory et al., 2004; Denli et al., 2004). Pre-miRNAs are subsequently exported to the cytoplasm and further processed by the endoribonuclease Dicer to produce mature miRNAs (Ha & Kim, 2014). Mature miRNAs are incorporated into the RNA-induced silencing complex (RISC), where they guide sequence-specific repression of target mRNAs (Figure 1) (Bartel, 2004; Bartel, 2009).

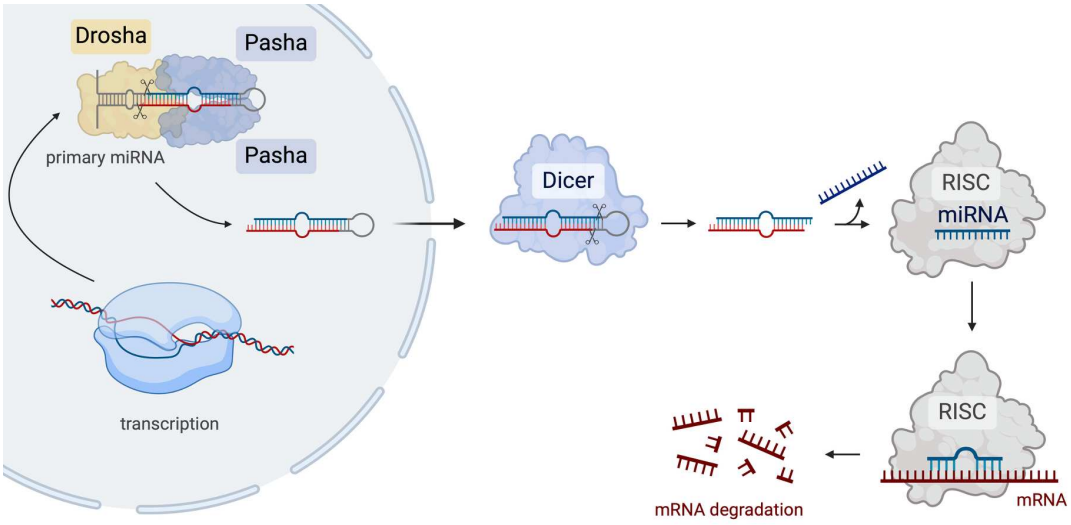

Figure 1. Overview of miRNA biogenesis and function. Primary miRNA transcripts (pri-miRNAs) are processed in the nucleus by the Drosha–Pasha Microprocessor complex to generate precursor miRNAs (pre-miRNAs), which are exported to the cytoplasm and further processed by Dicer into mature miRNAs. Mature miRNAs are incorporated into the RNA-induced silencing complex (RISC), where they guide sequence-specific repression of target mRNAs.

Advances in next-generation sequencing (NGS) technologies have enabled high-throughput measurement of both miRNA and mRNA expression across biological conditions. RNA sequencing (RNA-seq) provides quantitative information that can be used to identify differentially expressed genes and transcripts and has become a central tool in modern molecular biology (Conesa et al., 2016). By combining mRNA and small RNA sequencing and differential expression analysis, the regulatory impact of miRNAs can be accurately assessed on a genome-wide scale. Among the most widely used software packages for RNA-seq differential expression analysis is DESeq2, which models count data using a negative binomial distribution and employs shrinkage-based estimation procedures to improve dispersion and fold-change inference (Love et al., 2014). Because of its statistical rigor and flexibility, DESeq2 has become a standard framework for RNA-seq analysis across diverse experimental designs.

Despite these advances, most RNA-seq workflows analyze small RNA and mRNA datasets independently. While this approach effectively identifies differentially expressed features within each dataset, it often fails to directly address the regulatory relationships that connect miRNAs to their mRNA targets. This limitation is particularly important because regulation of mRNA expression represents one of the primary biological functions of miRNAs. Consequently, understanding how changes in miRNA abundance influence mRNA responses remains a critical challenge in transcriptomic analysis.

To address this challenge, several computational approaches have been developed to integrate miRNA and mRNA datasets. Most rely on predicted miRNA–mRNA interactions generated by sequence-based algorithms such as TargetScan (Lewis et al., 2005) and miRanda (Enright et al., 2003). These tools identify potential regulatory targets, although predicted interactions do not always correspond to biologically meaningful regulation and therefore require additional validation through expression-based and genetic analyses.

More recent integrative frameworks, including mirTarRnaSeq (Movassagh et al., 2022) and TimiRGeN (Patel et al., 2021), combine differential expression data with statistical approaches to evaluate support for predicted regulatory interactions. These tools are important advances in integrative transcriptomics; however, they are frequently implemented as standalone applications, often require substantial preprocessing, and may be difficult to incorporate into standard RNA-seq workflows. Reproducibility and workflow portability have become increasingly relevant challenges in computational biology as datasets and analytical complexity continue to grow (Perkel, 2019). Consequently, there is a growing need for integrated analysis frameworks that prioritize transparent, reproducible, and easily transferable workflows.

A major challenge in integrative RNA analysis is distinguishing true regulatory relationships from background noise (Bartel, 2009; Movassagh et al., 2022). Because miRNAs are generally expected to repress their targets, inverse expression patterns between miRNAs and mRNAs are often considered evidence of regulatory activity. However, these patterns can be obscured by indirect effects, incomplete target predictions, biological variability, and experimental noise.

These challenges highlight the need for a reproducible and flexible framework capable of integrating small RNA and mRNA datasets while remaining compatible with established differential expression workflows. Such a framework should support diverse input formats, provide interpretable visualizations, generate reproducible outputs, and facilitate quantitative evaluation of regulatory relationships.

In this study, I present RNA-Integrate, a modular RNA-seq workflow that combines DESeq2-based differential expression analysis with downstream integration of small RNA and mRNA expression data. The workflow is fully parameterized through external YAML configuration files, enabling reproducible execution and straightforward adaptation across datasets, organisms, and experimental designs. Following differential expression analysis, RNA-integrate links small RNAs and mRNAs using a user-defined gene-target table and generates integrative visualizations, including cosmic plots and slope plots, to facilitate interpretation of regulatory relationships.

In addition to visualization, RNA-integrate incorporates statistical evaluation through a one-sided binomial sign test that quantifies whether inverse small RNA–mRNA relationships occur more frequently than expected by chance. The workflow also produces publication-ready

figures, interactive HTML visualizations, results tables, and parameter snapshots that preserve complete records of analysis settings and outputs.

