PathwayKO: An R package for knock-out pathway enrichment analysis

September 30, 2020

Hannan Ai (allenai94@icloud.com)

1 State Key Laboratory for Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, 510275, P R China.

2 Department of Electrical & Computer Engineering, Grainger College of Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801-2918, USA.

Abstract

The *PathwayKO* package provides a novel solution to knock-out (KO) pathway enrichment analysis. This package contains multiple modules of preprocessing, ROC/AUC calculating, statistics analyzing, and visualizing. Outputs include the list of differentially expressed (DE) genes and the rendered true positive KO KEGG pathways that contain and are impacted by the KO gene at the system-level, and are correctly identified as significant by the pathway enrichment analysis. The ROC-curve based statistics analysis allows comparing the performance of pathway enrichment analysis methods in terms of AUC, partial AUCs, accuracy, precision and recall. This vignette illustrates the installation and common usages, by reproducing the figures presented in our publication as examples.

Website

https://github.com/allenaigit/pathwayko/

Maintainer

Hannan Ai (allenai94@icloud.com)

Acknowledgments

The work was financially supported by the Major Program of National Science and Technology of China (2014ZX0801105B-002) and the Supercomputing Program of National Natural Science Foundation of China (No. U1501501-534) to Y.A.

1. Cite our work

Please cite the *PathwayKO* paper when you use this package. This will help to support the maintenance and growth of open source projects.

Hannan Ai, Fanmei Meng and Yuncan Ai. PathwayKO: An integrated platform for deciphering the systems-level signaling pathways. (submitted)

2. Installation

It is assumed that R and Bioconductor have been installed correctly and accessible. Otherwise, please follow the instructions on R website and Bioconductor website, or contact your system administrator for additional help. It is strongly recommend that R and Bioconductor be upgraded to the latest available versions (4.0.0 or above for R and 3.11 or above for Bioconductor).

2.1 Acquiring *PathwayKO*

PathwayKO is available at Github Page

2.2 Acquiring Additional Supporting Data

Additional supporting data are necessary for demo if you intend to follow examples laid out in the *Preprocessing* section of this guide; otherwise you may opt not to download them.

Additional supporting data for demo are available here. *KO_GEO_datasets* contains raw GEO data obtained from NCBI database for demo purpose, and *CustomCDF_mmu* contains necessary chip-definition and annotation information to process the demo datasets. Additional custom CDFs can be acquired from MicroArray Lab.

2.3 Installation From Github

Installing from Github directly requires R package *devtools*, to install it, in your R session:

```
> install.packages("devtools")
> library(devtools)
> devtools::install_github("allenaigit/pathwayko")
```

2.4 Installation From Local Directory

Download "pathwayko-main.zip" from Github, unzip it. In your R session:

```
> install.packages("/path_to_unzipped_file/pathwayko-main", repo=NULL)
```

2.5 Dependencies

If R finds missing dependencies or needs additional packages during installation process, it should prompt you to install them in your R session. If the installation process failed, just install the missing packages as recommended, then try installing again. You may need multiple attempts at installing *PathwayKO* before being able to tackle all dependency issues.

3. Overview

The *PathwayKO* package is to provide a complete solution to knock-out (KO) pathway enrichment analysis via utilizing mouse GEO datasets (from NCBI database) and KEGG pathways (from KEGG) (Kanehisa et al., 2016).

The *PathwayKO* package contains multiple modules of preprocessing, ROC-AUC calculating, statistics analyzing, and visualizing. Each module can be executed separately or jointly as designed, completing a desired job. The general workflow of *PathwayKO* consists of the following stages:
- Preprocessing
- ROC-AUC Calculating
- Statistics Analyzing
- Visualizing

The ROC-curved based statistics analysis (Robin et al., 2011) approach is applied to allow comparing the performance of methods of pathway enrichment analysis in terms of AUC, partial AUCs, accuracy, precision and recall.

The *PathwayKO* package is flexible to incorporate custom methods and currently incorporates the state-of-the-art seven methods: SAFE (Barry et al., 2005), GSEA (Subramanian et al., 2005), GSA (Efron and Tibshirani 2007), ROntoTools_PE (Draghici et al., 2007), SPIA (Tarca et al., 2009), PADOG (Tarca et al., 2012) and ROntoTools_pDIS (Voichita et al., 2012).

By reproducing the thirteen figures presented in our publication as examples, we illustrate the general workflow of *PathwayKO* throughout the subsequent sections: (4) Preprocessing, (5) ROC-AUC Calculating, (6) Statistics Analyzing, and (7) Visualizing.

4. Preprocessing

`preprocess` processes a KO GEO dataset via Oligo (Carvalho and Irizarry, 2010) and Limma (Smyth, 2004; Ritchie et al., 2015) to yield intermediate outputs (pData.csv and PREP.RData) after RMA normalization.

Note: Examples in this section require additional data, not included in the *PathwayKO* R package, see the *Acquiring Additional Supporting Data* section for more information.

From the directory where demo datasets are located, launch R session.

```
> library(pathwayko)
> preprocess()

INFO: loading libraries...
INFO: series matrix files in current directory:
     [1] "./GSE146756_series_matrix.txt.gz"
     [2] "./GSE22873_series_matrix.txt.gz"
     [3] "./GSE35160_series_matrix.txt.gz"
     [4] "./GSE39864_series_matrix.txt.gz"
     [5] "./GSE411_series_matrix.txt.gz"
```

```
     [6] "./GSE6837-GPL1261_series_matrix.txt.gz"
     [7] "./GSE70302_series_matrix.txt.gz"
     [8] "./GSE8969_series_matrix.txt.gz"
     [9] "./GSE9037_series_matrix.txt.gz"
    [10] "./GSE95806_series_matrix.txt.gz"
Enter data index: 1
INFO: processing series matrix files...
INFO: series matrix processed.
INFO: CEL archives in current directory:
     [1] "./GSE146756_RAW.tar" "./GSE22873_RAW.tar"  "./GSE35160_RAW.tar"
     [4] "./GSE39864_RAW.tar"  "./GSE411_RAW.tar"    "./GSE6837_RAW.tar"
     [7] "./GSE70302_RAW.tar"  "./GSE8969_RAW.tar"   "./GSE9037_RAW.tar"
    [10] "./GSE95806_RAW.tar"
Enter data index: 1
Enter an user defined identifier: (e.g. GSE22873_MKO): GSE146756
INFO: processing CEL files...
    Reading in : GSM4405436_AY1_MoST.CEL.gz
    Reading in : GSM4405437_AY2_MoST.CEL.gz
    Reading in : GSM4405438_AY3_MoST.CEL.gz
    Reading in : GSM4405439_AY4_MoST.CEL.gz
    Reading in : GSM4405440_AY5_MoST.CEL.gz
    Reading in : GSM4405441_AY6_MoST.CEL.gz
    Reading in : GSM4405442_AY7_MoST.CEL.gz
    Reading in : GSM4405443_AY8_MoST.CEL.gz
    Reading in : GSM4405444_AY9_MoST.CEL.gz
    Reading in : GSM4405445_AY10_MoST.CEL.gz
INFO: all CEL files processed.
INFO: processing experiment designs...
INFO: Reference pData:
                                  title     organism platform       source
    GSM4405436 wild-type biological rep 1 Mus musculus  GPL6246 E13.5 embryo
    GSM4405437   LKB1 KO biological rep 1 Mus musculus  GPL6246 E13.5 embryo
    GSM4405438 wild-type biological rep 2 Mus musculus  GPL6246 E13.5 embryo
    GSM4405439   LKB1 KO biological rep 2 Mus musculus  GPL6246 E13.5 embryo
    GSM4405440 wild-type biological rep 3 Mus musculus  GPL6246 E13.5 embryo
    GSM4405441   LKB1 KO biological rep 3 Mus musculus  GPL6246 E13.5 embryo
    GSM4405442 wild-type biological rep 4 Mus musculus  GPL6246 E13.5 embryo
    GSM4405443   LKB1 KO biological rep 4 Mus musculus  GPL6246 E13.5 embryo
    GSM4405444 wild-type biological rep 5 Mus musculus  GPL6246 E13.5 embryo
    GSM4405445   LKB1 KO biological rep 5 Mus musculus  GPL6246 E13.5 embryo

Use this pData? (Y/N): Y
```

