CellEnrich Manual

2023.11.11 @hainct

Introduction

Testing environment:

1. Ubuntu 22.04.2 LTS
2. R version 4.3.0
3. RStudio 2023.03.0+386 "Cherry Blossom"
4. Rtools 40

NOTE: on a fresh installation, users may need to install some required interpreter compilers for the system to install other R packages:

- C++ compiler
- gfortran compiler (FYI: tips for [MAC](#) users or [other OS](#))
- Seurat >= version 5.0.0 is REQUIRED

Install Dependent Packages:

- Also, some package needs **install from sources which needs compilation,** and it needs proper version of [Rtools](#) installed
- It's recommended to install `devtools`, `BiocManager` packages first, to install following packages in Github / Bioconductor (not in CRAN).

```
library(devtools)
library(BiocManager)
```

Bioconductor Packages:
1. `SingleCellExperiment`
2. `Scran`

```
BiocManager::install('SingleCellExperiment')
BiocManager::install('scran')
```

Other Packages:
1. Seurat
2. farver
3. ggbiplot
4. waiter

```
install.packages('Seurat') # version >= 5.0.0 is REQUIRED!
install.packages('farver')
devtools::install_github('JohnCoene/waiter')
devtools::install_github('vqv/ggbiplot')
```

After installing these packages, we can install CellEnrich now.

Running this script will install most of the other dependent packages.

```
install_github('noobCoding/CellEnrich')
library(CellEnrich)
```

Download Data

CellEnrich provides example data and genesets.

Using this function, users can download them in the current working directory.

```
CellEnrich::DownloadData(type='test') # for available human
related data & genesets

## Use type='all' for all available data
```

Here is a list of data.

- Human Gene-sets
  - Reactome_2022.RData
  - WikiPathways_2021_Human.RData
  - KEGG_2021_Human.RData
  - humanGO.RData
  - humanGOBP.RData
  - humanGOCC.RData
  - humanGOMF.RData

- Mouse Gene-sets
  - WikiPathways_2019_Mouse.RData
  - KEGG_2021_Mouse.RData
  - mouseGO.RData
  - mouseGOBP.RData
  - mouseGOCC.RData
  - mouseGOMF.RData

- Human Data
  - pbmcClustInfo.RData ([Seurat PBMC guide](#))
  - pbmcData.RData
  - [DuoClustering](#)([link](#))
  - koh.RData
  - kohInfo.RData

- Mouse Data
  - [Seurat MCA guide](#)

Tutorial with PBMC 3k data

This tutorial will use PBMC analysis data from [Seurat](#).

To get a complete description of the experiment, refer [here](#)

We assume that all dependent packages are already installed. See **How to install CellEnrich** document.

Required Data

Pre-processed data is included in GitHub [repository](#).

Filename	Parameter	Type
pbmcData.RData	CountData	dgCMatrix
pbmcClustinfo.RData	GroupInfo	Character

Run these codes, users can see an interactive `shiny` page.

```
# Download data, if not downloaded
CellEnrich::DownloadData()

load("pbmcData.RData")
load("pbmcClustInfo.RData")

CountData <- pbmcData
GroupInfo <- pbmcClustInfo

# This will run CellEnrich
CellEnrich(CountData, GroupInfo)
```

Result

Set options before starting CellEnrich.

Possible options are :

- Strategy for Differential Expressed Genes.
- Plot Options
- Gene-set and their size cutoff to be used in the analysis.

After a few minutes, the analysis result will appear.

Test environment used:

- OS : Ubuntu 22.04 LTS
- CPU: iAMD® Ryzen 9 5950x 16-core processor × 32
- RAM: 64GB

CellEnrich consists of 5 modules.