By combining reproducible workflow design, differential expression analysis, integrative visualization, and statistical evaluation within a single framework, RNA-integrate provides a flexible and interpretable platform for investigating RNA regulatory networks. As a proof-of-principle, I applied the RNA-integrate workflow developed in this study, to small RNA and mRNA RNA-seq datasets from wild-type *C. elegans* and DGCR8/*pash-1* mutants. The *pash-1* mutation is a hypomorph that only partially impairs miRNA processing, resulting in modest reductions in mature miRNA abundance while maintaining organismal viability and overall animal health (Knittel et al., 2025). Therefore, these datasets allow for analysis of changes in miRNA target expression while limiting confounding effects resulting from developmental abnormalities. These data provide a proof-of-concept model for evaluating whether our integrative analysis workflow can recover expected regulatory relationships between miRNAs and their predicted mRNA targets.

METHODS

RNA Sequencing

mRNA sequencing was done for a separate, unpublished project. Briefly, total RNA was isolated from gravid wild-type and *pash-1* mutant adult *C. elegans* and treated with DNase (ThermoFisher AM2238). Ribosomal RNA was depleted using Ribo-Zero (Illumina 20020596). Libraries were prepared using the Illumina TruSeq Stranded Total RNA kit (Illumina 20020596). Samples were sequenced by Novogene on an Illumina NovaSeq X Plus platform using paired-end 150 bp reads. The resulting sequencing data were used as input for differential expression

and integrative analyses. The small RNA sequencing data was previously described (Knittel et al., 2025).

RNA-Integrate

RNA-integrate is a modular RNA-seq workflow developed to integrate differential expression results from small RNA and mRNA sequencing experiments. The workflow consists of three primary components: input processing and validation, DESeq2-based differential expression analysis, and downstream integration of small RNA–mRNA regulatory relationships. All workflow components can be controlled through external YAML configuration files, allowing reproducible execution across datasets, organisms, and experimental designs without modification of source code.

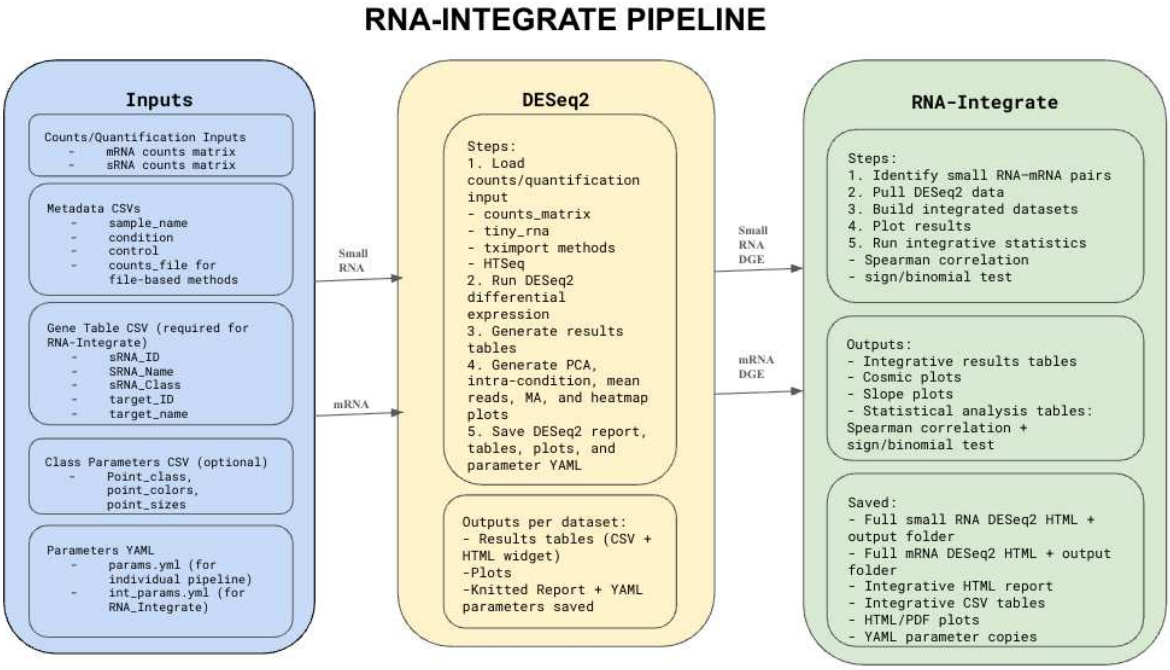

Figure 2. Overview of the RNA-integrate workflow. The workflow accepts mRNA and small RNA quantification inputs, metadata files, gene-target tables, and optional class-parameter files. Independent differential expression analyses are performed using DESeq2, generating normalized counts, fold changes, adjusted p-values, and quality-control visualizations. Differential expression results are then integrated through validated small RNA–mRNA target pairs to generate integrative results tables, cosmic plots, slope plots, and binomial sign-test

statistics. Reproducible outputs, including interactive HTML visualizations and YAML parameter snapshots, are archived automatically for analysis and reporting.

As illustrated in Figure 2, mRNA and small RNA datasets are processed independently using DESeq2 to generate normalized counts, log2 fold changes, adjusted p-values, and quality-control metrics. These outputs are subsequently linked through a user-defined gene-target table to generate paired small RNA–mRNA datasets. The integrative module then generates visualizations and statistical summaries that facilitate interpretation of potential regulatory relationships. All workflow steps are governed through YAML configuration files, ensuring reproducibility and transparent documentation of analysis parameters.

Software Design and Implementation

RNA-integrate was developed in R as a modular workflow composed of independent analysis modules that communicate through standardized tabular outputs. The workflow was designed to separate differential expression analysis from downstream integrative analyses, allowing individual components to be executed independently while maintaining compatibility across outputs. Major workflow functions include input validation, data preprocessing, differential expression analysis, target-pair integration, visualization generation, statistical evaluation, and automated report generation. This modular architecture facilitates maintenance, extensibility, and adaptation to future analytical approaches.

Input Data and Experimental Design

RNA-integrate supports multiple RNA-seq quantification formats commonly used in transcriptomic workflows. Accepted input formats include count matrices, HTSeq count files (Anders et al., 2015), RSEM quantifications (Li & Dewey, 2011), Salmon quantifications (Patro et al., 2017), Kallisto quantifications (Bray et al., 2016), and tinyRNA output files (Tate et al., 2023).

Required inputs include:

- Expression count data for mRNA and/or small RNA datasets
- Metadata files containing sample identifiers and experimental conditions
- Gene-target tables defining predicted small RNA–mRNA relationships

Optional inputs include gene annotation tables and class-parameter files used for annotation and visualization customization.

Metadata files define experimental conditions and identify control groups when applicable. These variables are used to construct the design matrix for differential expression analysis and determine comparison groups automatically.

Gene-target tables contain small RNA identifiers and corresponding mRNA target identifiers used to connect differential expression results across datasets. Optional columns containing gene names, target names, and classification information can also be included for downstream annotation and visualization.