Note: Users should inspect the displayed pData and come up with appropriate keywords that can be used to separate the KO samples from control samples. Otherwise, you can make changes to the pData csv file generated in your output dataset directory (e.g. GSE146756_preprocess_output), then enter "N" in R to update pData. In the case of GSE146756, the keyword to filter KO samples we choose is "*KO*" and control samples "*wild-type*" from the "*title*" column.

```
Use this pData? (Y/N): Y
                                  title     organism platform       source
```

```
    GSM4405436 wild-type biological rep 1 Mus musculus  GPL6246 E13.5 embryo
    GSM4405437   LKB1 KO biological rep 1 Mus musculus  GPL6246 E13.5 embryo
    GSM4405438 wild-type biological rep 2 Mus musculus  GPL6246 E13.5 embryo
    GSM4405439   LKB1 KO biological rep 2 Mus musculus  GPL6246 E13.5 embryo
    GSM4405440 wild-type biological rep 3 Mus musculus  GPL6246 E13.5 embryo
    GSM4405441   LKB1 KO biological rep 3 Mus musculus  GPL6246 E13.5 embryo
    GSM4405442 wild-type biological rep 4 Mus musculus  GPL6246 E13.5 embryo
    GSM4405443   LKB1 KO biological rep 4 Mus musculus  GPL6246 E13.5 embryo
    GSM4405444 wild-type biological rep 5 Mus musculus  GPL6246 E13.5 embryo
    GSM4405445   LKB1 KO biological rep 5 Mus musculus  GPL6246 E13.5 embryo

Enter case keyword for pattern matching: (e.g. "MKO"): KO
Enter control keyword for pattern matching: (e.g. "WT"): wild-type
Enter which column to match: title(1), source(2): 1
INFO: selected case samples:
    [1] "GSM4405437" "GSM4405439" "GSM4405441" "GSM4405443" "GSM4405445"
INFO: selected control samples:
    [1] "GSM4405436" "GSM4405438" "GSM4405440" "GSM4405442" "GSM4405444"
Continue? (Y/N): Y
```

Note: Samples designated as the KO and control groups will have their respective GSM sample number (e.g. GSM4405437) displayed. If the KO and control groups are correctly labeled, enter "Y" to continue; otherwise enter "N" to try a new pair of keywords.

```
Continue? (Y/N): Y
Enter KO gene symbol:
    (This is case sensitive and underscore deliminated, e.g. Myd88_Ager): Stk11
INFO: Matched genes:
       SYMBOL  ENTREZID                   GENENAME
    1  Stk11    20869     serine/threonine kinase 11
Continue? (Y/N): Y
```

Note: Information on the knock-out gene should be referred to the GEO information. In the case of GSE146756, the *Stk11* gene was KOed. Some genes may have more than one common name. For Stk11, Stk11ip, 1200014D22Rik, BB131189, L, LIP1, LKB, LKB1IP are aliases seen in literatures referring to the same gene. However, the annotation package only recognizes the official gene symbol.

```
Continue? (Y/N): Y
INFO: matched EntrezID(s):
    [1] "20869"
INFO: experiment design processed.
INFO: Performing RMA...
Background correcting... OK
Normalizing... OK
Calculating Expression
INFO: RMA done...
INFO: Saving data object...
INFO: Data object saved to
    GSE146756_preprocess_output/GSE146756_Stk11_RMA_PREP.RData.
INFO: preprocess successfully completed. Exiting...
    [1] TRUE
```

OUTPUTS:

```
- ./GSE146756_preprocess_output/
    - GSE146756_pData.csv
    - GSE146756_Myd88_RMA_PREP.RData
```

5. ROC-AUC Calculating

Pathwayko invoking **utilities**, **highedges**, **functions** and **methods** calculates the scores of all edges (Hanoudi et al., 2017) in a known global KEGG graph (Zhang and Wiemann, 2009) to yield the distribution of edge scores, and thus automatically determines a change point on the distribution of edge scores by the change-point analysis method (Killick and Eckley, 2014). Outputs include the KO gene-associated sub-graph, the list of differential expression (DE) genes, and the list of true positive KO KEGG pathways.

By **Pathwayko**, we build a ROC curve for each method on each KO GEO dataset, and compute AUC and partial AUCs for a region separately focusing on the 90–100% of specificity and sensitivity in both original and corrected (McClish, 1989) formats. Multiple ROC curves (with AUCs) can intuitively display a superiority among methods. True positive rates, true negative rates, false positive rates and false negative rates are calculated followed the definitions (Nguyen et al., 2019), and transformed into sensitivity, specificity, accuracy, precision and recall (Robin et al., 2011), plus the Youden's best p-value threshold (Youden, 1950). This main process will yield a bunch of results, which is the time-limiting step. For example, it will take about 20 minutes to complete one KO GEO dataset with the incorporated 7 methods, working on a DELL workstation with 32 cores and 192 GB memory.

Ten preprocessed datasets are included in the *PathwayKO* package located at "inst/extdata/geodata" within the directory where *PathwayKO* was installed.

