


# How to run gene set enrichment analysis on the results of DESeq2 of differential gene expression analysis with using Broads GSEA app (GSEAPreranked)

Simone Oberhänsli  
IBU Uni Bern  
2018-08-24

# Download



GenePattern GSEA Moderated estimation of f... IBU Pro

**GSEA**  
Gene Set Enrichment Analysis

GSEA Home Downloads Molecular Signatures Database

**Overview**

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether a predefined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes).

From this website you can:

- Download the GSEA software and additional resources to analyze, annotate and interpret enrichment results.
- Explore the Molecular Signatures Database (MSigDB), a collection of annotated gene sets for use with GSEA software.
- View documentation describing GSEA and MSigDB.

**What's New**

16-Jul-2018: MSigDB 6.2 released. This is a minor release that includes updates to gene set annotations, corrections to miscellaneous errors, and a handful of new gene sets. See the [release notes](#) for more information.

19-Oct-2017: MSigDB 6.1 released. See [release notes](#) for more information, including important corrections to gene sets in the C3 collection.

11-Aug-2017: Four new CHIP files are now available for use with data specified with Ensembl IDs, which are commonly used for gene expression derived from RNA-Seq data. More details are [here](#).

01-Jul-2017: The production version of GSEA Desktop v3.0 is now available! It's open-source on [GitHub](#), features SVG plots, Cytoscape 3.3+ support for Enrichment Maps, heatmap dataset export, and more.

06-Apr-2017: Version 6.0 of the Molecular Signatures Database (MSigDB) is now available under a Creative Commons license, with additional terms for some sub-collections of gene sets. The release also includes updates to the C3 motif gene sets, and some other minor additions and corrections. See the [Release Notes](#) for details.

Follow @GSEA\_MSigDB

**We're hiring!** We are looking for a curator to join the GSEA-MSigDB project in the Mesirov Lab at UC San Diego.

**Molecular Profiles**

Gene Set Data

**License Terms**

GSEA and MSigDB are released under the [Creative Commons Attribution-NonCommercial-ShareAlike license](#). Please register to download the MSigDB gene sets and to view the MSigDB gene sets. Please email [gsea@broadinstitute.org](#) for your email address. Please email [gsea@broadinstitute.org](#) for your email address. Please email [gsea@broadinstitute.org](#) for your email address.

**Contributors**

GSEA and MSigDB are released under the [Creative Commons Attribution-NonCommercial-ShareAlike license](#). Please register to download the MSigDB gene sets and to view the MSigDB gene sets. Please email [gsea@broadinstitute.org](#) for your email address. Please email [gsea@broadinstitute.org](#) for your email address. Please email [gsea@broadinstitute.org](#) for your email address.

**Citing GSEA**

To cite your use of GSEA, please cite the following paper: Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, D.L., Gillette, M.A., Paul, S., Omer, S.G., Zhang, L., Esposito, D., et al. (2005), PNAS 102, 8336-8341.

- log in here:  
<http://software.broadinstitute.org/gsea/login.jsp>
- click download
- choose memory option suitable for your computer