Preprocessing and Data Preparation

Prior to analysis, all inputs undergo validation to ensure compatibility with workflow requirements. Input files are checked for existence, required metadata fields, and proper formatting. Gene identifiers are standardized, duplicate identifiers are resolved automatically through suffix assignment, and sample ordering is synchronized between metadata and expression matrices.

For datasets generated using the tinyRNA workflow (Tate et al., 2023), annotation fields including feature identifiers, feature names, and classification categories are automatically extracted and standardized. Identifier prefixes and formatting inconsistencies are removed to improve compatibility with downstream integration and gene-target matching.

Only target pairs supported by observed expression data are retained for downstream analysis. Small RNAs absent from the small RNA dataset, and mRNA targets absent from the mRNA dataset are excluded automatically during preprocessing.

Differential Expression Analysis

Differential expression analysis was performed using DESeq2 (Love et al., 2014), which models count data using a negative binomial distribution. For all supported input formats, counts were normalized to account for differences in sequencing depth, and dispersion parameters were estimated to characterize biological variability.

Size factors were estimated using the median-of-ratios method implemented in DESeq2. Gene-specific dispersion estimates were stabilized using empirical Bayes shrinkage. Differential expression testing was performed using Wald statistics, and p-values were adjusted using the Benjamini–Hochberg procedure to control the false discovery rate.

Genes were classified as significantly differentially expressed when they met both of the following criteria:

- Adjusted p-value < 0.05
- Absolute log2 fold change > log2(1.3)

These thresholds were user-configurable through workflow parameter files.

Quality Control and Diagnostic Visualizations

To evaluate data quality and differential expression performance, RNA-integrate generates multiple quality-control visualizations.

Principal Component Analysis (PCA)

Principal component analysis was performed using variance-stabilized transformed counts generated by DESeq2. PCA plots were used to assess clustering of biological replicates and separation of experimental conditions.

Intra-Condition Scatterplots

Pairwise scatterplots comparing log2-transformed normalized counts between biological replicates were generated to evaluate within-condition reproducibility.

Mean Expression Plots

Mean expression plots display average normalized counts across conditions and allow visualization of global expression patterns.

MA Plots

MA plots display the relationship between average expression and log2 fold change and highlight significantly differentially expressed genes. These plots are generated using DESeq2 (Love et al., 2014).

Heatmaps

Interactive heatmaps were generated to visualize expression patterns across samples and conditions. Heatmaps could be produced for complete datasets or restricted to selected gene classes.

Integration of Small RNA and mRNA Datasets

Following independent differential expression analyses, small RNA and mRNA datasets were integrated using the user-supplied gene-target table.

Target Pair Filtering

Only target pairs meeting both of the following criteria were retained:

- Small RNA detected within the small RNA dataset
- Target mRNA detected within the mRNA dataset

This filtering step ensured that all analyzed relationships were supported by observed expression measurements.

Combined Dataset Construction

For each validated target pair, RNA-integrate compiles:

- Small RNA log2 fold change
- mRNA log2 fold change
- Small RNA adjusted p-value
- mRNA adjusted p-value
- Normalized expression values
- Gene and target annotations when available

The resulting paired dataset served as the foundation for integrative analyses.

Visualization of Integrative Results

Cosmic Plots

Cosmic plots were developed to visualize regulatory relationships between small RNAs and mRNA targets. Each point represents a predicted interaction, with small RNA log2 fold change displayed on the x-axis and mRNA log2 fold change displayed on the y-axis.

Interactions were classified into four significance categories:

1. Significant small RNA and significant mRNA
2. Significant small RNA and non-significant mRNA
3. Non-significant small RNA and significant mRNA
4. Non-significant small RNA and non-significant mRNA

Point color and transparency were used to distinguish significance classes, while point size represented mean mRNA expression. Interactive implementations additionally allowed users to selectively display significance categories and inspect individual interactions through hover annotations.

Slope Plots

Slope plots were generated to visualize paired expression changes between small RNAs and their target mRNAs. Each line connects the fold change of a small RNA with the fold change of its associated target mRNA. Negative slopes indicate inverse expression patterns consistent with canonical repression, whereas positive slopes indicate direct relationships.

Statistical Analysis of Regulatory Relationships

Binomial Sign Test

To evaluate enrichment of inverse regulatory relationships, target pairs were classified as either inverse or direct according to the direction of miRNA and mRNA fold changes. Inverse relationships were defined as miRNA and mRNA fold changes having opposite signs, while direct relationships were defined as fold changes sharing the same sign. Only pairs in which both small RNAs and mRNAs satisfied significance thresholds were included in the analysis.

A one-sided exact binomial sign test was performed under the null hypothesis that inverse and direct relationships occur with equal probability ($p = 0.5$). The observed proportion of inverse relationships was compared against this expectation.

Reproducibility and Output Generation

For each workflow execution, RNA-integrate automatically generates:

- Differential expression results tables (CSV)
- Integrative results tables (CSV)

- Publication-quality figures (PDF and PNG)
- Interactive visualizations (HTML)
- YAML parameter snapshots

Outputs are organized into date-stamped directories to facilitate reproducibility, preserve analysis provenance, and simplify comparison across experiments.

RESULTS

Overview of Differential Expression Analysis

As a proof-of-principle, we applied RNA-integrate to the analysis of small RNA and mRNA sequencing datasets from a partial loss-of-function *pash-1* mutant and wild-type *C. elegans*. The *pash-1* mutant is relatively healthy but has an ~50% reduction in miRNA levels, although the impact on mRNA expression has not been explored (Knittel et al., 2025). We first evaluated the overall data structure to assess the quality of the datasets using principal component analysis (PCA) on both small RNA and mRNA data. For the small RNA dataset, PCA revealed clear separation between *pash-1* and wild-type samples along the first principal component, which accounted for the majority of the variance (Figure 3). Replicates within each condition mostly clustered together, indicating strong within-group consistency and minimal technical variability. This separation suggests that the *pash-1* mutation produces a measurable and consistent effect on small RNA expression profiles.