5.1 For Single Dataset

We use the preprocessed dataset, GSE146756, as example. Set your working directory to where the preprocessed "_PREP.RData" files are at (e.g.,GSE146756_preprocess_output), in your R session:

```
> pathwayko()

INFO: starting pathwayko...
INFO: Acquiring user parameters...
INFO: files in current working directory:
    [1] "./GSE146756_preprocess_output/GSE146756_Stk11_RMA_PREP.RData"
Enter preprocessed data index: 1
INFO: directories in current working directory:
    [1] "./KO_GEO_datasets"
    [2] "./GSE146756_preprocess_output"
    [3] "../R/x86_64-pc-linux-gnu-library/4.0/pathwayko/extdata/mmuKEGGxml"
Enter KEGG xml directory index: 3
```

Note: The *Pathwayko* package comes with cached mouse KEGG pathway xml files; you can also download your KEGG pathway xmls from KEGG.

```
INFO: SPIA data in current working directory:
    [1] "../R/x86_64-pc-linux-gnu-library/4.0/pathwayko/extdata/mmuSPIA.RData"
Enter SPIA data index: 1
Run SPIA?: (Y/N) Y
Run ROntoTools PE?: (Y/N) Y
Run ROntoTools PDIS?: (Y/N) Y
Run PADOG?: (Y/N) Y
Run GSA?: (Y/N) Y
Run SAFE?: (Y/N) Y
Run GSEA?: (Y/N) Y
How to choose DE genes? ('HES1','HES2','HES3' for HighEdgeS lower,
    change point, upper bound, 'CLA' for Classical)
Note: geneset-based methods will not use DE genes regardlessly: HES1
Lower bound of specificity (X-axis) in partial ROC? (e.g. 90): 90
Lower bound of sensitivity (Y-axis) in partial ROC? (e.g. 90): 90
INFO: preprocessing...
INFO: 505 genes are considered differentially expressed.
INFO: target pathways:
    [1] "03460" "04068" "04140" "04150" "04151" "04152" "04211" "04530" "04920"
INFO: pathway analysis...
INFO: plotting ROC...
INFO: pathwayko completed.
    [1] TRUE
```

OUTPUTS:

```
./GSE146756_result_output/
    - GSE146756_RMA_Myd88_HES1_SUM.RData
    - GSE146756_RMA_Myd88_HES1_HighEdgeS_plot.pdf
    - GSE146756_RMA_Myd88_HES1_ROC_combined.pdf
    ......
```

5.2 For Multiple Datasets

We use the preprocessed datasets, GSE146756 and GSE22873, as examples. Parameters used for GSE22873 preprocessing are:

```
Enter case keyword for pattern matching: (e.g. "MKO"): MKO
Enter control keyword for pattern matching: (e.g. "WT"): WT
Enter which column to match: title(1), source(2): 1
Enter KO gene symbol:
    (This is case sensitive and underscore delimited, e.g. Myd88_Ager): Myd88
INFO: Matched genes:
      SYMBOL    ENTREZID                                        GENENAME
    1  Myd88    17874           myeloid differentiation primary response gene 88
```

Set your working directory to where these two preprocessed "_PREP.RData" files are at, in your R session:

```
> pathwayko_batch()

INFO: Acquiring user parameters...
INFO: files in current working directory:
    [1] "./GSE146756_preprocess_output/GSE146756_Stk11_RMA_PREP.RData"
    [2] "./GSE22873_preprocess_output/GSE22873_Myd88_RMA_PREP.RData"
Process all?: (Y/N) Y
INFO: successfully queued 2 RData files.
INFO: directories in current working directory:
    [1] "./KO_GEO_datasets"
    [2] "./KO_GEO_datasets/GSE146756_preprocess_output"
    [3] "./KO_GEO_datasets/GSE146756_result_output"
    [4] "./KO_GEO_datasets/GSE22873_preprocess_output"
    [5] "../R/x86_64-pc-linux-gnu-library/4.0/pathwayko/extdata/mmuKEGGxml"
Enter KEGG xml directory index: 5
INFO: SPIA data in current working directory:
    [1] "../R/x86_64-pc-linux-gnu-library/4.0/pathwayko/extdata/mmuSPIA.RData"
Enter SPIA data index: 1
Run SPIA?: (Y/N) Y
Run ROntoTools PE?: (Y/N) Y
Run ROntoTools PDIS?: (Y/N) Y
Run PADOG?: (Y/N) Y
Run GSA?: (Y/N) Y
Run SAFE?: (Y/N) Y
Run GSEA?: (Y/N) Y
How to choose DE genes? ('HES1','HES2','HES3' for HighEdgeS lower,
    change point, upper bound, 'CLA' for Classical)
Note: geneset-based methods will not use DE genes regardlessly: HES1
Lower bound of specificity (X-axis) in partial ROC? (e.g. 90): 90
Lower bound of sensitivity (Y-axis) in partial ROC? (e.g. 90): 90
```

Note: Expect 20 minutes per dataset and up to 2GB RAM usage per method selected. Total 10 KO GEO datasets will be completed in around 3 hrs on a DELL workstation with 32 cores and 192 GB memory.

```
INFO: batch job ( 1 / 2 )...
INFO: starting pathwayko...
INFO: preprocessing...
INFO: 505 genes are considered differentially expressed.
INFO: target pathways:
    [1] "03460" "04068" "04140" "04150" "04151" "04152" "04211" "04530" "04920"
INFO: pathway analysis...
INFO: plotting ROC...
INFO: pathwayko completed.
INFO: batch job ( 1 / 2 ) completed.

INFO: batch job ( 2 / 2 )...
INFO: starting pathwayko...
INFO: preprocessing...
INFO: 122 genes are considered differentially expressed.
INFO: target pathways:
     [1] "04010" "04064" "04620" "04621" "05132" "05133" "05134" "05135" "05140"
```

```
    [10] "05142" "05143" "05144" "05145" "05152" "05161" "05162" "05164" "05168"
    [19] "05169" "05170" "05235"
INFO: pathway analysis...
INFO: plotting ROC...
INFO: pathwayko completed.
INFO: batch job ( 2 / 2 ) completed.
    [1] TRUE
```

OUTPUTS:

```
- ./GSE146756_result_output/
    - GSE146756_RMA_Myd88_HES1_SUM.RData
    - GSE146756_RMA_Myd88_HES1_Histogram_plot.pdf
    - GSE146756_RMA_Myd88_HES1_HighEdgeS_plot.pdf
    - GSE146756_RMA_Myd88_HES1_SPIA_ROC_plot.pdf
    - GSE146756_RMA_Myd88_HES1_SPIA_CI_plot.pdf
    - GSE146756_RMA_Myd88_HES1_SPAI_pAUC_plot.pdf
    - GSE146756_RMA_Myd88_HES1_ROC_combined.pdf
    ......
- ./GSE22873_result_output/
    - GSE22873_RMA_Myd88_HES1_SUM.RData
    - GSE22873_RMA_Myd88_HES1_Histogram_plot.pdf (Figure 1)
    - GSE22873_RMA_Myd88_HES1_HighEdgeS_plot.pdf (Figure 2)
    - GSE22873_RMA_Myd88_HES1_SPIA_ROC.pdf (Figure 3)
    - GSE22873_RMA_Myd88_HES1_SPIA_CI.pdf (Figure 4)
    - GSE22873_RMA_Myd88_HES1_SPIA_pAUC.pdf (Figure 5)
    - GSE22873_RMA_Myd88_HES1_ROC_combined.pdf (Figure 6)
    ......
```

