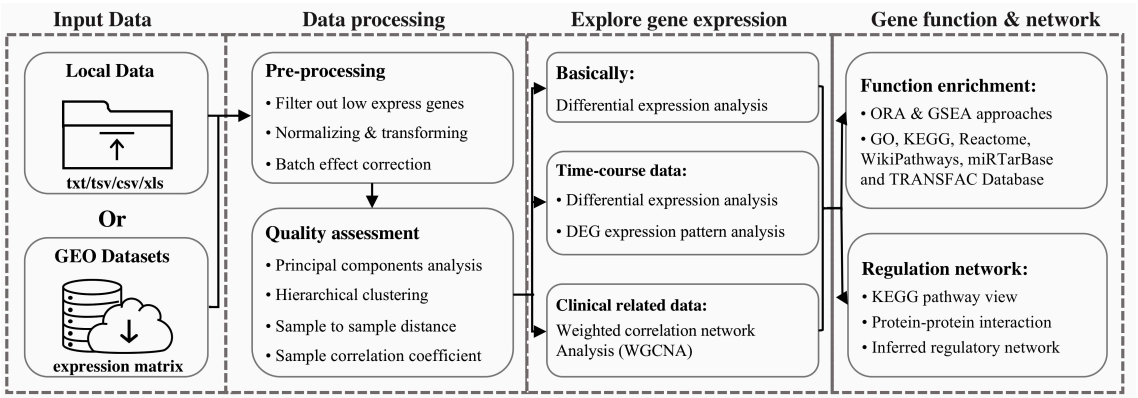
# QRAP: an R package and Shiny app for interactive RNA sequencing data analysis

