Toolkit for quick and low-cost differential expression analysis of RNA-seq data

Alexis Norris[1], Mamatha Garige[2], Susmita Ghosh[2], Carole Sourbier[2], Heather Lombardi[1], and Mayumi Miller[3]
[1]Office of New Animal Drug Evaluation (CVM); [2]Office of Biotechnology Products (CDER); [3]Office of Research (CVM)

Abstract

Bulk RNA sequencing (RNA-seq) is a powerful and widely used tool to identify biomarkers. The most common RNA-seq analysis is differential expression (DE) analysis, for which there are well-benchmarked analysis tools available. However, these tools are relatively inaccessible to scientists without bioinformatics expertise. Thus, we sought to create an end-to-end toolkit for DE analysis that could be quickly employed by all FDA scientists, using existing resources and with minimal cost to the user.

Our toolkit makes use of user-friendly, open-source interfaces: Galaxy, precisionFDA, and RStudio. First, precisionFDA or CDRH's Galaxy is used to trim the sequencing reads, align, and generate gene expression counts. Then, RStudio software is used to identify differentially expressed genes and pathways. Finally, RStudio's Shiny interactive dashboards are used to explore and visualize the results.

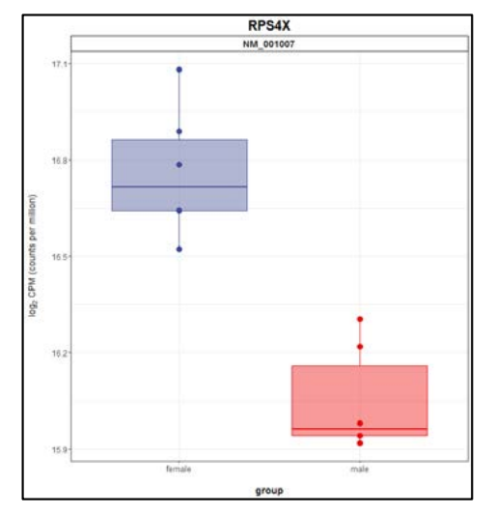
The entire analysis can be performed on a laptop, and typical analyses (6-12 samples) can be completed within one day. CDRH usage is low cost ($5 per CPU day) and Galaxy runs in a web browser without additional cost. RStudio software is free and can be run on PC, Mac, and Linux computers. The Shiny dashboards allow users to generate custom, publication quality figures and tables without costly bioinformatics support.

We have created a toolkit to perform RNA-seq analysis without the need for bioinformatics expertise. The analysis is low cost, quick, reproducible, and available to FDA scientists to support regulatory research and evaluation.

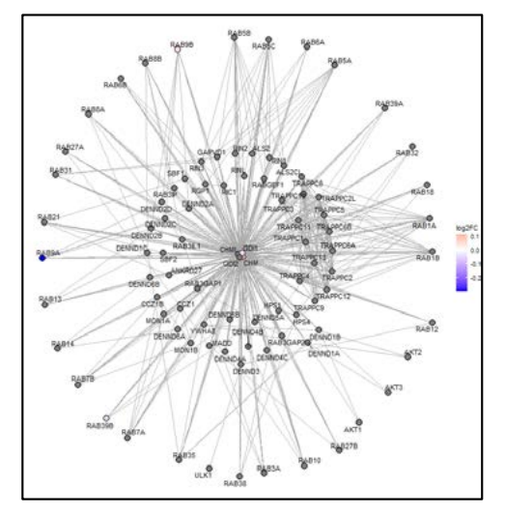
Introduction

Biomarker discovery with RNA-seq
Differential expression analysis of RNA-seq data identifies genes which are dysregulated between groups. Pathways with differentially expressed genes indicate dysregulated pathways. These dysregulated genes and pathways are potential biomarkers for susceptibility, diagnosis, prognosis, treatment response, etc.