## Downloads

**We're hiring!** We are looking for a curator to join the GSEA-MSigDB project in the Mesirov Lab at UC San Diego.

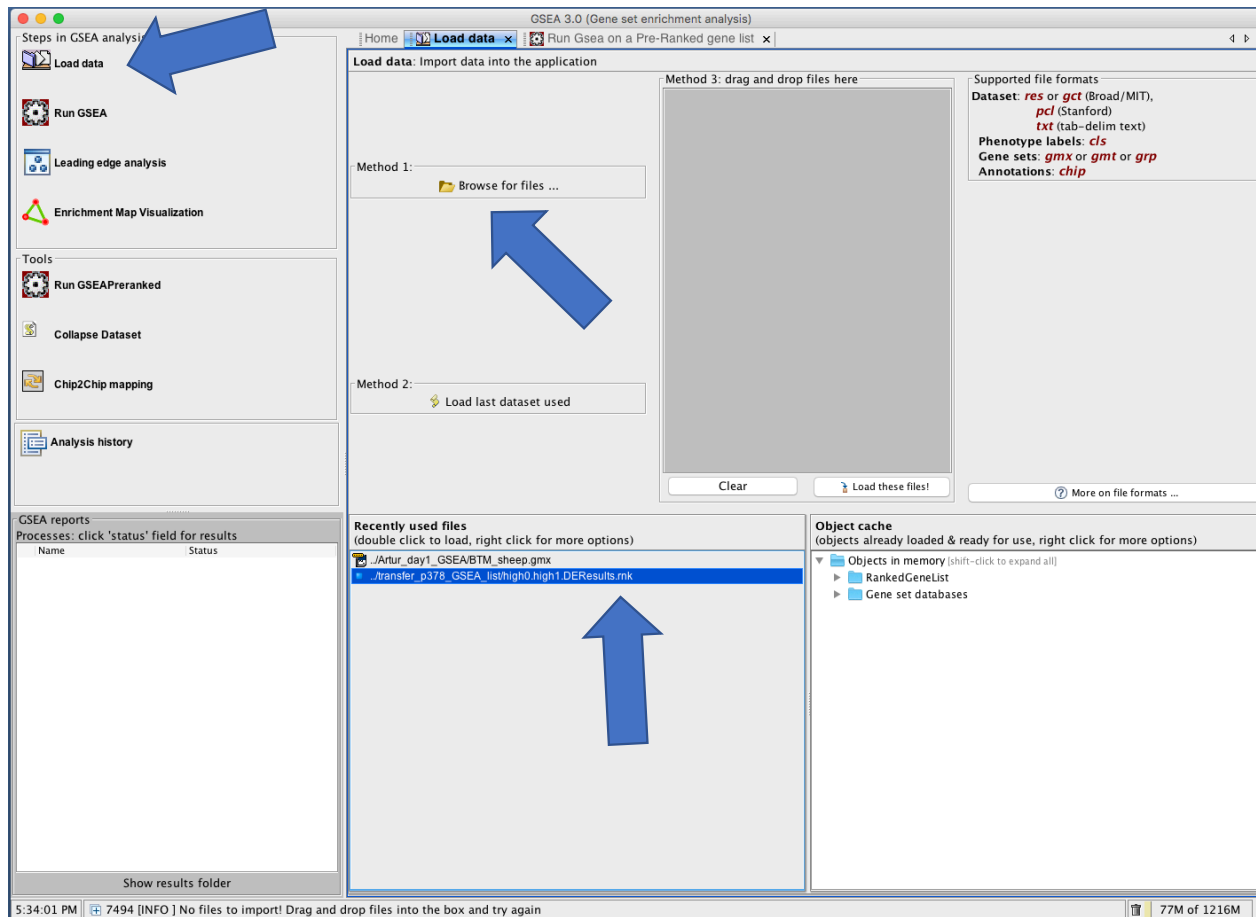
**Software**

There are several options for GSEA software. All options implement exactly the same algorithm. Usage recommendations and installation instructions are listed below. Current Java implementations of GSEA require Java 8.

See the [license terms](#) page for details about the license for the GSEA software and source code. Please note that the license terms vary for different versions of the software.