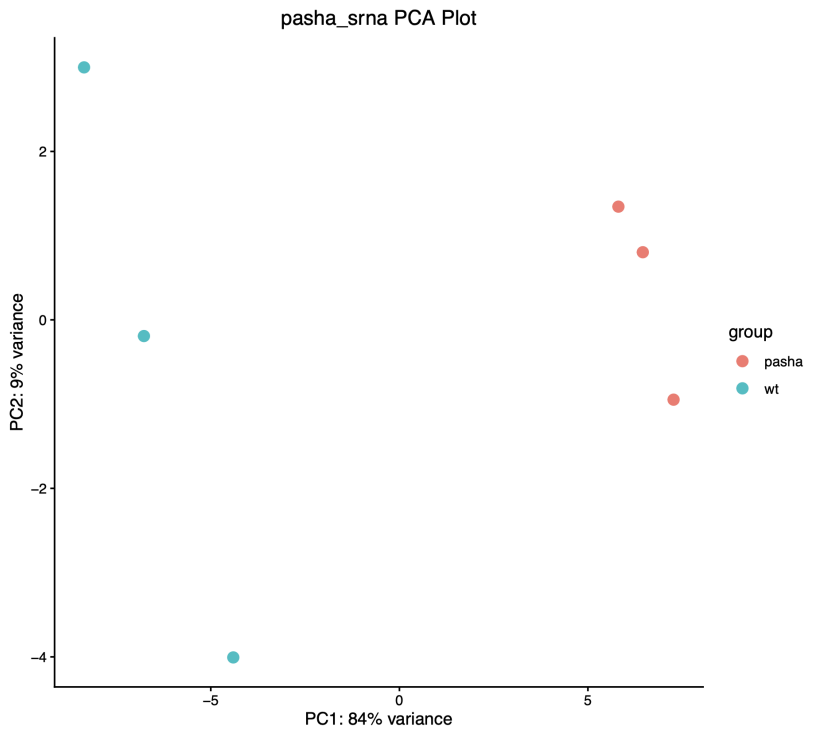

Figure 3. PCA plot of small RNA expression data in *pash-1* versus wild-type conditions. Each point represents a sample, colored by condition. Principal components were calculated from normalized expression values, with separation between groups indicating differences in small RNA expression profiles.

A similar pattern was seen in the mRNA dataset, where PCA showed clear clustering of samples by condition (Figure 4). The first principal component explained a large portion of the variance, supporting strong transcriptional differences between *pash-1* and wild-type samples. Replicates also clustered closely together, indicating high reproducibility, consistent with what was observed in the small RNA data.

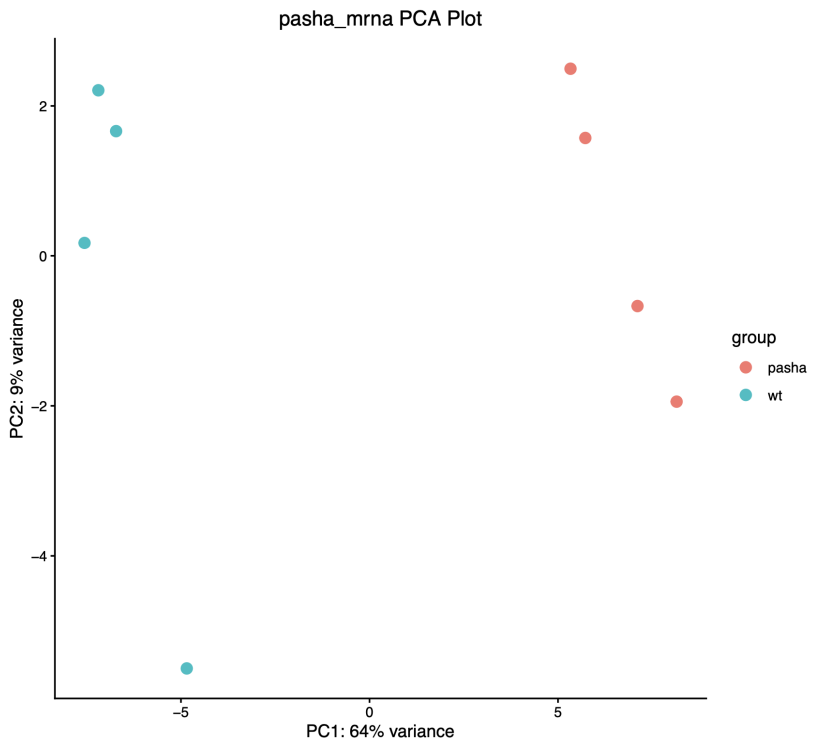

Figure 4. PCA plot of mRNA expression data in *pash-1* versus wild-type conditions. Each point represents a sample, colored by condition. Principal component analysis was performed on normalized mRNA expression values, with separation along the first principal component indicating transcriptional differences between conditions.

In the small RNA dataset, most features were centered around a log2 fold change of zero, suggesting little overall change between conditions (Figure 5). However, a subset showed clear positive or negative fold changes, indicating differential expression in the *pash-1* mutant. Variability in fold-change estimates was higher at lower expression levels, which is expected due to increased noise and reduced statistical power for low-count features.