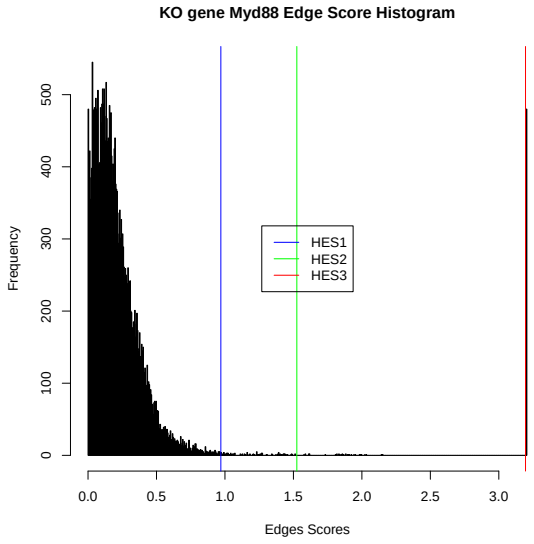

Figure 1: GSE22873_RMA_Myd88_HES1_Histogram_plot.pdf

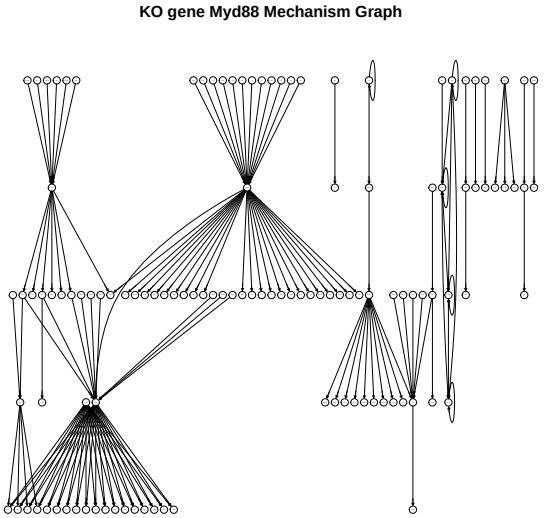

Figure 2: GSE22873_RMA_Myd88_HES1_HighEdgeS_plot.pdf

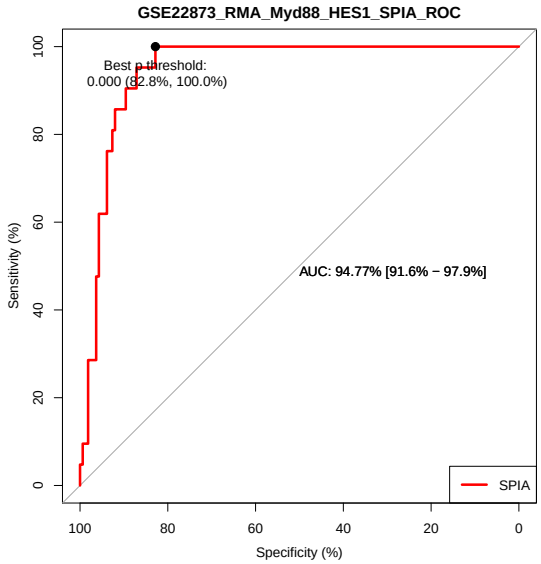

Figure 3: GSE22873_RMA_Myd88_HES1_SPIA_ROC_plot.pdf

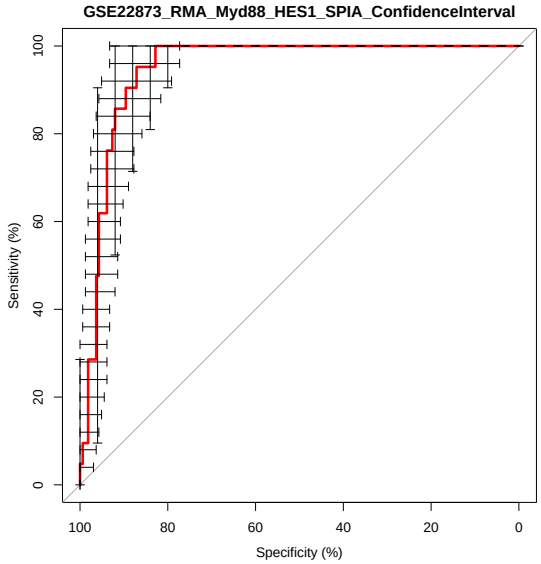

Figure 4: GSE22873_RMA_Myd88_HES1_SPIA_CI_plot.pdf

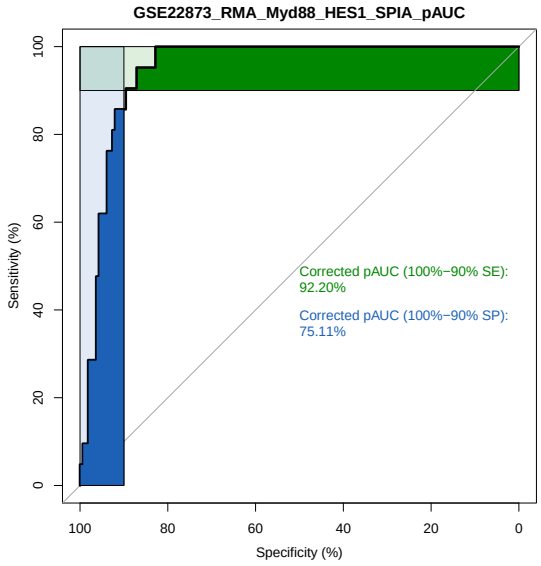

Figure 5: GSE22873_RMA_Myd88_HES1_SPIA_pAUC_plot.pdf

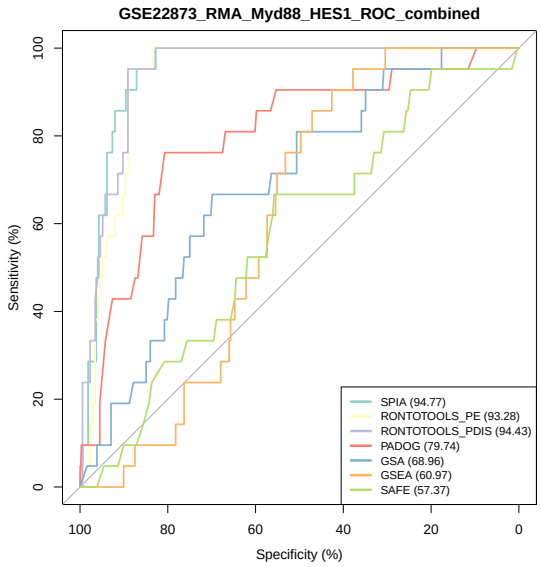

Figure 6: GSE22873_RMA_Myd88_HES1_ROC_combined.pdf

6. Statistics Analyzing

6.1 `combineresult()`

`combineresult` yields STATS.RData, summarizing the intermediate outputs such as true positive rates (TPR), true negative rates (TNR), false positive rates (FPR) and false negative rates (FNR) calculated by the definitions (Nguyen et al., 2019), plus the Youden's best p-value threshold (Youden, 1950), sensitivity (TPR), specificity (TNR), accuracy, precision and recall (Robin et al., 2011).