1. Parameter Selection

ⓘ CellEnrich

Options

Methods
- ● CellEnrich - HALF median (coef = 0.5)
- ○ CellEnrich - Median (coef = 1)
- ○ CellEnrich - Mixture

Scatter Plot
- ○ PCA
- ○ TSNE
- ● UMAP

Top-N dims
50

Minimum Gene-set Size
15

Maximum Gene-set Size
500

Pathway Frequency
0.1

Q-value threshold
0.05

Genesets
- ● Reactome
- ○ Human-WikiPathway
- ○ Human-KEGG
- ○ Human-GO
- ○ Human-GO-BP
- ○ Human-GO-CC
- ○ Human-GO-MF
- ○ Mouse-WikiPathway
- ○ Mouse-KEGG
- ○ Mouse-GO
- ○ Mouse-GO-BP
- ○ Mouse-GO-CC
- ○ Mouse-GO-MF

User defined geneset

BROWSE...

RData format is required!

RUN

2. Scatter Area

- The top-left plot is a scatter plot created with PCA, TSNE or UMAP and ggplot2 that the user selected in the option.
- The top-right plot is a histogram plot created with high charter to see a distribution of Group / Cell labels.
- The bottom-left plot shows the significance level (darker is more significant) of enriched pathways dominant in each cell of a group.
- The bottom-right plot shows the significance level of enriched pathways in the whole data.

- Users can use the designed buttons to change the enriched criteria:
    1. ODDS RATIO button will colorize the enriched cell in each group with the **highest** odds ratio.
    2. FREQUENCY button will colorize frequently enriched cells in each group.

- Users also can use the designed buttons to download analyzed figures and results.

3. Pathways Module

This pathway module will show significant pathways for each group based on the Frequency (of pathways in a group), Size, OddsRatio and Q-value (of pathways) in table format.

For each group, **only 1 pathway** can be selected to show in the scatter plot at a time.

B | SELECT

Geneset	Frequency	Size	Qvalue	OddRatio
ANTIGEN PROCESSING AND PRESENTATION OF EXOGENOUS PEPTIDE ANTIGEN VIA MHC CLASS II	343	27	20	79.2278
ER TO GOLGI TRANSPORT VESICLE MEMBRANE	343	51	20	79.0949
ENDOCYTIC VESICLE MEMBRANE	340	45	20	77.6131
POSITIVE REGULATION OF T CELL ACTIVATION	336	33	20	70.1417
TRANSPORT VESICLE MEMBRANE	315	27	20	42.898
ADAPTIVE IMMUNE RESPONSE	299	159	20	15.4034
CLATHRIN-COATED ENDOCYTIC VESICLE MEMBRANE	290	42	20	11.781
B CELL RECEPTOR SIGNALING PATHWAY	98	42	20	9.3323
POSITIVE REGULATION OF T CELL MEDIATED CYTOTOXICITY	291	21	20	8.6265
B CELL ACTIVATION	52	34	20	8.2109

Previous 1 2 3 Next

CD8 T | SELECT

Geneset	Frequency	Size	Qvalue	OddRatio
NATURAL KILLER CELL MEDIATED CYTOTOXICITY	28	22	13.7615	5.4114
T CELL RECEPTOR SIGNALING PATHWAY	40	84	15.9546	4.6352
IMMUNOLOGICAL SYNAPSE	98	42	20	3.1694
RRNA PROCESSING	257	146	5.8533	2.9291
MHC CLASS I PROTEIN BINDING	33	17	6.8354	2.7643
T CELL ACTIVATION	65	55	11.5	2.7447
VESICLE	236	129	3.9055	1.8295
PROTEOLYSIS INVOLVED IN PROTEIN CATABOLIC PROCESS	101	44	5.3877	1.7347
RIBOSOMAL LARGE SUBUNIT ASSEMBLY	245	17	2.0803	1.5907
UNFOLDED PROTEIN BINDING	117	94	3.5068	1.5049