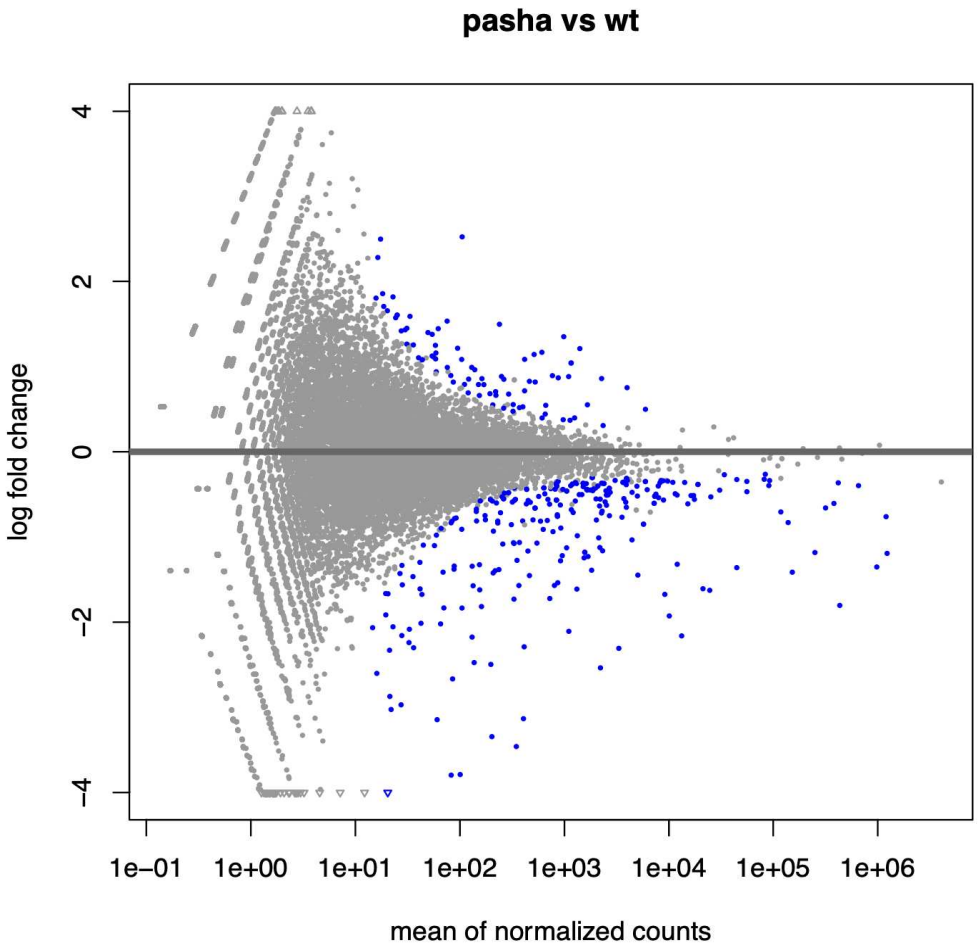

Figure 5. MA plot of small RNA differential expression in *pash-1* versus wild-type conditions. The x-axis shows the mean of normalized counts on a log scale, and the y-axis shows log2 fold change. Points represent small RNA features, with significant features highlighted.

The mRNA MA plot showed a similar pattern, with most genes near zero and a subset showing significant changes (Figure 6). Compared to the small RNA data, significant genes were more broadly distributed across expression levels, indicating more widespread transcriptional changes. Variability was highest at lower expression levels.

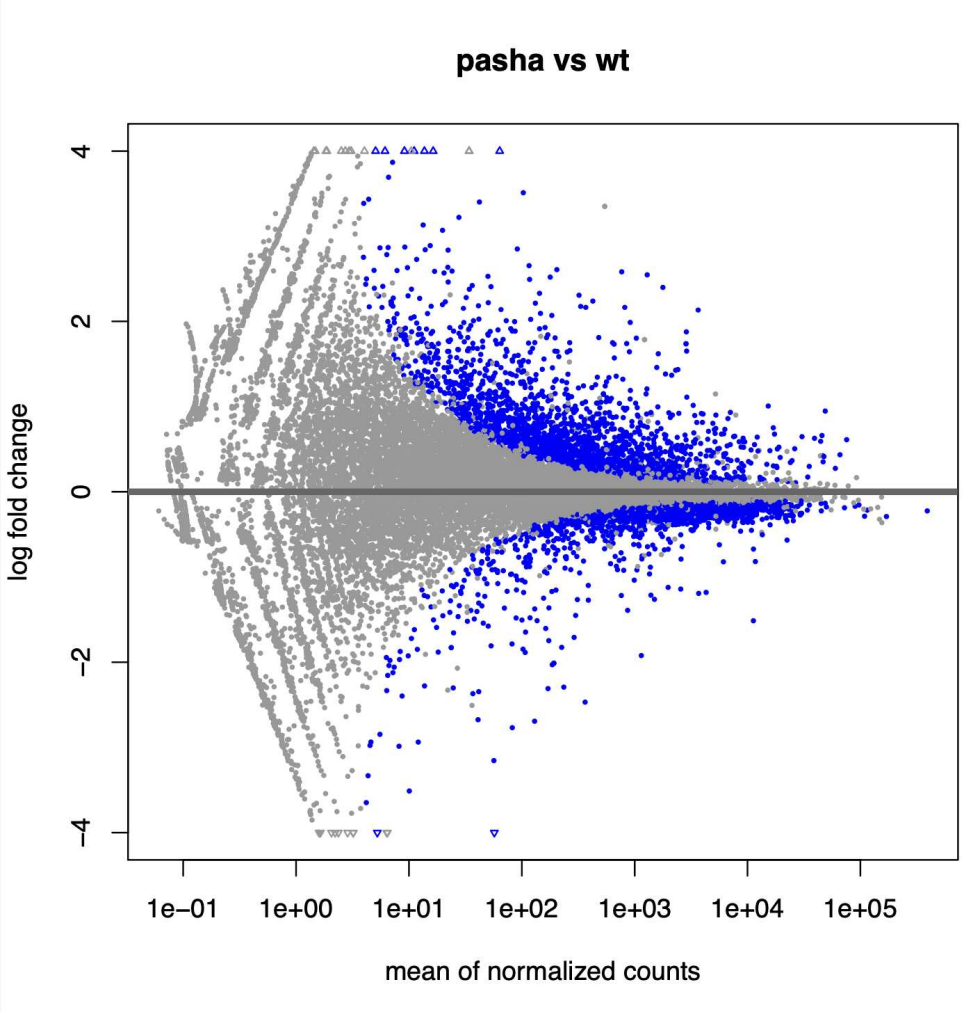

Figure 6. MA plot of mRNA differential expression in *pash-1* versus wild-type conditions. The x-axis shows the mean of normalized counts on a log scale, and the y-axis shows log2 fold change. Points represent mRNA features, with significant genes highlighted.

Overall, both the small RNA and mRNA datasets showed clear differences between conditions, with distinct clustering and subsets of significantly differentially expressed features. Replicates grouped closely together, and the separation between conditions suggests the data are consistent. These results provide a strong foundation for the integrative analysis of miRNA-mRNA regulatory relationships explored in the following section.

Integrative Analysis of Small RNA and mRNA Relationships

To evaluate regulatory relationships between miRNAs and their predicted mRNA targets, integrative analysis is performed using paired fold-change values derived from DESeq2 results within the RNA-integrate framework. These relationships are visualized using a cosmic plot, which displays miRNA and mRNA log2 fold changes for each predicted interaction.