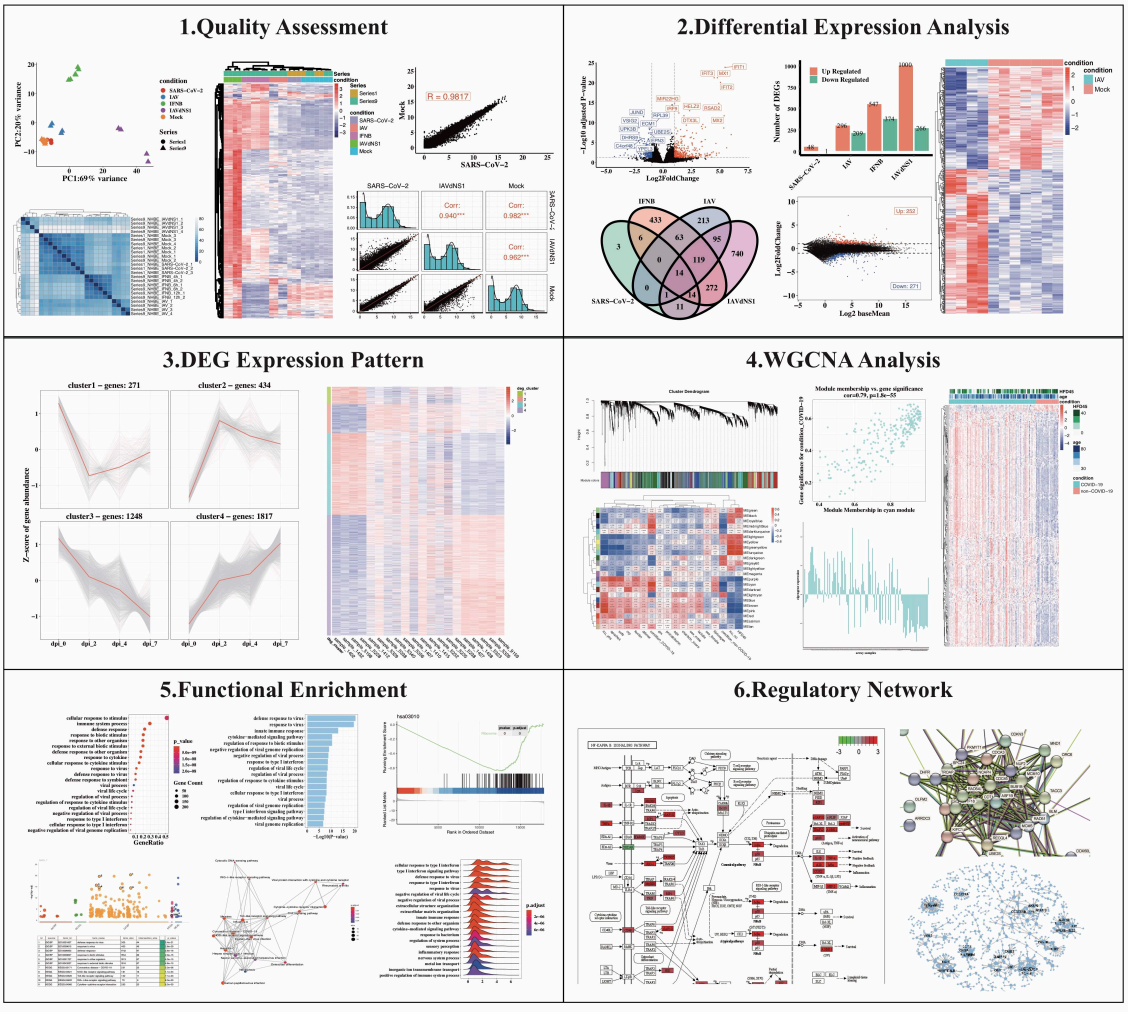
Shixue Gou

2022-12-05

# Contents

# Preface


Input Data
Data processing
Explore gene expression
Gene function & network
Local Data
txt/tsv/csv/xls
Or
GEO Datasets
expression matrix
Pre-processing
• Filter out low express genes
• Normalizing & transforming
• Batch effect correction
Quality assessment
• Principal components analysis
• Hierarchical clustering
• Sample to sample distance
• Sample correlation coefficient
Basically:
Differential expression analysis
Time-course data:
• Differential expression analysis
• DEG expression pattern analysis
Clinical related data:
Weighted correlation network Analysis (WGCNA)
Function enrichment:
• ORA & GSEA approaches
• GO, KEGG, Reactome, WikiPathways, miRTarBase and TRANSFAC Database
Regulation network:
• KEGG pathway view
• Protein-protein interaction
• Inferred regulatory network


1.Quality Assessment
2.Differential Expression Analysis
3.DEG Expression Pattern
4.WGCNA Analysis
5.Functional Enrichment
6.Regulatory Network

# Chapter 1

# Introduction

## 1.1 Introduction

Motivation: RNA-sequencing (RNA-seq) has become the most commonly used tool in life science research for exploring whole transcript profiles. The advancement of next-generation sequencing has promoted a large amount of RNA-seq data. However, the popularity of bioinformatics lags far behind the generation of sequencing data, resulting in the inability of most researchers to analyze RNA-seq data. Although a large number of tools are currently available for RNA-seq analysis, data uploading, analysis, and visualization through an interactive interface are more acceptable to researchers than command-line codes.

Results: We designed an interactive RNA-seq analysis toolkit based on the R Shiny package, named QRAP (Quick RNA-seq Analysis Platform), which can easily accomplish RNA-seq data analysis and visualization through a user-friendly interface on the web page. As a comprehensive RNA-seq analysis tool, QRAP can support the analysis of publicly available and user-generated data, which include regular RNA-seq data, time-course RNA-seq data, and clinically relevant RNA-seq data, and provide functional annotation for approximately 500 species.

Availability and implementation: As an open source R package, QRAP can be freely accessed at https://github.com/gsx-ucas/QRAP.

## 1.2 Installation

### 1.2.1 R package

To install QRAP, R version 4.0 or greater is required. We also recommend installing R Studio.

#### 1.2.1.1 Install the release version of QRAP

```
# Enter commands in R (or R studio, if installed)
# Install the remotes package install.packages('devtools')
devtools::install_github("gsx-ucas/QRAP")
```

#### 1.2.1.2 Install the development version of QRAP

```
# Enter commands in R (or R studio, if installed)
# Install the remotes package install.packages('devtools')
devtools::install_github("gsx-ucas/QRAP", ref = "dev")
```

### 1.2.2 Docker image

We provide docker images for QRAP via dockerhub. To pull the latest image using the command line:

```
# Enter commands in shell
docker pull goushixue/qrap:latest
```

## 1.3 Get started

### 1.3.1 Launch the QRAP application in R or Rstudio

```
library(QRAP) # loading the QRAP library to R environment
startQRAP() # launch the QRAP application to web browser
```

Then you would get the link to activate your browser:

```
> library(QRAP)
> startQRAP()

Listening on http://127.0.0.1:4986
```

Use the link http://127.0.0.1:4986 to access the interactive analysis interface. Note that the port (4986) should change to yours.