Dysregulated genes



Disrupted pathways



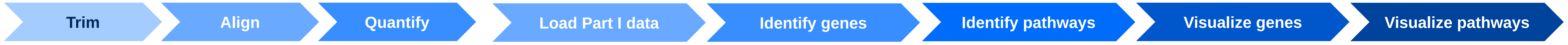
Overview of toolkit analysis
The toolkit has two parts and makes use of user-friendly, open-source interfaces that are publicly available: Galaxy, precisionFDA, and RStudio. Part I is performed using either Galaxy (**Figure 1**) or precisionFDA. Part II is performed using RStudio (**Figure 2**). Code and instructions are available at https://git.fda.gov/alexis.norris/rnaseq_de_toolkit

Materials and Methods

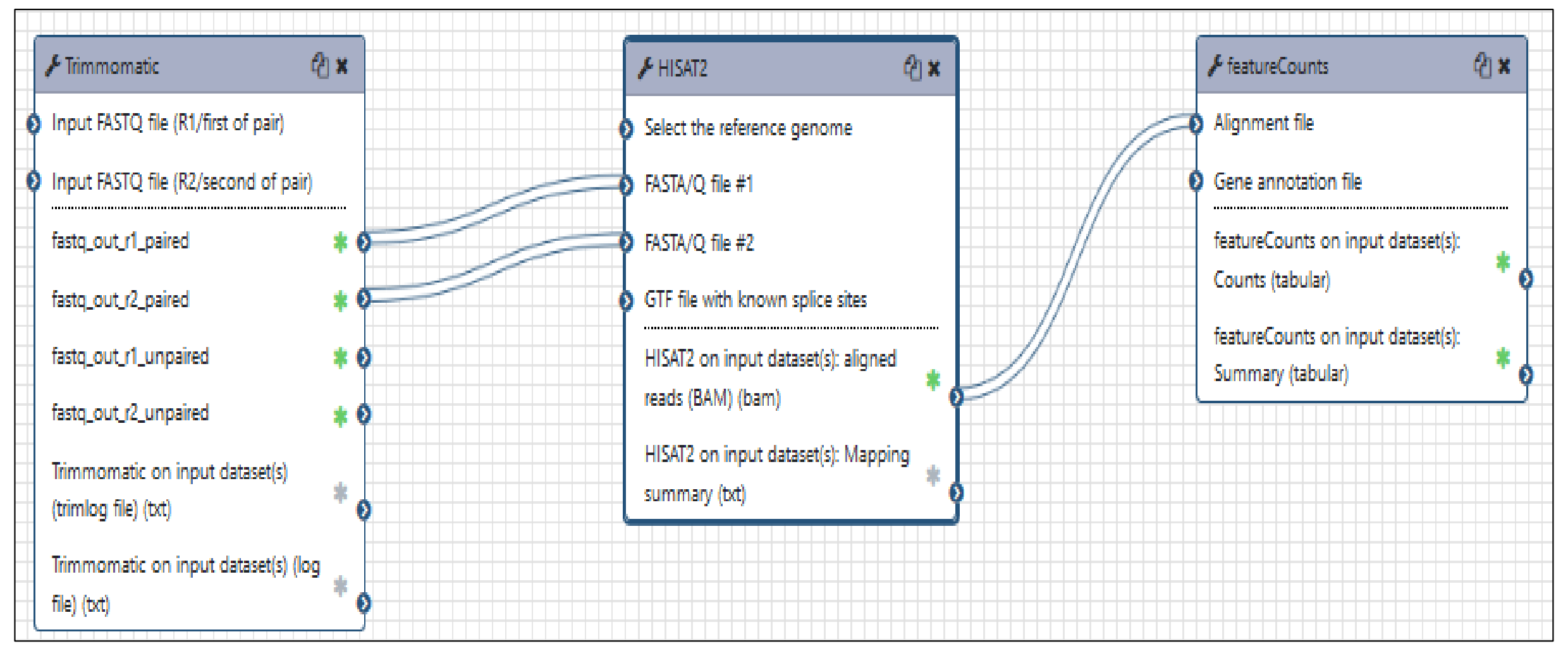
Dataset used to demonstrate the toolkit
We used polyA-captured Illumina paired-end (75bp x 2) reads from Pertea, *et al.* Nature Methods 2016 (https://www.nature.com/articles/nprot.2016.095). The files are subsets of Geuvadis data, for chrX of six female and 6 male samples, available to download at ftp://ftp.ccb.jhu.edu/pub/RNAseq_protocol.