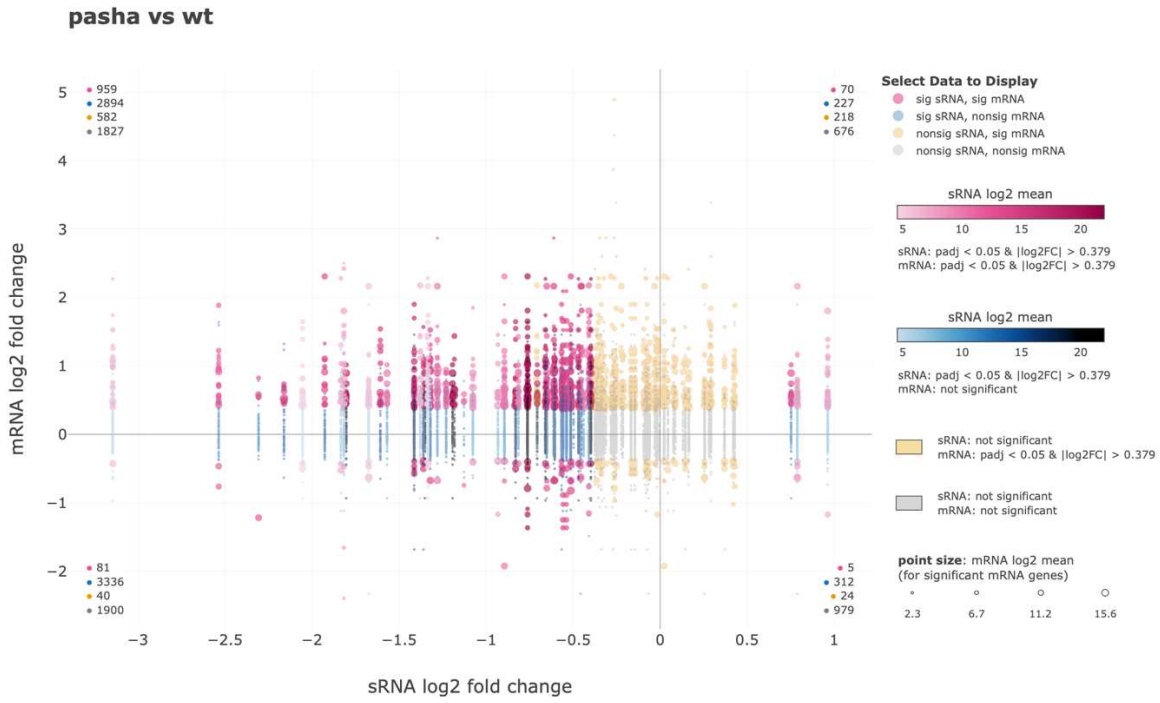

Figure 7. Cosmic plot of predicted miRNA–mRNA interactions in *pash-1* versus wild-type animals. Each point represents a validated miRNA–mRNA target pair. The x-axis displays miRNA log2 fold change, and the y-axis displays mRNA log2 fold change. Interactions are grouped according to statistical significance of the small RNA and mRNA components. Point color intensity reflects miRNA mean expression, while point size corresponds to mRNA log2 mean expression for significant mRNA targets. Quadrant summaries provide counts of interactions within each fold-change direction category, allowing rapid identification of inverse and direct regulatory relationships.

The cosmic plot provided a comprehensive overview of miRNA–mRNA regulatory relationships in our *pash-1* mutant datasets (Figure 7). Interactions were distributed across all four significance categories; however, significant miRNA–mRNA pairs were concentrated within quadrants exhibiting opposite fold-change directions. This pattern is consistent with the

expected repressive activity of miRNAs on their target transcripts. The enhanced visualization framework implemented in these cosmic plots enabled interactions to be simultaneously evaluated according to fold change, statistical significance, and expression abundance. Point color intensity reflects mean miRNA expression, whereas point size represents mean mRNA expression among significant targets. Interactive filtering functionality further allows significance classes to be examined independently if the user chooses. Together, these features facilitate the analysis of miRNA-target pairs on a genome-wide scale.

Although interactions were observed in all quadrants, the enrichment of significant inverse relationships suggested that disruption of *pash-1* preferentially affected miRNA-mediated regulatory pathways rather than producing random transcriptional changes throughout the dataset.

Slope Plot Analysis

To further examine the directional relationship between miRNAs and their predicted mRNA targets, RNA-integrate generates slope plots that directly connect fold-change values for each small RNA-mRNA pair.

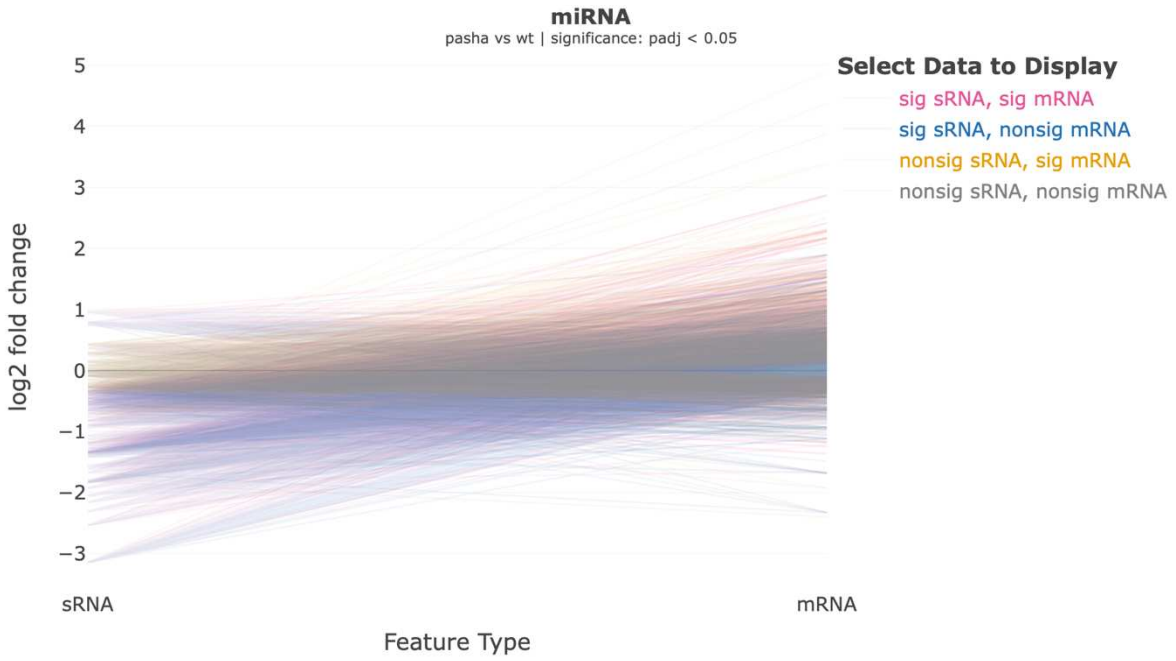

Figure 8. Slope plot of miRNA–mRNA expression relationships in *pash-1* versus wild-type animals. Each line represents a predicted miRNA–mRNA target pair and connects the log2 fold change of the miRNA to the log2 fold change of its associated mRNA target. Lines are colored according to significance classification. Negative slopes indicate inverse relationships consistent with canonical miRNA-mediated repression, whereas positive slopes indicate direct relationships. Interactive filtering allows individual significance classes to be examined independently.

Slope plots provide a complementary view of small RNA–mRNA relationships by directly connecting fold-change values for individual target pairs. Unlike the cosmic plot, which summarizes pairwise interactions within a two-dimensional fold-change space, the slope plot emphasizes the directional relationship between each miRNA and its associated mRNA target.

Within the *pash-1* datasets, a large proportion of significant interactions exhibited negative slopes, indicating opposite directions of expression change between miRNAs and mRNAs (Figure 8). These patterns are consistent with the established role of miRNAs as negative regulators of gene expression. Positive slopes were also observed, suggesting either

indirect regulatory effects, incomplete target predictions, or biological processes not driven by repression mechanisms.