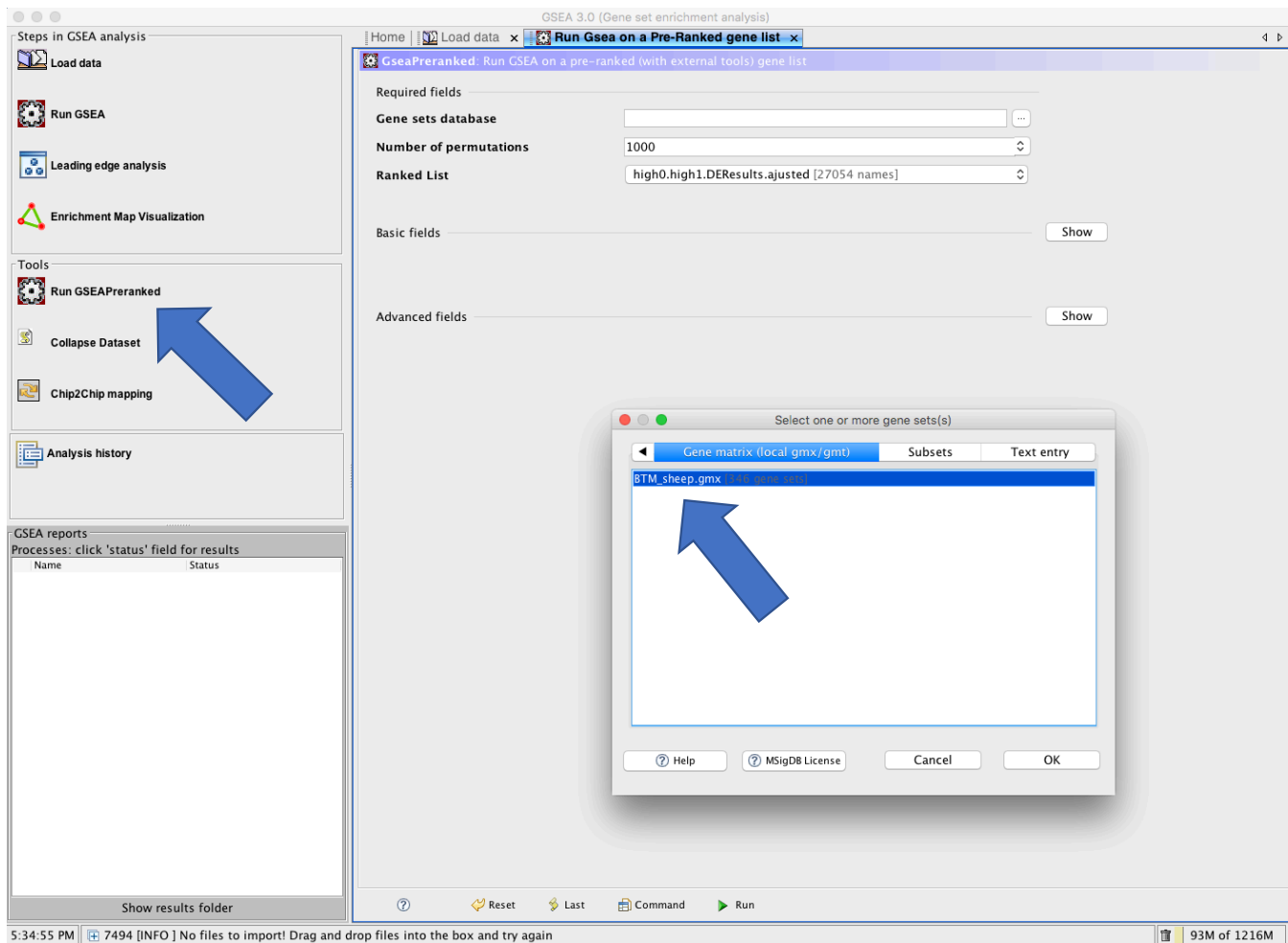
Software	Features	Launch with
<b>javaGSEA Desktop Application</b>	<ul style="list-style-type: none"><li>Easy-to-use graphical user interface.</li><li>Runs on any desktop computer (Windows, macOS, Linux etc.) that supports Java 8. <b>Oracle Java is recommended as there are known issues when running with OpenJDK. Java 9 and higher are not supported at this time.</b></li><li>Produces richly annotated reports of enrichment results.</li><li>This release is open source under a <a href="#">BSD-style license</a>. The source is available on our <a href="#">GitHub repository</a>. The changes are noted in the <a href="#">Release Notes</a>.</li><li>We recommend using a memory configuration smaller than your computer's total memory.</li></ul>	Launch with 1GB (for 32 or 64-bit Java) memory: <a href="#">Launch</a>
<b>javaGSEA Java Jar file</b>	<ul style="list-style-type: none"><li>Command line or offline usage. See our <a href="#">User Guide</a> for details.</li><li>Runs on any platform that supports Java 8. <b>Oracle Java is recommended as there are known issues when running with OpenJDK. Java 9 and higher are not supported at this time.</b></li><li>We recommend using the 'Launch' buttons above instead of this mode for most users.</li></ul>	<a href="#">download gsea-3.0.jar</a>