Part I: Quantify gene expression
1. Trim reads (trimmomatic)
2. Align reads to the reference genome (hisat2)
3. Quantify gene expression (featureCounts)

Part II: Identify differentially expressed genes and pathways; visualize results
1. Identify differentially expressed genes (edgeR)
2. Identify differentially expressed pathways (clusterProfiler)
3. Visualize results (RShiny)



Trim → Align → Quantify → Load Part I data → Identify genes → Identify pathways → Visualize genes → Visualize pathways

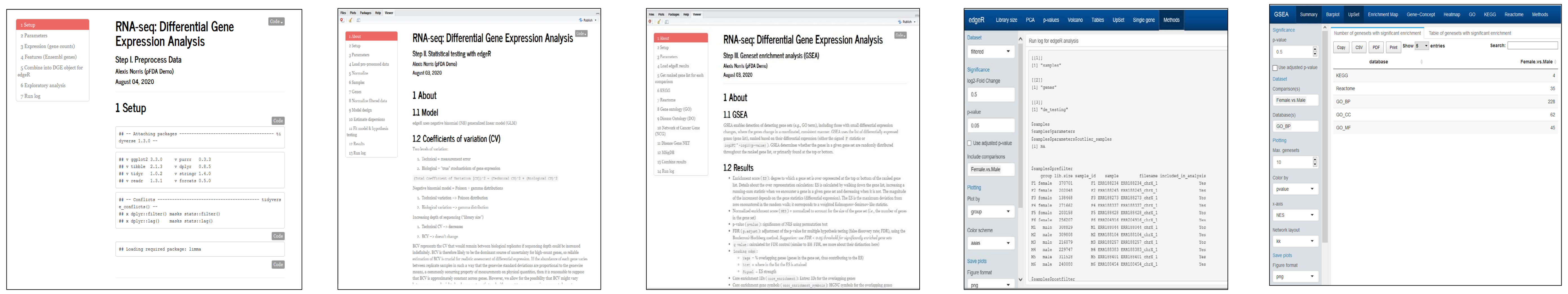


Trimmomatic
- Input FASTQ file (R1/first of pair)
- Input FASTQ file (R2/second of pair)
- fastq_out_r1_paired
- fastq_out_r2_paired
- fastq_out_r1_unpaired
- fastq_out_r2_unpaired
- Trimmomatic on input dataset(s) (trimlog file) (txt)
- Trimmomatic on input dataset(s) (log file) (txt)

HISAT2
- Select the reference genome
- FASTA/Q file #1
- FASTA/Q file #2
- GTF file with known splice sites
- HISAT2 on input dataset(s): aligned reads (BAM) (bam)
- HISAT2 on input dataset(s): Mapping summary (txt)

featureCounts
- Alignment file
- Gene annotation file
- featureCounts on input dataset(s): Counts (tabular)
- featureCounts on input dataset(s): Summary (tabular)

Figure 1. Part I using Galaxy. Public instance available at www.usegalaxy.org; FDA instance available from CDRH HPC group.