Previous 1 2 Next

CD14+ Mono | SELECT

Geneset	Frequency	Size	Qvalue	OddRatio
SECRETORY GRANULE LUMEN	477	100	20	35.0065
FICOLIN-1-RICH GRANULE LUMEN	476	118	20	30.1316
TERTIARY GRANULE LUMEN	444	41	20	28.9028
ANTIGEN PROCESSING AND PRESENTATION OF EXOGENOUS PEPTIDE ANTIGEN VIA MHC CLASS II	442	27	20	17.4529
COLLAGEN-CONTAINING EXTRACELLULAR MATRIX	358	134	20	14.3061
ANTIGEN PROCESSING AND PRESENTATION	448	39	20	13.0168
POSITIVE REGULATION OF INFLAMMATORY RESPONSE	253	78	20	8.9201
MHC CLASS II PROTEIN COMPLEX	402	15	20	7.8069
ENDOCYTIC VESICLE MEMBRANE	391	45	20	7.2544
NEUTROPHIL CHEMOTAXIS	164	47	20	6.5747

Previous 1 2 3 4 5 ... 10 Next

DC | SELECT

Geneset	Frequency	Size	Qvalue	OddRatio
ER TO GOLGI TRANSPORT VESICLE MEMBRANE	31	51	7.7155	28.823
COPII-COATED ER TO GOLGI TRANSPORT VESICLE	5	34	5.7082	25.5263
ENDOCYTIC VESICLE MEMBRANE	30	45	10.5129	24.7688
PROTON TRANSMEMBRANE TRANSPORT	31	80	5.2115	18.3829
LYSOSOMAL MEMBRANE	27	294	12.7768	18.342
TRANSPORT VESICLE MEMBRANE	26	27	12.8556	17.1138
LYSOSOME	25	369	12.414	15.4234
CLATHRIN-COATED ENDOCYTIC VESICLE MEMBRANE	28	42	10.3106	15.3559
OXIDATIVE PHOSPHORYLATION	31	20	4.263	14.8654
RESPONSE TO ENDOPLASMIC RETICULUM STRESS	4	73	3.655	14.3516

Previous 1 2 3 4 5 ... 11 Next

User chosen Pathways

CELLULAR RESPONSE TO TUMOR NECROSIS FACTOR @NK

RRNA PROCESSING @Naive CD4 T

COPII-COATED ER TO GOLGI TRANSPORT VESICLE @DC

RESPIRASOME @Memory CD4 T

TERTIARY GRANULE LUMEN @FCGR3A+ Mono

TERTIARY GRANULE LUMEN @CD14+ Mono

| PLOT THE SELECTED PATHWAYS | CLEAR LIST | PLEASE SWITCH ANY ELEMENT'S POSITION TO ACTIVATE THE PLOT BUTTON |

NOTE: please switch any element's position to activate the plot button

After clicking the 'PLOT THE SELECTED PATHWAYS' button, the top cell enriched by the selected pathways will be shown in the Scatter & Bar session.

The corresponding legend will also be displayed. And it can be downloaded separately.

| Scale	Pathway	Group
High — Low	B CELL ACTIVATION	B
High — Low	POSITIVE REGULATION OF INFLAMMATORY RESPONSE	CD14+ Mono
High — Low	NATURAL KILLER CELL MEDIATED CYTOTOXICITY	CD8 T
High — Low	COPII-COATED ER TO GOLGI TRANSPORT VESICLE	DC
High — Low	SPECIFIC GRANULE MEMBRANE	FCGR3A+ Mono
High — Low	T CELL ACTIVATION	Memory CD4 T
High — Low	CELLULAR RESPONSE TO INTERLEUKIN-4	Naive CD4 T
High — Low	MONOCYTE CHEMOTAXIS	NK
High — Low	PLATELET FORMATION	Platelet

4. BiPlot & HeatMap Module

This section shows the BiPlot presentation of the correlation between categories and top significant pathways.