### 1.3.2 Launch the QRAP application by docker image in shell

```
docker run  -p 3838:3838 goushixue/qrap # use the 3838 port
```

Then you would get the output like this:

```
mac@mac-2:~$ docker run  -p 3838:3838 goushixue/qrap

R version 4.2.2 (2022-10-31) -- "Innocent and Trusting"
Copyright (C) 2022 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> QRAP::startQRAP(port = 3838, host = '0.0.0.0')
```

Use the link http://localhost:3838/ to access the interactive analysis interface.

### 1.3.3 Access the interactive analysis interface

Just start your analysis by clicking, clicking, clicking...


## Abstract

We designed an interactive RNA-seq analysis toolkit based on R Shiny package, named QRAP (Quick RNA-seq Analysis Platform), which can easily accomplish RNA-seq data analysis and visualization through an intuitive graphical interface on the web page. As a comprehensive RNA-seq analysis tool, QRAP can support to analyze publicly available and user-generated data, which include regular RNA-seq data, time-course RNA-seq, data and clinically relevant RNA-seq data, and provide function annotation for approximately 500 species.

# Chapter 2

# Data input and pre-processing

## 2.1 Data input

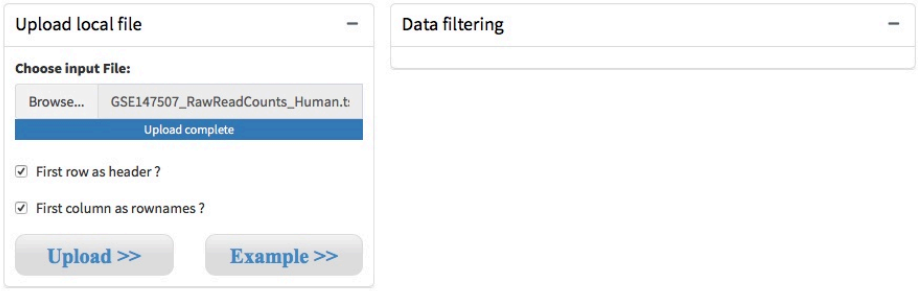
QRAP can support to analyze publicly available and user-generated data. There are two action buttons, 'Get Started Local' and 'Get Started GEO', can activate corresponding data upload and processing pipeline.


### 2.1.1 Upload local file

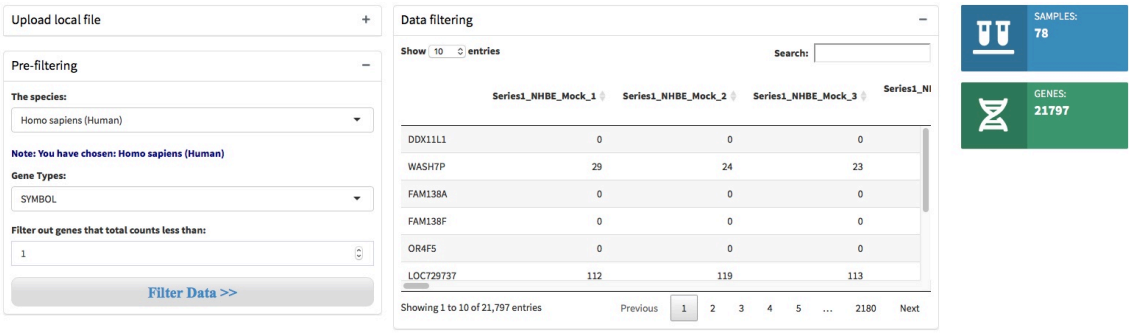
For user generated local files, click the 'Get Started Local' button to enter the data upload page.

***Parameters in this section:***

- Choose input file: browser and upload the expression matrix file, accept .csv/.tsv/.tab/.txt format.
- First row as header ? This means use the first row of the expression matrix as column names, often is samples names.
- First column as rownames ? This means use the first column of the expression matrix as row names, often is gene names.




After select the local files and set the parameters, click the 'Upload' button to preview the expression matrix.