1 Setup
2 Parameters
3 Expression (gene counts)
4 Features (Ensembl genes)
5 Combine into DGE object for edgeR
6 Exploratory analysis
7 Run log

RNA-seq: Differential Gene Expression Analysis
Step I. Preprocess Data
Alexis Norris (pFDA Demo)
August 04, 2020

1 Setup

```
## -- Attaching packages --------------------------------------- ti
dyverse 1.3.0 --
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
## -- Conflicts ------------------------------------------ tidyvers
e_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## Loading required package: limma
```

1 About
2 Setup
3 Parameters
4 Load pre-processed data
5 Normalize
6 Samples
7 Genes
8 Normalize filtered data
9 Model design
10 Estimate dispersions
11 Fit model & hypothesis testing
12 Results
13 Run log

RNA-seq: Differential Gene Expression Analysis
Step II. Statistical testing with edgeR
Alexis Norris (pFDA Demo)
August 03, 2020

1 About

1.1 Model

edgeR uses negative binomial (NB) generalized linear model (GLM)

1.2 Coefficients of variation (CV)

Two levels of variation:
1. Technical = measurement error
2. Biological = "true" stochasticism of gene expression

(Total Coefficient of Variation [CV])^2 = (Technical CV)^2 + (Biological CV)^2

Negative binomial model = Poisson + gamma distributions
1. Technical variation -> Poisson distribution
2. Biological variation -> gamma distribution

Increasing depth of sequencing ("library size")
1. Technical CV -> decreases
2. BCV -> doesn't change

BCV represents the CV that would remain between biological replicates if sequencing depth could be increased indefinitely. BCV is therefore likely to be the dominant source of uncertainty for high-count genes, so reliable estimation of BCV is crucial for realistic assessment of differential expression. If the abundance of each gene varies between replicate samples in such a way that the genewise standard deviations are proportional to the genewise means, a commonly occurring property of measurements on physical quantities, then it is reasonable to suppose that BCV is approximately constant across genes. However, we allow for the possibility that BCV might vary

1 About
2 Setup
3 Parameters
4 Load edgeR results
5 Get ranked gene list for each comparison
6 KEGG
7 Reactome
8 Gene ontology (GO)
9 Disease Ontology (DO)
10 Network of Cancer Gene (NCG)
11 Disease Gene NET
12 MSigDB
13 Combine results
14 Run log

RNA-seq: Differential Gene Expression Analysis
Step III. Geneset enrichment analysis (GSEA)
Alexis Norris (pFDA Demo)
August 03, 2020

1 About

1.1 GSEA

GSEA enables detection of detecting gene sets (e.g., GO term), including those with small differential expression changes, where the genes change in a coordinated, consistent manner. GSEA uses the list of differentially expressed genes (gene list), ranked based on their differential expression (either the signed F statistic or log2FC * -log10(p-value)). GSEA determines whether the genes in a given gene set are randomly distributed throughout the ranked gene list, or primarily found at the top or bottom.

1.2 Results

- Enrichment score (ES): degree to which a gene set is over-represented at the top or bottom of the ranked gene list. Details about the over representation calculation: ES is calculated by walking down the gene list, increasing a running-sum statistic when we encounter a gene in a given gene set and decreasing when it is not. The magnitude of the increment depends on the gene statistics (differential expression). The ES is the maximum deviation from zero encountered in the random walk; it corresponds to a weighted Kolmogorov-Smirnov-like statistic.
- Normalized enrichment score (NES) = normalized to account for the size of the gene set (i.e., the number of genes in the gene set)
- p-value (pvalue): significance of NES using permutation test
- FDR (p.adjust): adjustment of the p-value for multiple hypothesis testing (false discovery rate; FDR), using the Bonferroni-Hochberg method. Suggestion: use FDR < 0.05 threshold for significantly enriched gene sets
- q value: calculated for FDR control (similar to BH-FDR, see more about their distinction here)
- Leading edge:
  - Tags = % overlapping genes (genes in the gene set, thus contributing to the ES)
  - List = where in the list the ES is attained
  - Signal = ES strength
- Core enrichment IDs (core_enrichment): Entrez IDs for the overlapping genes
- Core enrichment gene symbols (core_enrichment_symbols): HGNC symbols for the overlapping genes

Figure 2. Part II. To explore the results (both tables and visually) and generate high-quality figures for publication, Shiny Apps can be deployed using RStudio. https://rstudio.com/