We will use the ten demo datasets after applying ROC-AUC Calculating with the same parameters as examples. Before statistics analyzing can be applied to multiple datasets, all ROC-AUC Calculating results have to be combined into a single object.

Note: Only ROC-AUC Calculating results of the same parameters should be combined for further analysis, combining ROC-AUC Calculating results of different parameters will cause errors and/or inconclusive analysis.

Set your working directory to where all "_SUM.RData" files are located, in your R session:

```
> combineresult()

INFO: files in current working directory:
     [1] "./GSE146756_result_output/GSE146756_RMA_Stk11_HES1_SUM.RData"
     [2] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_SUM.RData"
     [3] "./GSE35160_result_output/GSE35160_RMA_Nfe2l2_HES1_SUM.RData"
     [4] "./GSE39864_result_output/GSE39864_RMA_Gata3_HES1_SUM.RData"
     [5] "./GSE411_400_RMA_Socs3_HES1_SUM.RData"
     [6] "./GSE6837_result_output/GSE6837_RMA_Ikbke_HES1_SUM.RData"
     [7] "./GSE70302_result_output/GSE70302_RMA_Il1a_HES1_SUM.RData"
```

```
     [8] "./GSE8969_result_output/GSE8969_RMA_Keap1_HES1_SUM.RData"
     [9] "./GSE9037_result_output/GSE9037_RMA_Irak4_HES1_SUM.RData"
    [10] "./GSE95806_result_output/GSE95806_RMA_Trp53_HES1_SUM.RData"
Process all?: (Y/N) Y
INFO: successfully loaded 10 RData files.
INFO: combining results...
INFO: all process completed.
    [1] TRUE
```

OUTPUTS:

```
- ./combinedresults/
    - 2020-09-30.STATS.RData
    - Summary.AUC.csv
    - Summary.pAUC_SE.csv
    - Summary.pAUC_SP.csv
    - Summary.parameters.csv
```

6.2 `violinplot()`

`violinplot` displays 12 key metrics including the Youden's best p-value threshold, specificity (i.e., TNR), sensitivity (i.e., TPR), FDR, FPR, FNR, accuracy, precision, recall, AUC, pAUC_SP and pAUC_SE across 7 methods when benchmarked on the mouse 10 KO GEO datasets (GSE22873_Myd88, GSE411_Socs3, GSE6837_Ikbke, GSE8969_Keap1, GSE35160_Nfe2l2, GSE39864 Gata3, GSE70302 Il1a, GSE95806 Trp53, GSE146756 Stk11, GSE9037 Irak4; refer to Supplementary Table S1 in our publication) against the available mouse 333 KEGG pathways.

```
> violinplot()
```

```
INFO: summary files in current working directory:
    [1] "./combinedresults/2020-09-30.STATS.RData"
Enter stats data index: 1
```

OUTPUTS:

```
- ./ViolinPlots/
    - plot.STATS5.comb.pdf (Figure 7)
```

6.3 `wilcoxtest()`

`wilcoxtest` conducts the "paired", pairwise comparisons between each pair of two methods under study when benchmarked on the same 10 KO datasets. **`pairwise.wilcox.test`**(from the R `stats` package) is used with corrections (BH FDR adjustment of p-value) for multiple testing under the two-sided mode.

```
> wilcoxtest()
```

```
INFO: summary files in current working directory:
    [1] "./combinedresults/2020-10-08.STATS.RData"
Enter stats data index: 1
INFO: starting...
```