The significance-based classification implemented within RNA-integrate allows different categories of interactions to be evaluated independently, improving interpretability of complex datasets. Visualization of the *pash-1* datasets using slope plots demonstrated that while individual interactions varied substantially in magnitude, inverse regulatory relationships were consistently enriched throughout the integrated dataset (Figure 8).

Statistical Evaluation of Regulatory Patterns

Sign Test for Directional Enrichment

To quantify the enrichment of inverse regulatory relationships, RNA-integrate classifies small RNA–mRNA target pairs according to the direction of their fold changes. Pairs exhibiting opposite fold-change directions are classified as inverse relationships, whereas pairs exhibiting the same fold-change direction are classified as direct relationships. To determine whether the proportion of direct or inverse relationships exceeds random expectation, a one-sided binomial sign test is performed under the null hypothesis that inverse and direct relationships occur with equal probability ($p = 0.5$).

Among 1,115 significant miRNA–mRNA pairs identified in the *pash-1* datasets, 964 exhibited inverse relationships and 151 exhibited direct relationships, corresponding to an observed inverse proportion of 0.865. The resulting p-value ($p = 1.01 \times 10^{-145}$) indicates that his represents a highly significant enrichment of inverse relationships. These findings are consistent with the expected role of miRNAs as negative regulators of gene expression and provide quantitative support for the regulatory patterns observed in both the cosmic and slope plot visualizations.

Software Output and Workflow Performance

RNA-integrate successfully generates a comprehensive collection of reproducible outputs for both differential expression and integrative analyses. For each workflow execution, with a single command the pipeline automatically produced differential expression results tables, quality-control visualizations, integrative results tables, statistical summaries, publication-quality figures, interactive HTML visualizations, and parameter records.

The differential expression component generates normalized count tables, principal component analysis plots, intra-condition scatterplots, mean expression plots, MA plots, and heatmaps for both small RNA and mRNA datasets. These outputs provided validation of sample quality, normalization performance, and differential expression results. The integrative component generates paired small RNA–mRNA datasets together with cosmic plots, slope plots, and statistical summaries describing regulatory relationships.

A major feature of RNA-integrate is its emphasis on reproducibility. For each analysis run, all outputs are organized automatically into date-stamped directories containing both generated results and YAML parameter snapshots. These parameter files preserved complete records of workflow settings, significance thresholds, visualization options, and input file locations, allowing portability and reproducibility.

The workflow additionally generates interactive HTML visualizations that enabled users to explore complex datasets through significance-based filtering and hover-enabled feature inspection. These interactive outputs complement the static PDF figures by providing greater flexibility for exploratory analysis while maintaining publication-ready visualizations for reporting purposes.

RNA-integrate functions not only as an analytical framework for investigating miRNA–mRNA regulatory relationships but also as a reproducible software platform designed for flexible application across diverse RNA-seq datasets and experimental systems.

Summary of Results

We developed RNA-integrate to identify and visualize biologically meaningful regulatory relationships while providing a reproducible framework for integrated analysis of small RNA and mRNA expression data. In a proof-of-principle analysis using *pash-1* mutant *C. elegans* datasets, we identified widespread potential miRNA–mRNA regulatory interactions, paving the way for future studies demonstrating direct regulatory relationships with miRNAs and their predicted targets.

DISCUSSION

The primary goal of this project was to develop a reproducible framework for integrated analysis of miRNA and mRNA datasets. Rather than replacing established differential expression methodologies, RNA-integrate was designed to extend the DESeq2 framework by incorporating downstream integration, visualization, and statistical evaluation of small RNA–mRNA relationships (Love et al., 2014). Many existing differential expression workflows frequently analyze these data types independently, requiring substantial downstream processing to evaluate regulatory relationships. RNA-integrate addresses this limitation by combining differential expression analysis, target-pair integration, visualization, and statistical evaluation within a unified workflow. Compared with existing integrative frameworks such as mirTarRnaSeq (Movassagh et al., 2022) and TimiRGeN (Patel et al., 2021), RNA-integrate emphasizes

workflow reproducibility through external YAML parameterization, support for multiple RNA-seq quantification formats, automated validation procedures, and generation of publication-ready outputs within a single framework. These features facilitate adoption across diverse experimental designs while reducing the amount of manual preprocessing required for integrative analyses.

Several design features contribute to the flexibility and reproducibility of the software. Increasing emphasis has been placed on reproducible, transparent, and reusable computational workflows in modern bioinformatics, consistent with the FAIR principles for scientific data and software stewardship (Wilkinson et al., 2016). The workflow is fully parameterized through YAML configuration files, allowing analyses to be reproduced without modification of source code. Support for multiple quantification formats, including counts matrices, HTSeq, Salmon, Kallisto, RSEM, and tinyRNA outputs, increases applicability across diverse experimental systems. Automated validation of input files and target-pair relationships further improves reliability by reducing opportunities for user error during preprocessing.

The visualization framework implemented in RNA-integrate represents an additional contribution of this work. Cosmic plots provide simultaneous assessment of fold-change direction, statistical significance, and expression abundance, while slope plots allow direct visualization of relationships between individual miRNAs and their associated mRNA targets. Interactive HTML implementations further support exploratory analysis by allowing significance classes to be selectively displayed and individual interactions to be examined in detail. Interactive visualization has become an increasingly valuable component of biological data analysis because it enables users to explore complex datasets beyond the limitations of static figures (Sievert, 2020). Together, these visualizations provide an intuitive approach for interpreting complex regulatory datasets.

Despite these strengths, several limitations should be acknowledged. First, the workflow relies on user-supplied target predictions, which may contain false-positive interactions. Consequently, not all predicted miRNA–mRNA pairs necessarily represent true biological relationships. Incorporation of experimentally validated interaction databases or additional target-confidence metrics could improve interpretation and reduce uncertainty associated with target prediction.

Second, the current implementation focuses primarily on pairwise regulatory relationships between individual miRNAs and mRNA targets. Gene regulation frequently occurs within larger interconnected networks involving multiple regulatory molecules, feedback mechanisms, and temporal effects. Future development could incorporate network-based approaches, pathway-level analyses, or time-series datasets to provide a more comprehensive view of regulatory dynamics.

Additional opportunities for future development include incorporation of alternative statistical frameworks, integration of experimentally validated target resources, expanded support for additional RNA classes, and development of graphical user interfaces that further improve accessibility for users without programming experience.