Biplot between top pathways and groups
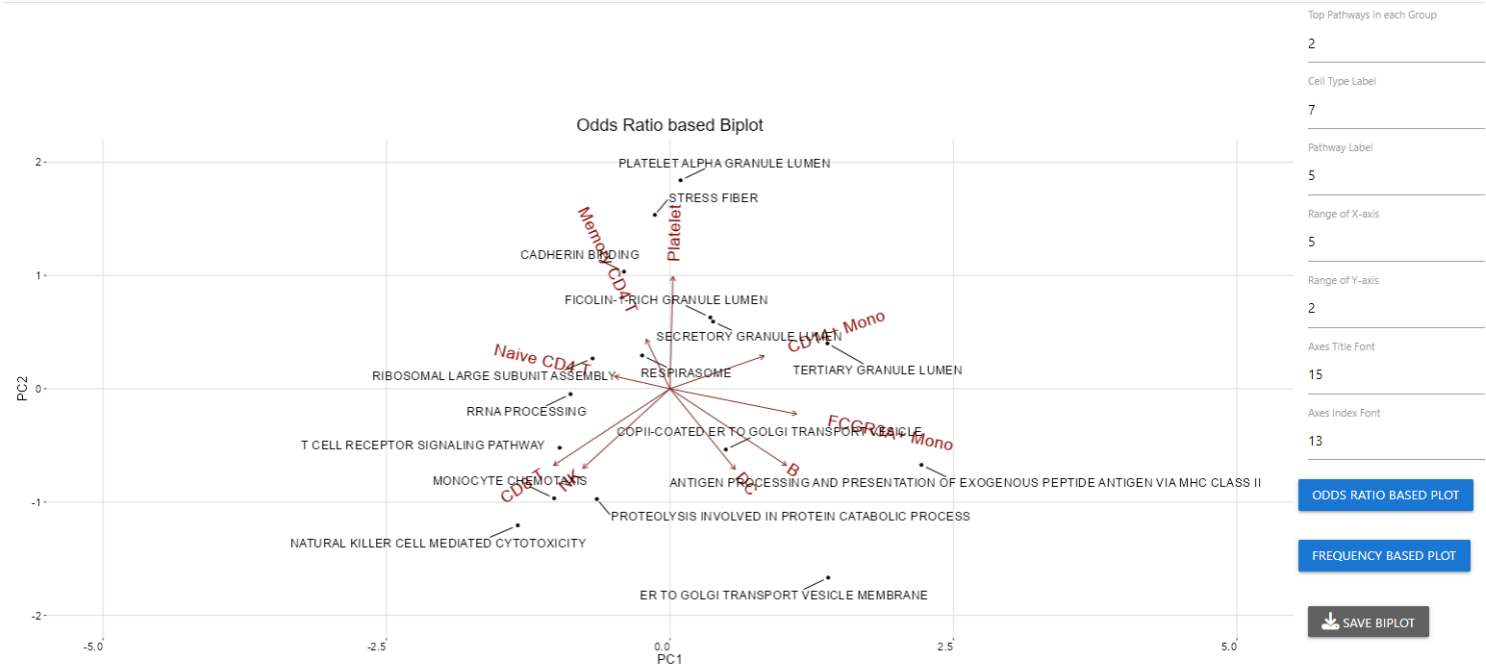
Users can check the BiPlot under the perspective of **FREQUENCY** or **ODDS RATIO.**
This will also update the corresponding HeatMap.
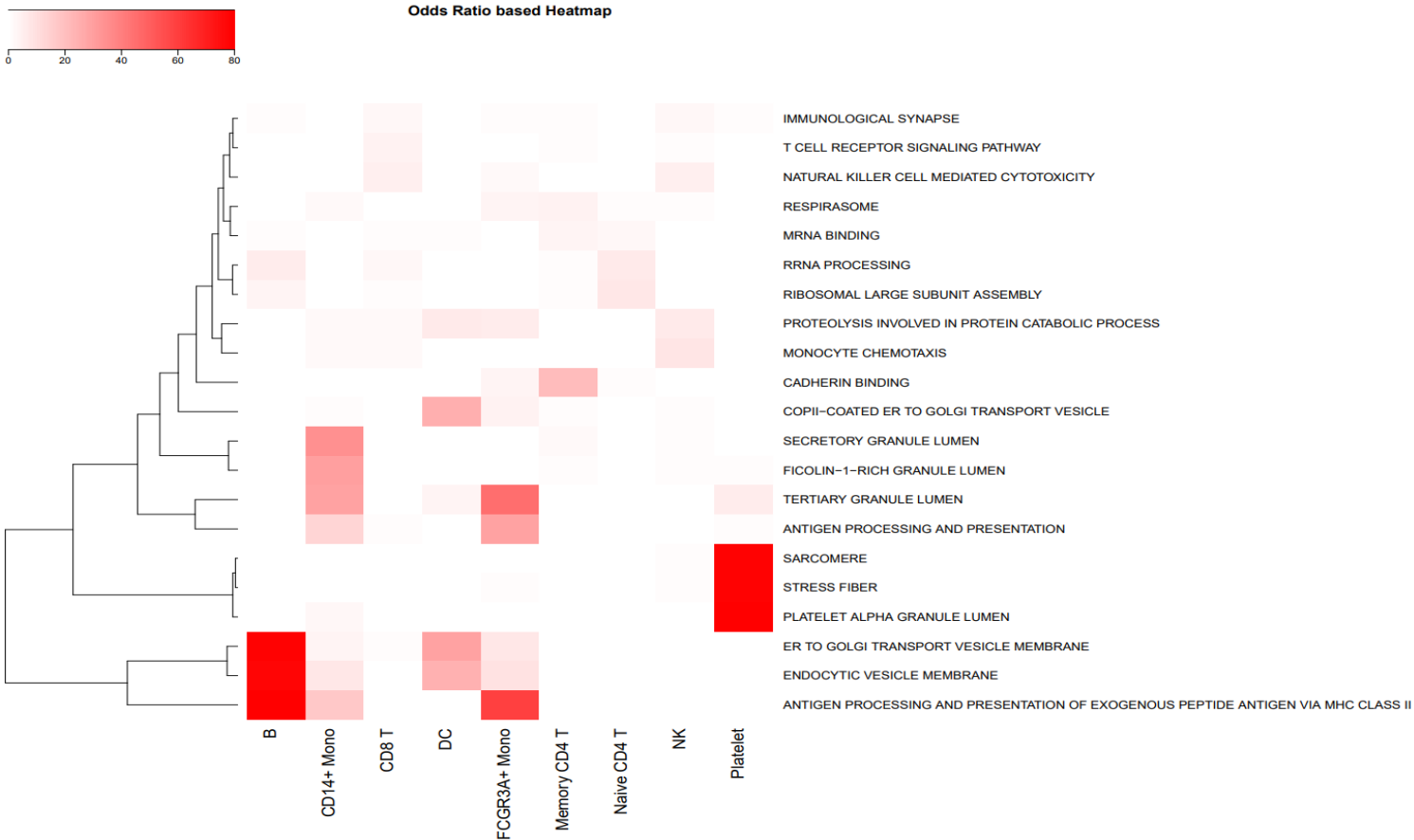
5. Marker Module