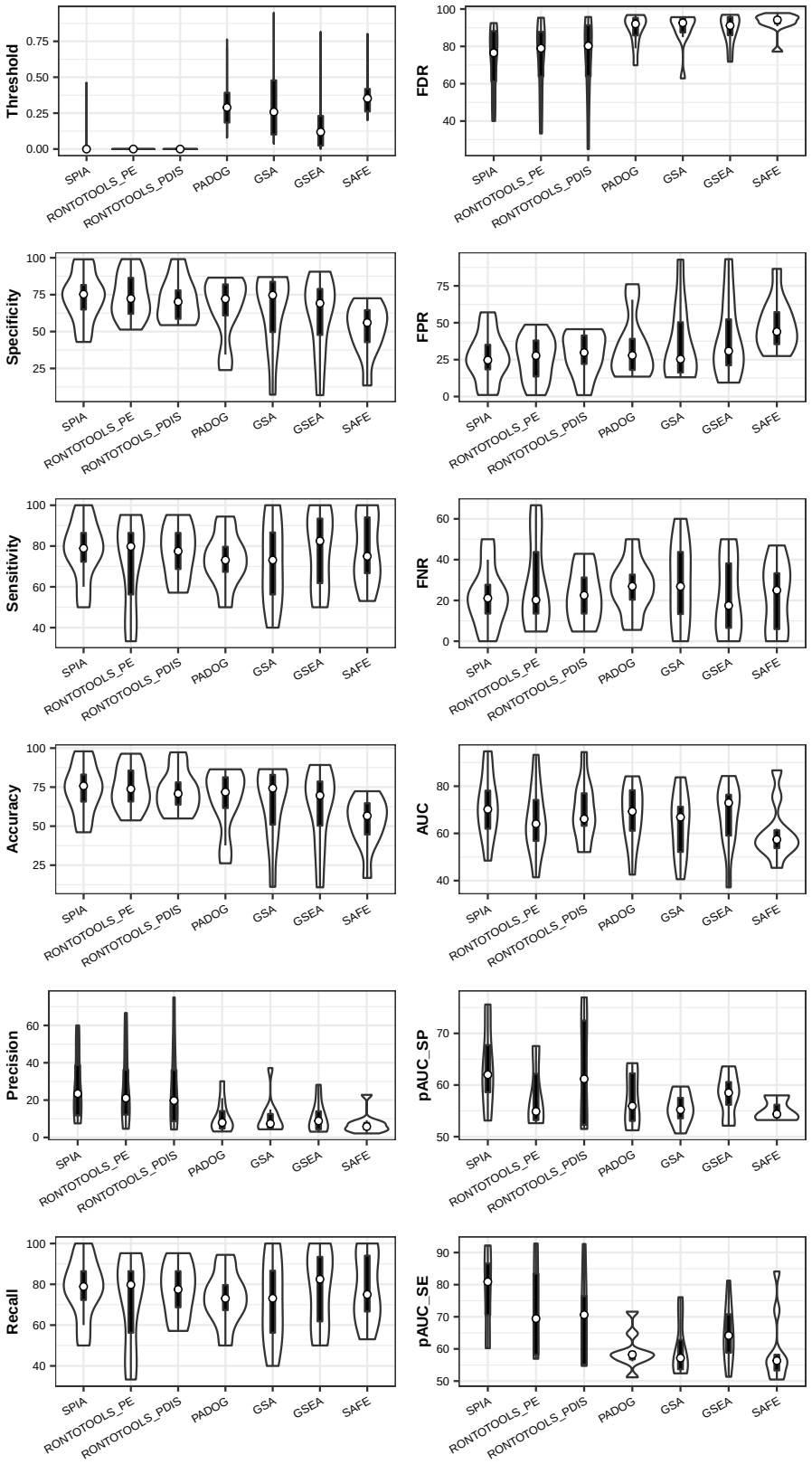

Figure 7: plot.STATS5.comb.pdf

```
INFO: completed.
```

OUTPUTS:

```
- ./Wilcoxtests/
    - wilcox.out.txt
```

Table 1 Pairwise comparisons using Wilcoxon signed rank test with continuity correction

data: AUC and METHOD *P*-value adjustment method: BH

AUC	GSA	GSEA	PADOG	ROntoTools_pDIS	ROntoTools_PE	SAFE
GSEA	0.88	-	-	-	-	-
PADOG	0.44	0.93	-	-	-	-
ROntoTools_pDIS	0.79	1.00	0.88	-	-	-
ROntoTools_PE	0.93	0.88	0.91	0.97	-	-
SAFE	0.79	0.79	0.44	0.79	0.86	-
SPIA	0.79	0.93	0.82	0.79	0.41	0.44

Table 2 Pairwise comparisons using Wilcoxon signed rank test with continuity correction

data: pAUC_SP_cor and METHOD *P*-value adjustment method: BH

AUC	GSA	GSEA	PADOG	ROntoTools_pDIS	ROntoTools_PE	SAFE
GSEA	1.00	-	-	-	-	-
PADOG	0.88	1.00	-	-	-	-
ROntoTools_pDIS	0.53	1.00	0.58	-	-	-
ROntoTools_PE	1.00	0.55	1.00	0.58	-	-
SAFE	0.58	0.88	1.00	1.00	1.00	-
SPIA	0.27	0.41	0.27	1.00	0.27	0.88

Table 3 Pairwise comparisons using Wilcoxon signed rank exact test

data: pAUC_SE_cor and METHOD *P*-value adjustment method: BH

AUC	GSA	GSEA	PADOG	ROntoTools_pDIS	ROntoTools_PE	SAFE
GSEA	0.90	-	-	-	-	-
PADOG	0.92	0.44	-	-	-	-
ROntoTools_pDIS	0.44	0.90	0.44	-	-	-
ROntoTools_PE	0.44	0.92	0.44	0.44	-	-
SAFE	0.92	0.47	0.92	0.44	0.44	-
SPIA	0.44	0.44	0.44	0.92	1.00	0.44

6.4 evidenceplot()

evidenceplot generates two-dimensional evidence plots of the probability of random distribution and accumulated perturbation as characterized by SPIA (Tarca et al., 2009), which illustrate the substantial changes. We use GSE22873 as an example:

```
> evidenceplot()

INFO: summary files in current working directory:
     [1] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_SUM.RData"
Enter summary data index: 1
INFO: choose method:
    [1] "SPIA"
Enter method index: 1
INFO: completed.
```

OUTPUTS:

```
- ./GSE22873_EvidencePlot_output/
    - GSE22873_RMA_Myd88_HES1_SPIA.evidence_plot.pdf (Figure 8)
```

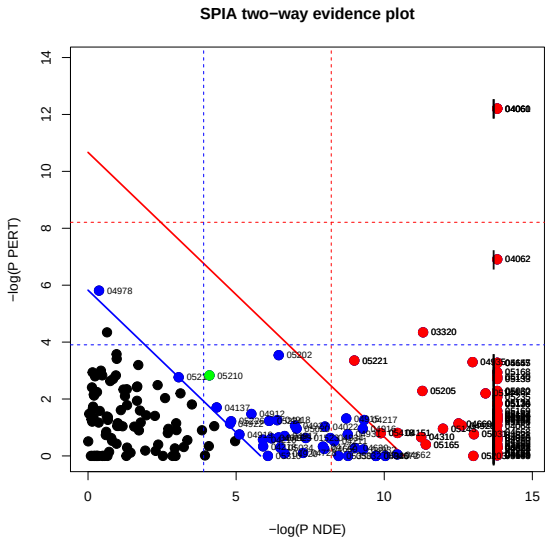

Figure 8: GSE22873_RMA_Myd88_HES1_SPIA.evidence_plot.pdf

6.5 roctest()

The ROC curve-based statistical hypothesis testing is applied to compare two ROC curves (Robin et al., 2011) for the ROC curves themselves (by venkatraman), for the full AUC and for the partial pAUCs (both by bootstrap). For example, **roctest** conducts the ROC curve-based statistical hypothesis testing, which suggests that SPIA significantly outperforms PADOG in terms of the two ROC curves themselves ($p<0.05$), AUC ($p<0.01$), pAUC_SP ($p<0.1$) and pAUC_SE ($p<0.001$) when benchmarked on GSE22873 (KO Myd88).

```
> roctest()

INFO: summary files in current working directory:
     [1] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_SUM.RData"
Enter summary data index: 1
INFO: methods run:
    [1] "RONTOTOOLS_PE"   "RONTOTOOLS_PDIS" "SPIA"            "PADOG"
    [5] "GSA"             "SAFE"            "GSEA"
Enter index of method one: 3
INFO: first method selected: SPIA
Enter index of method two: 4
INFO: second method selected: PADOG
INFO: processing...
      |======================================================================| 100%
      |======================================================================| 100%
      |======================================================================| 100%
      |======================================================================| 100%
INFO: all process completed.
```

OUTPUTS:

- ./GSE22873_roctest_output/
    - ROC_test.venkatraman.ROCcurves.pdf (Figure 9)
    - ROC_test.bootstrap.AUC.pdf (Figure 10)
    - ROC_test.bootstrap.pAUC_SP.pdf (Figure 11)
    - ROC_test.bootstrap.pAUC_SE.pdf (Figure 12)