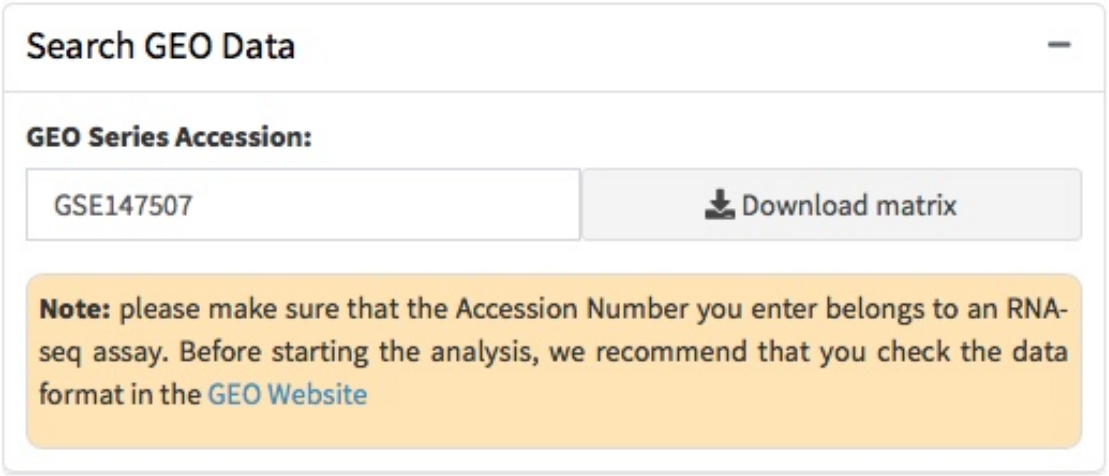



Click 'Example' will upload the example dataset internal, transcriptomes of SARS-Cov-2 infected normal human bronchial epithelial cells (GSE147507).

### 2.1.2 Pull down GEO datasets

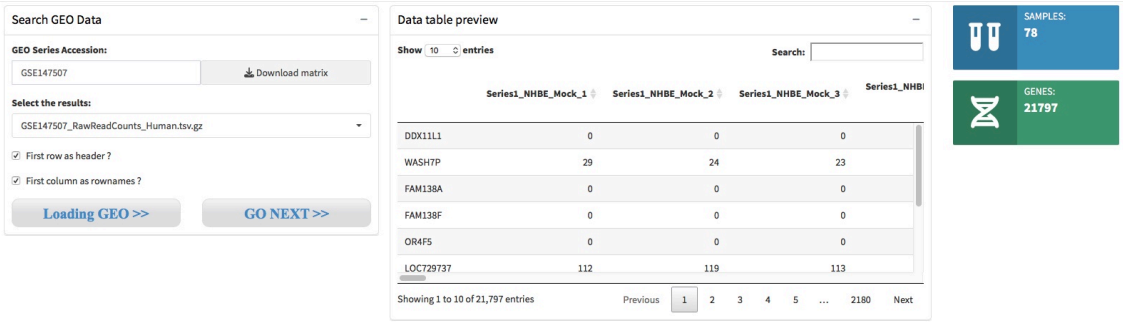
The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes comprehensive sets of microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community.

After you input the GEO Acession Number and active the 'Download matrix' button, We will download the value matrix tables within GDSxxx or supplementary files within GSExxx, these files will store in the working directory of the R project you created.



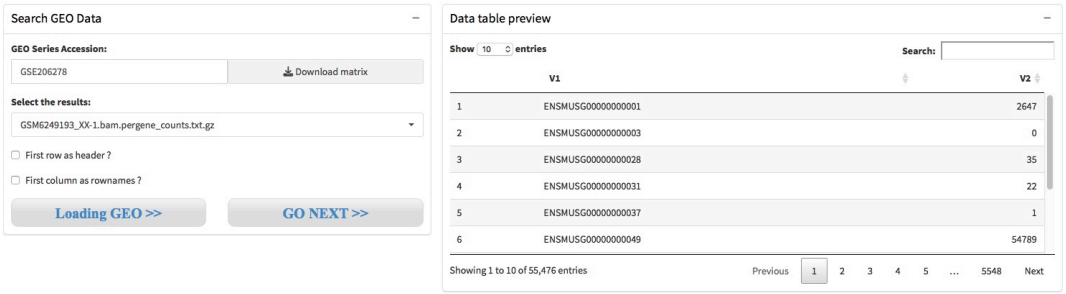

When the files download accomplished, there will show the download files name in the Parameter setting panel, and you should select file(s) that contain interested gene expression matrix and

active the 'Loading GEO' button to preview the matrix.