MarkerGenes

DE from each Cell specific

Group	Top	genes	FDR
All	All	All	All
1	1	CD3D	0
1	1	RPS27A	0
1	1	RPS27	0
1	3	LTB	0
1	5	RPL9	0
1	2	RPL21	0
1	3	RPS3A	0
1	2	LDHB	0
1	4	RPS25	0
1	2	CD3E	0

Previous 1 2 3 4 5 ... 36 Next

DE - Pathway from each Cell specific

genes	Count	Group	Top	FDR
All	All	All	All	All
ACTN1	1	0	4	0
CD3D	2	0	1	0
CD3E	2	0	2	0
CD3G	1	0	7	0
CD8B	3	0	3	0
IL7R	1	0	5	0
LCK	2	0	5	0
LDHB	1	0	2	0
LEF1	1	0	8	0
RPL10	1	0	9	0

Previous 1 2 3 4 5 ... 28 Next

This marker module will show Differentially Expressed genes in the following:

1. Each group (using findMarker in scran)
2. Each group and pathway-specific (using Fisher's exact test)

Developers

- Hai Nguyen (hainct@unist.ac.kr)
  Author, maintainer

- Jinhwan Kim
  Author

- Dougu Nam (dougnam@unist.ac.kr)
  Author, project manager

License

- MIT + file LICENSE