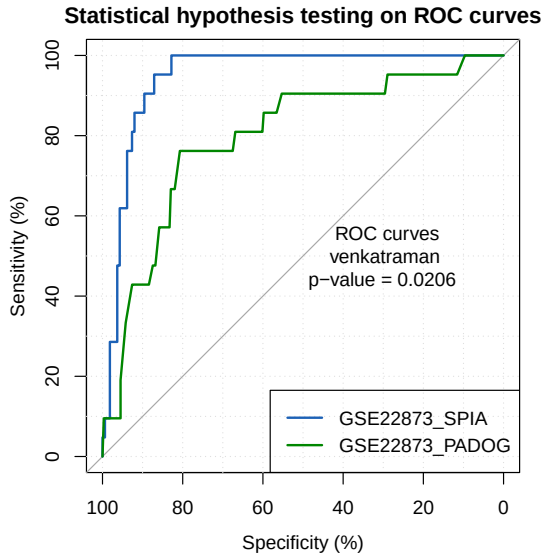

Figure 9: ROC_test.venkatraman.ROCcurves.pdf

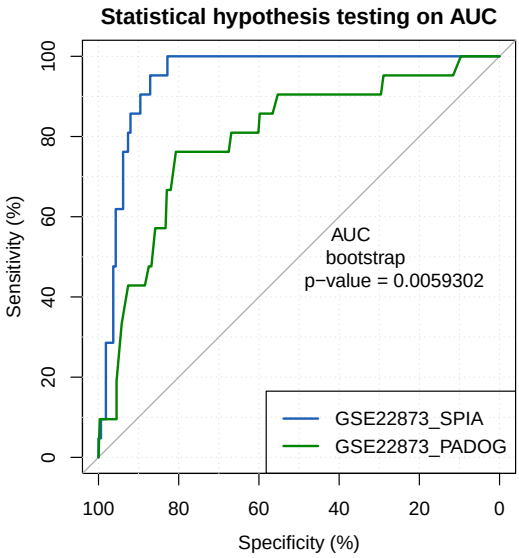

Figure 10: ROC_test.bootstrap.AUC.pdf

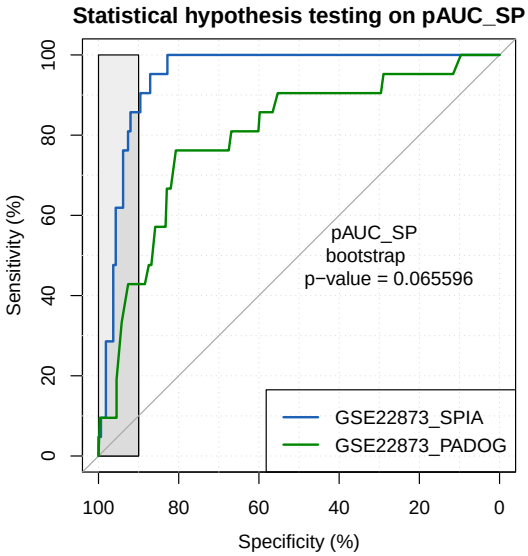

Figure 11: ROC_test.bootstrap.pAUC_SP.pdf

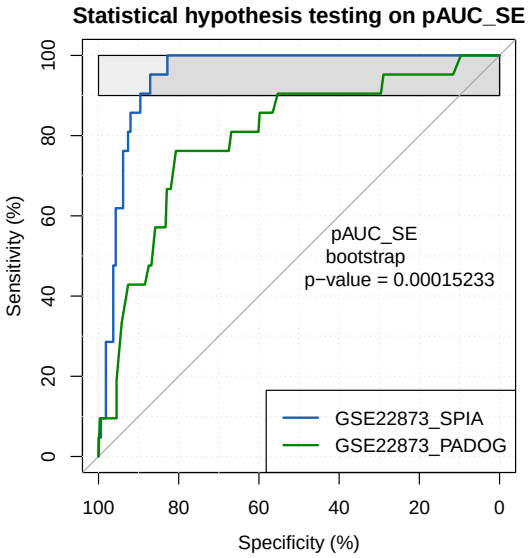

Figure 12: ROC_test.bootstrap.pAUC_SE.pdf

7. Visualizing

7.1 `filtertrue()`

`filtertrue` extracts the true positive KO KEGG pathways that contain and are impacted by the same KO gene. For example, total 21 KO KEGG pathways are identified by SPIA from GSE22873 (KO Myd88).

```
> filtertrue()

INFO: directories in current directory:
    [1] "./GSE22873_result_output"
Enter data index: 1
INFO: csv files processed in ./GSE22873_result_output :
    [1] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_GSA.csv"
    [2] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_GSEA.csv"
    [3] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_PADOG.csv"
    [4] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_RONTOTOOLS_PDIS.csv"
    [5] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_RONTOTOOLS_PE.csv"
    [6] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_SAFE.csv"
    [7] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_SPIA.csv"
INFO: generating GSE22873_RMA_Myd88_HES1_GSA.PostProc.TRUE.csv...
INFO: generating GSE22873_RMA_Myd88_HES1_GSEA.PostProc.TRUE.csv...
INFO: generating GSE22873_RMA_Myd88_HES1_PADOG.PostProc.TRUE.csv...
INFO: generating GSE22873_RMA_Myd88_HES1_RONTOTOOLS_PDIS.PostProc.TRUE.csv...
INFO: generating GSE22873_RMA_Myd88_HES1_RONTOTOOLS_PE.PostProc.TRUE.csv...
INFO: generating GSE22873_RMA_Myd88_HES1_SAFE.PostProc.TRUE.csv...
INFO: generating GSE22873_RMA_Myd88_HES1_SPIA.PostProc.TRUE.csv...
INFO: completed.
```

7.2 pathwayview()

pathwayview renders the true positive KO KEGG pathways with the up- and down-regulated DE genes (Luo et al., 2013), e.g., Toll-like receptor signaling pathway, to highlight the KO KEGG pathway that contains and is impacted by the KO Myd88 gene.

```
> pathwayview()

INFO: summary files in current working directory:
    [1] "./GSE22873_result_output/GSE22873_RMA_Myd88_HES1_SUM.RData"
Enter stats data index: 1
INFO: preprocessed files in current directory:
    [1] "./GSE22873_Myd88_RMA_PREP.RData"
Enter preprocessed data index: 1
INFO: directories in current directory:
    [1] "./R/x86_64-pc-linux-gnu-library/4.0/pathwayko/extdata/mmuKEGGxml"
Enter KEGG data index: 1
INFO: processing...
INFO: all process completed.
```