# 1. Load data

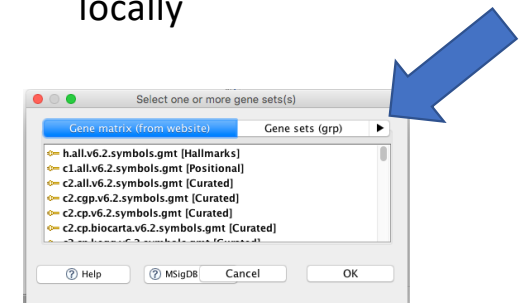


- click “load data”
- import your files using methods 1,2 or 3. These files are:
  - ranked gene list (required)
  - custom gene set (optional)
- a message will tell you if upload was successful. It should say “there were NO errors”.
- you can now see the uploaded files in the bottom left field

## 2. Choose gene set



- click on “Run GSEAPreranked”
- click on the button with the three dots right to “Gene sets database”
- choose the gene set you would like to use
- if you want to use a custom gene set (loaded in step 1) click on the arrow to get to “Gen matrix (local)” where you can choose the gene sets which have been loaded locally



### 3. Run analysis

GSEA 3.0 (Gene set enrichment analysis)

Steps in GSEA analysis

- Load data
- Run GSEA
- Leading edge analysis
- Enrichment Map Visualization

Tools

- Run GSEAPreranked
- Collapse Dataset
- Chip2Chip mapping
- Analysis history

GSEA reports

Processes: click 'status' field for results

Name	Status
------	--------

Show results folder

Home | Load data | **Run GSEA on a Pre-Ranked gene list**

Gseapreranked: Run GSEA on a pre-ranked (with external tools) gene list

Required fields

Gene sets database: 78\_Artur\_Sheep\_RNA\_seq\_vac/Artur\_day1\_GSEA/BTM\_sheep.gmx

Number of permutations: 1000

Ranked List: high0.high1.DEResults.adjusted [27054 names]

Basic fields

Analysis name: some\_descriptive\_title

Enrichment statistic: classic

Max size: exclude larger sets: 500

Min size: exclude smaller sets: 15

Save results in this folder: /Users/simone/Projects

Advanced fields

Normalization mode: meandiv

Alternate delimiter:

Create SVG plot images: false

Make detailed gene set report: true

Plot graphs for the top sets of each phenotype: 20

Seed for permutation: timestamp

Make a zipped file with all reports: false

Reset Last Command Run

5:36:21 PM 7494 [INFO] No files to import! Drag and drop files into the box and try again 128M of 1216M

- blue arrows: adjust the parameters and fill out text fields
- adjust additional parameters if you wish to. For more information check the [GSEA user guide](#)

## 4. Successful run

The screenshot displays the GSEA 3.0 (Gene set enrichment analysis) software interface. The main window is titled "GSEA 3.0 (Gene set enrichment analysis)" and shows a tab for "Run GSEA on a Pre-Ranked gene list". The interface is divided into several sections:

- Steps in GSEA analysis:** A sidebar on the left lists the steps: Load data, Run GSEA, Leading edge analysis, and Enrichment Map Visualization.
- Tools:** A sidebar on the left lists the tools: Run GSEAPreranked, Collapse Dataset, and Chip2Chip mapping.
- Analysis history:** A sidebar on the left shows the history of analyses.
- GSEA reports:** A table at the bottom left shows the status of the analysis. A blue arrow points to the "Success" status.

The main panel displays the configuration for the "Run GSEA on a Pre-Ranked gene list" analysis. The configuration is organized into sections:

- Required fields:**
  - Gene sets database: 78\_Artur\_Sheep\_RNA\_seq\_vac/Artur\_day1\_GSEA/BTM\_sheep.gmx
  - Number of permutations: 1000
  - Ranked List: high0.high1.DEResults.adjusted [27054 names]
- Basic fields:**
  - Analysis name: some\_descriptive\_title
  - Enrichment statistic: classic
  - Max size: exclude larger sets: 500
  - Min size: exclude smaller sets: 15
  - Save results in this folder: /Users/simone/Projects
- Advanced fields:**
  - Normalization mode: meandiv
  - Alternate delimiter:
  - Create SVG plot images: false
  - Make detailed gene set report: true
  - Plot graphs for the top sets of each phenotype: 20
  - Seed for permutation: timestamp
  - Make a zipped file with all reports: false

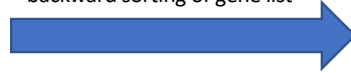
The status bar at the bottom indicates the time is 5:37:35 PM, the command is 0206 [INFO] Parsed from unigene / gene symbol: 38870, and the memory usage is 357M of 3148M.