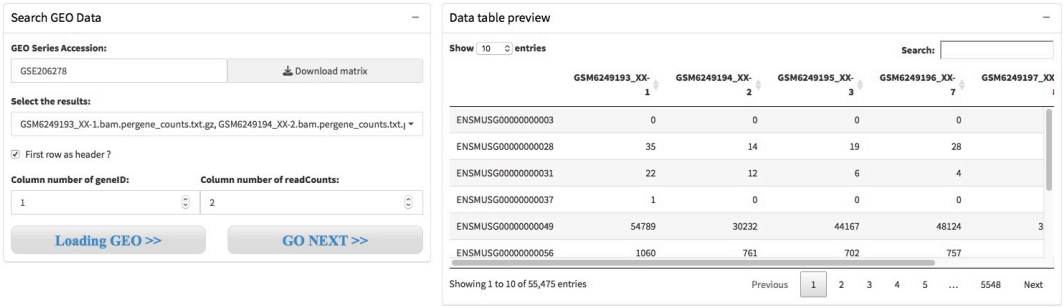



Please Note that if the files are in a tar archive format, such as htseq-count generated results, these files will contain Gene ID and Gene Expression Value of each sample, respectively. Therefore, you need to provid the column number of the Gene ID and Gene Expression Value to help merge the files to generate an analysis ready gene expression matrix.

- Step 1. Choose a single file, and leave the 'First row as header' and 'First column as rownames' unchecked, then click the button 'Loading GEO' to preview what the file contained.

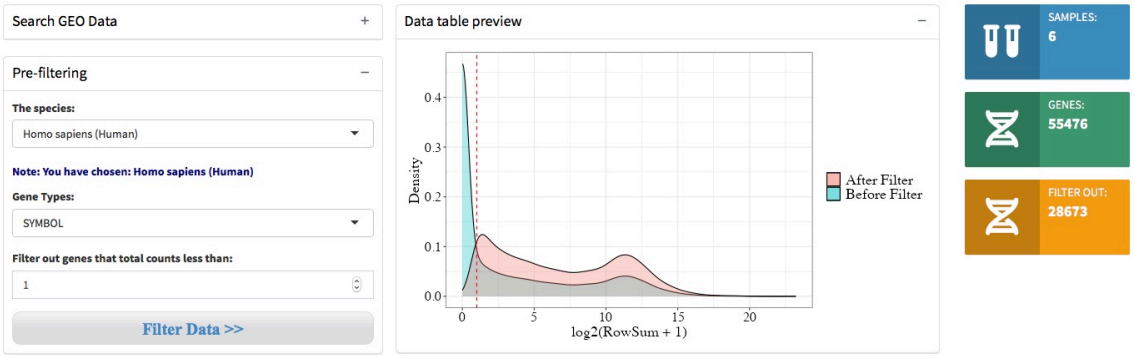



- Step 2. Choose all or interested files, set the column number of geneID and readCounts, then click the button 'Loading GEO' to generate and preview the analysis ready expression matrix.




## 2.2 Data pre-processing

QRAP starts with a read-counts matrix or GEO accession number, and then filters out low expression genes under an given threshold. We then specify the organisms and filtered out low expression genes.




**Note the Gene Types:**

| SYMBOL | ENSEMBL | ENTREZID |
|---|---|---|
| GAPDH | ENSG00000111640 | 2597 |
| TP53 | ENSG00000141510 | 7157 |

# Chapter 3

# Data quality exploring

We describe our methods in this chapter.

## 3.1 principal component analysis (PCA)

## 3.2 hierarchical clustering heatmap

## 3.3 sample-to-sample heatmap

## 3.4 sample correlation coefficient

# Chapter 4

# Differential expression analysis

Some *significant* applications are demonstrated in this chapter.

## 4.1 Extract DEGs

## 4.2 Visualize DEGs

# Chapter 5

# DEG expression pattern detection

We have finished a nice book.

## 5.1 Detect DEG expression pattern

## 5.2 Visualize DEG expression pattern

# Chapter 6

# weighted correlation network analysis

We have finished a nice book.

## 6.1 Data preparation

## 6.2 Soft threshold detection

## 6.3 Gene module detection

## 6.4 Module-Traits relationship

## 6.5 MM vs. GS scatterplot

## 6.6 Module gene expression visualization

# Chapter 7

# functional enrichment analysis

We have finished a nice book.

## 7.1 Gprofiler API

## 7.2 ClusterProfiler

### 7.2.1 ORA

### 7.2.2 GSEA

# Chapter 8

# Gene regulatory network

We have finished a nice book.

## 8.1 KEGG Pathview

## 8.2 PPI network

## 8.3 GENIE3 inffered network

# Chapter 9

# Summary of genes and functions

We have finished a nice book.

## 9.1 summarize genes

## 9.2 summarize functions