Our integrative analyses of wild-type and *pash-1* mutants performed using RNA-integrate provided a proof-of-principle, while also yielding a valuable resource for exploring miRNA-mRNA interactions in *C. elegans*. The results revealed a strong enrichment of inverse relationships between miRNAs and their predicted mRNA targets. Among significant target pairs, 86.5% exhibited opposite fold-change directions, and the resulting binomial sign test demonstrated that this pattern was highly unlikely to occur by chance. These findings are consistent with the established role of miRNAs as post-transcriptional repressors of gene

expression (Bartel, 2004; Bartel, 2009) and support the expectation that disruption of miRNA biogenesis leads to repression of target mRNAs.

Although inverse relationships were strongly enriched, substantial variability remained among individual target pairs. Both the cosmic and slope plot visualizations revealed interactions spanning multiple significance categories and fold-change directions. This variability likely reflects the complexity of post-transcriptional regulation, where individual targets differ in binding affinity, regulatory context, transcript stability, and contributions from additional regulatory pathways. As a result, not all predicted interactions are expected to produce measurable expression changes under a given experimental condition.

Overall, this study demonstrates that RNA-integrate provides an effective and reproducible framework for integrating small RNA and mRNA differential expression analyses. Application of the workflow to a *pash-1* mutant dataset recovered regulatory patterns consistent with established models of miRNA-mediated repression while simultaneously demonstrating the utility of the software for visualization and interpretation of complex RNA regulatory relationships. By combining reproducible workflow design, flexible data integration, interactive visualization, and statistical evaluation within a single platform, RNA-integrate provides a useful resource for future investigations of RNA regulatory networks across diverse biological systems.

CONCLUSION

This study presents RNA-Integrate, a reproducible workflow for integrated analysis of miRNA and mRNA expression data. By combining differential expression analysis, target-pair integration, visualization, and statistical evaluation within a single framework, RNA-Integrate

provides an accessible and flexible platform for investigating miRNA-mediated regulatory relationships.

Application of the workflow to a *pash-1* mutant dataset in *Caenorhabditis elegans* demonstrated significant enrichment of inverse miRNA–mRNA relationships, consistent with established models of miRNA-mediated repression. These findings provide proof-of-principle that RNA-Integrate can recover biologically meaningful regulatory patterns while facilitating interpretation through integrative visualizations and statistical analysis.

As transcriptomic datasets continue to increase in size and complexity, reproducible computational workflows will remain essential for extracting meaningful biological insight. RNA-Integrate provides a foundation for future integrative analyses across diverse organisms and experimental systems and offers a practical framework that can be expanded to incorporate additional RNA classes, validated regulatory interactions, and network-level analyses.

DATA AVAILABILITY

The small RNA sequencing data used in this study are publicly available through the NCBI Gene Expression Omnibus (GEO) under accession number GSE263914. Additional processed outputs generated during analysis, including differential expression results and integrative analysis tables, are available through the RNA-integrate repository and associated supplementary materials where applicable.

CODE AVAILABILITY

RNA-integrate is available as open-source software through the Montgomery Laboratory GitHub repository: https://github.com/MontgomeryLab/RNA-integrate

The repository contains the RNA-integrate workflow described in this study, including the DESeq2 differential expression module, integrative analysis module, example datasets, YAML parameter templates, demonstration workflows, and user documentation.

The tinyRNA workflow used for small RNA quantification is available through the Montgomery Laboratory GitHub repository: https://github.com/MontgomeryLab/tinyRNA

REFERENCES

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq: A Python framework to work with high-throughput sequencing data. *Bioinformatics, 31*(2), 166–169. https://doi.org/10.1093/bioinformatics/btu638

Bartel, D. P. (2004). MiRNAs: Genomics, biogenesis, mechanism, and function. *Cell, 116*(2), 281–297.

Bartel, D. P. (2009). MiRNAs: Target recognition and regulatory functions. *Cell, 136*(2), 215–233.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology, 34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology, 17,* 13. https://doi.org/10.1186/s13059-016-0881-8

Denli, A. M., Tops, B. B. J., Plasterk, R. H. A., Ketting, R. F., & Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature, 432*(7014), 231–235. https://doi.org/10.1038/nature03049

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., & Marks, D. S. (2003). MiRNA targets in Drosophila. *Genome Biology, 5*(1), R1. https://doi.org/10.1186/gb-2003-5-1-r1

Gregory, R. I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., &

Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature, 432*(7014), 235–240. https://doi.org/10.1038/nature03120

Ha, M., & Kim, V. N. (2014). Regulation of miRNA biogenesis. *Nature Reviews Molecular Cell Biology, 15*(8), 509–524. https://doi.org/10.1038/nrm3838

Knittel, T. L., Montgomery, B. E., Reed, K. J., Chong, M. C., Isolehto, I. J., Cafferty, E. R., Smith, M. J., Sprister, R. A., Magelky, C. N., Scherman, H., Ketting, R. F., & Montgomery, T. A. (2025). Interdependence of Pasha and Drosha for localization and function of the Microprocessor in *Caenorhabditis elegans*. *Nature Communications, 16,* 5595. https://doi.org/10.1038/s41467-025-60721-5

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Rådmark, O., Kim, S., & Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature, 425*(6956), 415–419. https://doi.org/10.1038/nature01957

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are miRNA targets. *Cell, 120*(1), 15–20. https://doi.org/10.1016/j.cell.2004.12.035

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics, 12*, 323. https://doi.org/10.1186/1471-2105-12-323

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology, 15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Movassagh, M., Morton, S. U., Hehnly, C., Smith, J., Doan, T. T., Irizarry, R., Broach, J. R., Schiff, S. J., Bailey, J. A., & Paulson, J. N. (2022). mirTarRnaSeq: An R/Bioconductor

statistical package for miRNA–mRNA target identification and interaction analysis. *BMC Genomics, 23*, 439. https://doi.org/10.1186/s12864-022-08558-w

Patel, K., Chandrasegaran, S., Clark, I. M., Proctor, C. J., Young, D. A., & Shanley, D. P. (2021). TimiRGeN: R/Bioconductor package for time series miRNA–mRNA integration and analysis. *Bioinformatics, 37*(20), 3604–3609 https://doi.org/10.1093/bioinformatics/btab377

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods, 14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Perkel, J. M. (2019). Workflow systems turn raw data into scientific knowledge. *Nature,* 573(7772), 149–150. https://doi.org/10.1038/d41586-019-02619-z

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny.* CRC Press. https://doi.org/10.1201/9780429447273

Tate, A. J., Brown, K. C., & Montgomery, T. A. (2023). tiny-count: A counting tool for hierarchical classification and quantification of small RNA-seq reads with single-nucleotide precision. *Bioinformatics Advances, 3*(1), vbad181. https://doi.org/10.1093/bioadv/vbad065

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3,* 160018. https://doi.org/10.1038/sdata.2016.18