miRNA Differential Expression Analysis Pipeline

A comprehensive R-based pipeline for analyzing miRNA differential expression using edgeR, with enhanced visualizations and modular structure.

Overview

This pipeline provides a complete workflow for miRNA differential expression analysis, including:

- Quality control and filtering
- Differential expression testing
- Multiple visualization types (heatmaps, volcano plots, MA plots, BCV plots)
- Statistical summaries and reports
- Modular design for easy customization
- Simplified execution via wrapper script

Requirements

R Version

- R >= 4.0.0

Required R Packages

The pipeline will automatically install missing packages, but you can install them manually:

```
 # Bioconductor packages
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install(c("edgeR", "ComplexHeatmap"))

# CRAN packages
install.packages(c("ggplot2", "dplyr", "pheatmap", "RColorBrewer",
                   "gridExtra", "knitr", "ggrepel", "VennDiagram",
                   "tidyr", "viridis", "circlize"))
```

File Structure

```
project_directory/
    modules/
        setup_packages.r # Package installation and loading
        data_handling.r # Data loading and validation
        normalization_filtering.r # Data filtering and normalization
        diagnostic_module.r # BCV plots and diagnostic functions
        differential_expression.r # Core DE analysis functions
        heatmap_module.r # Heatmap generation functions
        volcano_plots_module.r # Volcano plot functions
        ma_plots_module.r # MA plot functions
        venn_diagram_module.r # Venn diagram functions
        summary_module.r # Summary statistics and reporting
        main_analysis.r # Main analysis workflow
        main_analysis_module.r # Alternative analysis wrapper
    run_analysis.R # Minimal execution script (sources main_analysis.R)
    sample_info.csv # Sample metadata (REQUIRED)
    my_merged_results_reorganized.tsv # Count matrix (REQUIRED)
    mirna_readme.md # This file
    mirna_readme.pdf # PDF version of documentation
```

Input Files

1. Sample Information File (sample_info.csv)

Required columns:

- **Sample:** Sample identifiers (must match count matrix column names)
- **Group:** Experimental group (e.g., "Young_Leaves", "Mature_Roots")

Example format:

```
Sample,Group
Sample1,Young_Leaves
Sample2,Young_Leaves
Sample3,Mature_Leaves
Sample4,Mature_Leaves
Sample5,Young_Roots
Sample6,Mature_Roots
```

2. Count Matrix (my_merged_results_reorganized.tsv)

- Tab-separated file with miRNA names as row names
- Sample names as column headers (matching sample_info.csv)
- Raw count values

Example format:

```
 miRNA_ID    Sample1    Sample2    Sample3    Sample4
miR-001     150        200        180        120
miR-002     300        250        400        350
miR-003     50         80         60         45
```

Execution Methods

Method 1: Simple Execution (Recommended)

```
 # 1. Set your working directory to the project folder
setwd("path/to/your/project")

# 2. Run the minimal execution script which sources the main analysis
source("run_analysis.R")
```

Method 2: Direct Module Execution

```
 # Alternatively, you can source the main analysis module directly
source("modules/main_analysis.R")
```

Method 3: Custom Workflow

```
 # For advanced users: Create your own workflow by combining modules
source("modules/setup_packages.r")
source("modules/data_handling.r")
source("modules/differential_expression.r")
source("modules/volcano_plots_module.r")

# Load and process data
dge_data <- load_and_validate_data("sample_info.csv", "counts.tsv")
de_results <- perform_differential_expression(dge_data)
create_volcano_plots(de_results)
```

Analysis Parameters (Configured in main_analysis.R)

```
 # Statistical thresholds
STAT_THRESHOLD_TYPE <- "pvalue"  # "FDR" or "pvalue"
FDR_THRESHOLD <- 0.05
PVALUE_THRESHOLD <- 0.05
LOGFC_THRESHOLD <- 1

# Filtering parameters
MIN_CPM <- 5      # Minimum CPM for filtering
MIN_SAMPLES <- 3   # Minimum samples with CPM > MIN_CPM

# Visualization
TOP_N_HEATMAP <- 500       # Top N genes for heatmaps
TOP_N_VOLCANO_LABELS <- 10 # Top genes to label in volcano plots
```

Output Files

Diagnostic Plots

- `BCV_plots_comprehensive.png` - Dispersion analysis
- `MA_plots_[threshold].png` - MA plots

Expression Heatmaps

- `global_expression_heatmap_fixed.png` - Global patterns
- `DEM_heatmap_[comparison]_[threshold].png` - DEM-specific

Statistical Plots

- `volcano_[comparison]_[threshold].png` - Volcano plots
- `venn_diagram_DEMs.png` - Venn diagrams (when applicable)

Data Tables

- `DE_results_[comparison].txt` - Detailed results
- `normalized_counts_CPM.txt` - Normalized expression values

Reports

- `session_info.txt` - Environment details
- `comprehensive_results_summary.txt` - Analysis summary

Troubleshooting

Common Issues

File not found errors

- Verify file paths in run_analysis.R
- Check working directory is set correctly

Sample mismatch errors

- Ensure sample names match exactly between files
- Check for hidden characters or spaces

No DEMs found

- Adjust thresholds in main_analysis.R
- Reduce MIN_CPM or MIN_SAMPLES

Memory issues

- Decrease TOP_N_HEATMAP value
- Filter more aggressively before analysis

Version History

- **v4.1:** Simplified execution with wrapper script (June 2025)
- **v4.0:** Complete modular redesign (May 2025)
- **v3.1:** Added comprehensive documentation (March 2025)
- **v3.0:** Enhanced visualizations (January 2025)

Compatible with: R >= 4.0.0, edgeR >= 3.34.0