# How to create a ranked gene list for GSEA

	pvalue	adjusted pvalue
gene 1	...	...
gene 2	...	...
gene 3	...	...
...	...	...
gene 1000	...	...

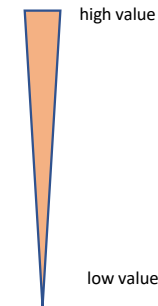
Results of differential gene expression analysis, e.g. with DESeq2. In this example the list is sorted according to gene name

backward sorting of gene list



	ranking metric
gene 3	...
gene 1	...
gene 1000	...
...	...
gene 2	...

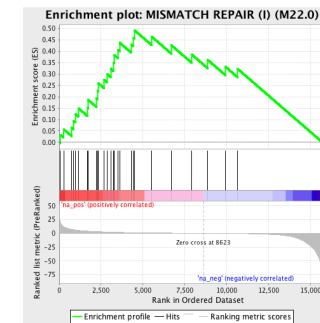
Ranked gene list according to ranking metric. Ranking metric can be e.g. pvalue, a score of a statistical test or log2Foldchange



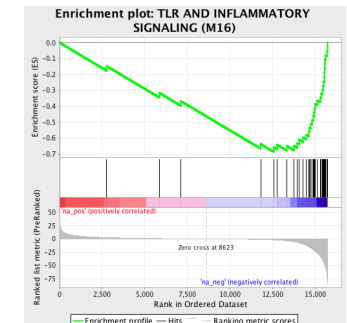
use this list as input for GSEAPreranked



important: choose your ranking metric depending on your research question. When analyzing results of DESeq2 it is handy to arrange the genes in the ranked list so that the significantly differentially expressed genes are located at the tails of the list, e.g. the sign. differentially expressed genes with a positive foldchange at the top and sign. differentially expressed genes with a negative fold change at the bottom of the list. Arranging genes in this way will make the interpretation of GSEA results easier. A possible metric to achieve this would be log2foldchange or a signed pvalue (see next slide)



enriched in differentially expressed genes with positive fold change



enriched in differentially expressed genes with negative fold change

# How to create a ranked gene list for GSEA (cont.)

In our view a suitable ranking metric for DESeq2 results is a “signed” pvalue or “signed” adjusted pvalue, which means that we add a sign (+ or -) to the pvalue in order to indicate the direction of fold change (positive or negative). For technical reasons, GSEA requires backward sorting. Therefore, we need to convert the pvalues so that the resulting “signed” pvalue is large for small pvalues and small for large pvalue (see below for example).

	Log2Foldchange	signFC	pvalue	-log(pvalue)*signFC
gene 3	positive	1	0.001	6.907755
gene 1	positive	1	0.002	6.214608
gene 1000	positive	1	0.009	4.710531
...				...
gene 2	negative	-1	0.001	-6.907755

indicates direction of fold change

the lower the more significant

the higher the more significant

	-log(pvalue)*signFC
gene 3	6.907755
gene 1	6.214608
gene 1000	4.710531
...	...
gene 2	-6.907755

Please note: When using adjusted pvalue as ranking metric there is a possibility of duplicate ranking (two genes have the same rank). This can happen if two (in rare cases three) genes have a very similar pvalue, which serves as bases for calculating the adjusted pvalue. GSEA does not resolve ties. In the case of a tie, the order of genes will be arbitrary. However, if two genes have almost the same pvalue, the order of how they appear in the ranking list is probably not that crucial. If you want to avoid the problem of duplicate ranking you can use the pvalue instead of the adjusted pvalue as ranking metric. The number of genes in the ranked gene list is most probably smaller than the DESeq2 result list. The reason is that DESeq2 applies a filter to genes with zero or very low expression and outliers. These genes will not have a pvalue and/or an adjusted pvalue (indicated with “NA” in the DESeq2 result list). For more information see [DESeq2 vignette](#).



# Important!

- the pre-ranked list must have the ending .rnk (NOT .txt)
- the pre-ranked list has to be sorted
- the pre-ranked list should have column headers
- the pre-ranked list can only consist of two columns