Results and Discussion

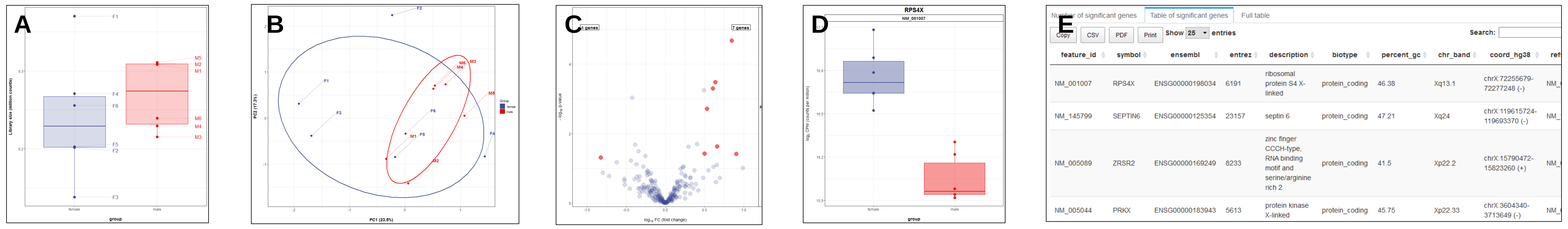


Figure 3. Examples of gene visualizations. **A.** library sizes; **B.** Principal Component Analysis (PCA) clustering; **C.** volcano plot; **D.** boxplot of expression for a gene of interest; and **E.** table of full DE results that can be filtered and exported.

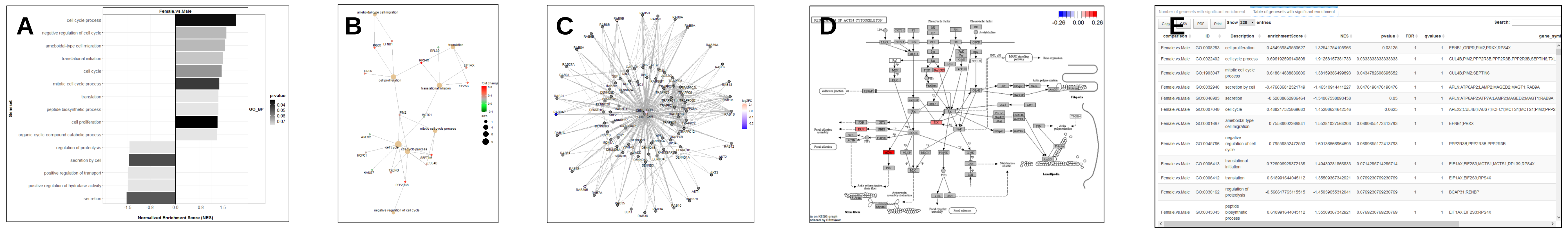


Figure 4. Examples of pathway visualizations. **A.** barplot of the top pathways; **B.** gene overlap (gene ~ concept map); **C.** Reactome pathway's network view; **D.** KEGG pathway; and **E.** table of full pathway results that can be filtered and exported.

Conclusion

RNA-seq DE Toolkit enables regulatory research

Quick
- Typical analyses can be completed within 1 day

Inexpensive
- Galaxy is free; CDRH usage is only $5/CPU day
- precisionFDA is free
- RStudio is free

Reproducible
- Methods are documented
- Methods, data, and Shiny Apps can be easily shared

Flexible
- The Galaxy portion is now also available on precisionFDA

Reduced barrier to RNA-seq analysis
- No coding required for Galaxy, Shiny Apps, and precisionFDA
- Minimal coding for RStudio

Current efforts

Expand user testing
- The toolkit has been tested by beta users on PC laptops
- We're currently testing on Mac and Linux computers

Eliminate the need to install RStudio software on laptop
- Run Rmd and RShiny apps on precisionFDA directly

For more information, contact:
Alexis.Norris@fda.hhs.gov or Mayumi.Miller@fda.hhs.gov