OUTPUTS:

```
- ./GSE22873_pathview_output/pdfs/
    - mmu04620.GSE22873.Myd88.SPIA.pdf (Figure 13)
    - mmu04010.GSE22873.Myd88.SPIA.pdf
    - mmu04064.GSE22873.Myd88.SPIA.pdf
    - mmu04621.GSE22873.Myd88.SPIA.pdf
    - mmu05132.GSE22873.Myd88.SPIA.pdf
    - mmu05133.GSE22873.Myd88.SPIA.pdf
    - mmu05134.GSE22873.Myd88.SPIA.pdf
    - mmu05135.GSE22873.Myd88.SPIA.pdf
    - mmu05140.GSE22873.Myd88.SPIA.pdf
    - mmu05142.GSE22873.Myd88.SPIA.pdf
    - mmu05143.GSE22873.Myd88.SPIA.pdf
    - mmu05144.GSE22873.Myd88.SPIA.pdf
    - mmu05145.GSE22873.Myd88.SPIA.pdf
    - mmu05152.GSE22873.Myd88.SPIA.pdf
    - mmu05161.GSE22873.Myd88.SPIA.pdf
    - mmu05162.GSE22873.Myd88.SPIA.pdf
    - mmu05164.GSE22873.Myd88.SPIA.pdf
    - mmu05168.GSE22873.Myd88.SPIA.pdf
    - mmu05169.GSE22873.Myd88.SPIA.pdf
    - mmu05170.GSE22873.Myd88.SPIA.pdf
    - mmu05235.GSE22873.Myd88.SPIA.pdf
```

References

Ai H., Meng F., and Ai Y. (2022) PathviewKO: An integrated platform for deciphering the systems-level signaling pathways. (submitted)

Barry W. T. et al. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics, 21, 1943¨C1949.

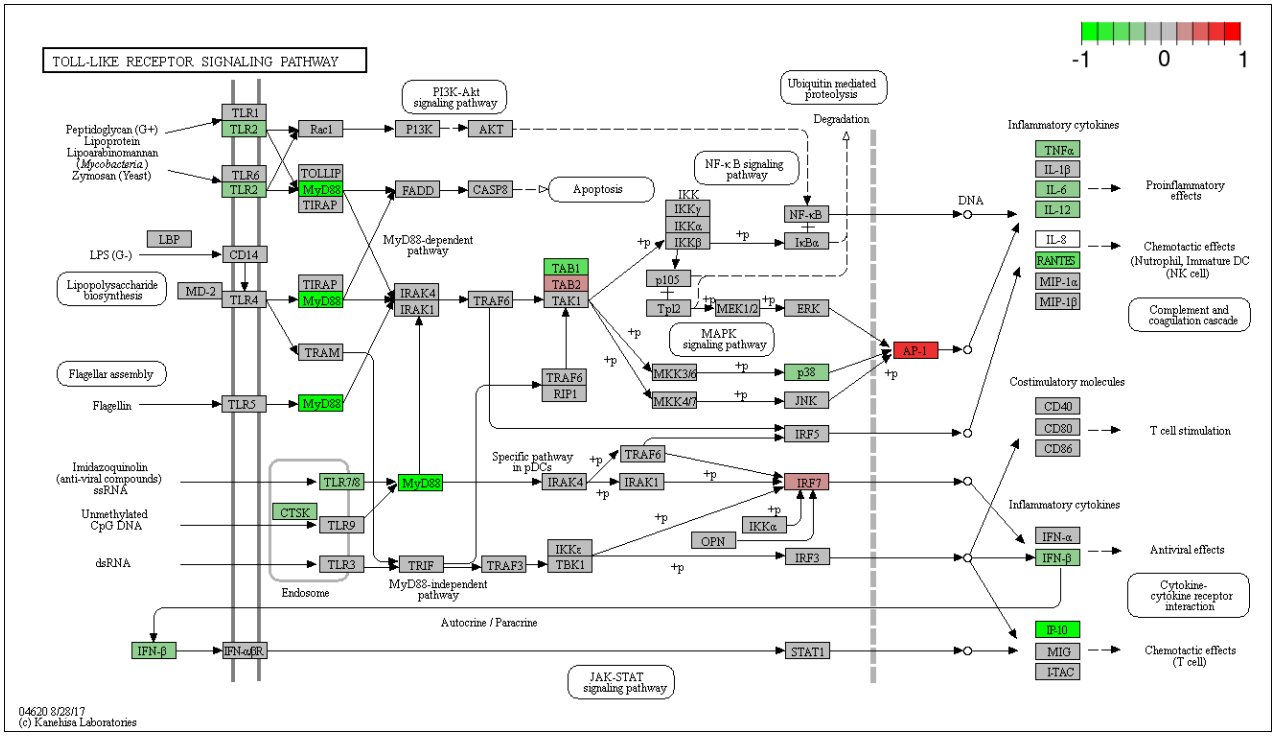

Figure 13: mmu04620.GSE22873.Myd88.SPIA.pdf

Carvalho B.S. and Irizarry R.A. (2010) A framework for oligonucleotide microarray preprocessing. Bioinformatics, 16, 2363¨C2367.

Draghici S. et al. (2007) A systems biology approach for pathway level analysis. Genome Res.,17,1537¨C1545.

Efron B. and Tibshirani R. (2007) On testing the significance of sets of genes. Ann. Appl. Stat., 1, 107¨C129.

Hanoudi S., et al. (2017) Identifying biologically relevant putative mechanisms in a given phenotype comparison. PLoS ONE 12, e0176950.

Kanehisa M. et al. (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D475¨CD462.

Killick R. and Eckley I. (2014) changepoint: An R package for changepoint analysis. J. Statistic Software. 58,1¨C19.

Luo W. and Brouwer C. (2013) Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Bioinformatics, 29,1830¨C1831.

McClish, D.K. (1989) Analyzing a portion of the ROC curve. Med Decis Making 9,190¨C195.

Nguyen T. et al. (2017) DANUBE: Data-driven meta-ANalysis using UnBiased Empirical

distributions - applied to biological pathway analysis. Proc IEEE., 105, 496¨C515.

Nguyen T. et al. (2018) Network-based approaches for pathway level analysis. Curr. Protocols Bioinformatics, 1¨C28.

Nguyen T. et al. (2019) Identifying significantly impacted pathways: a comprehensive review and assessment. Genome Biol., 20, 203.

Ritchie M.E., et al. (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research., 43, e47.

Robin X. et al. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, 77.

Smyth G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat. Appl. Genet. Mol. Biol., 3, DOI: 10.2202/1544-6115.1027.

Subramanian A. et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A, 102, 15545¨C15550.

Tarca A. L. et al. (2009) A novel signaling pathway impact analysis. Bioinformatics, 25, 75¨C82.

Tarca A. L. et al. (2012) Down-weighting overlapping genes improves gene set analysis. BMC Bioinformatics, 13, 136.

Youden W. J. (1950) Index for rating diagnostic tests. Cancer, 3, 32¨C35. DOI: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2¨C3.

Voichita C. et al. (2012) Incorporating gene significance in the impact analysis of signaling pathways. Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA), Dec. 2012, pp. 126¨C131.

Zhang J. D. and Wiemann S. (2009) KEGGgraph: a graph approach to KEGG pathway in R and bioconductor. Bioinformatics, 25, 1470¨C1471.