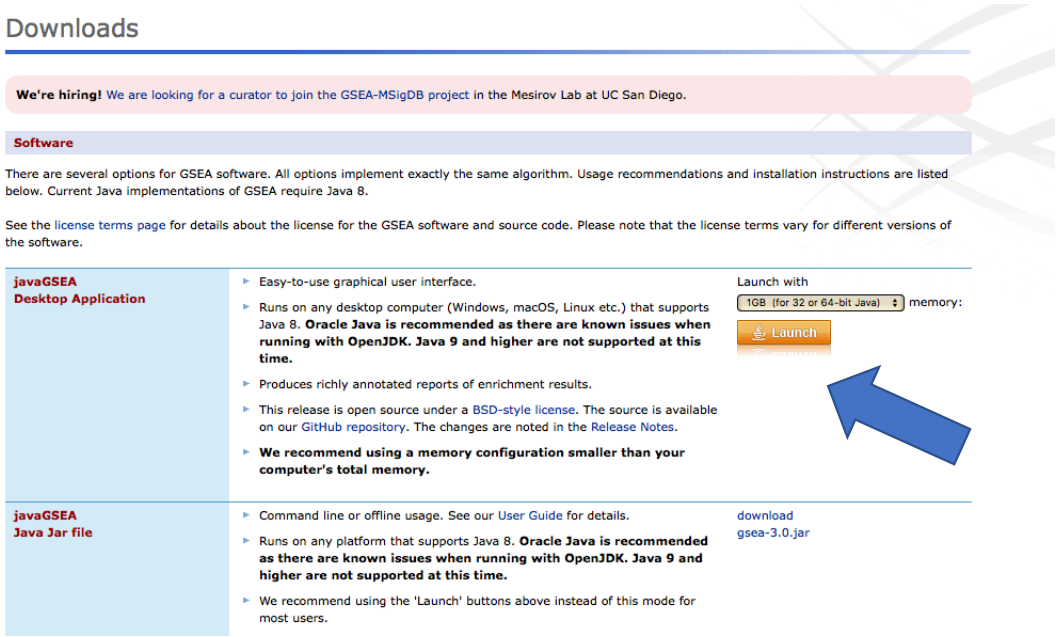
How to run gene set enrichment analysis on the results of DESeq2 of differential gene expression analysis with using Broads GSEA app (GSEAPreranked)

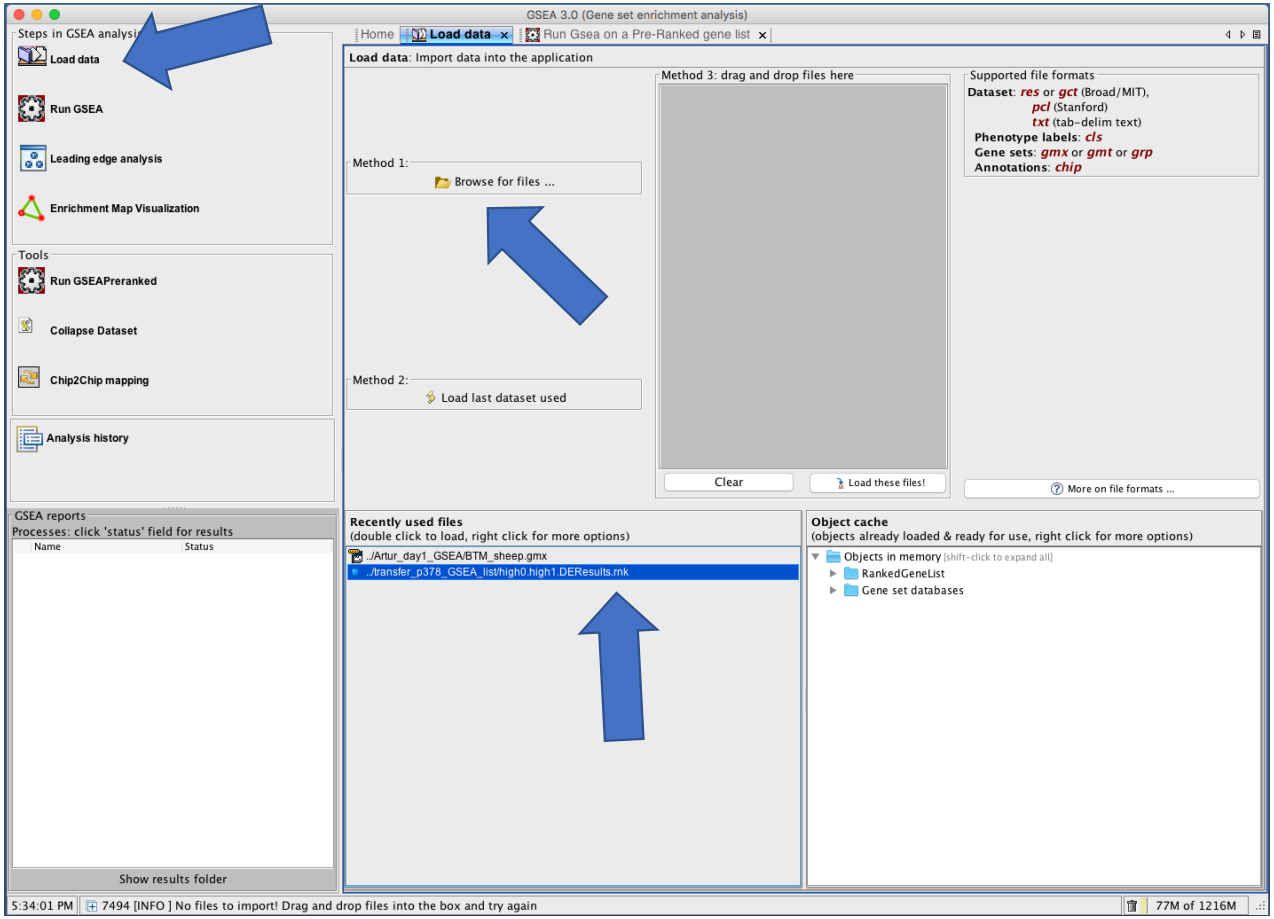
Simone Oberhänsli
IBU Uni Bern
2018-08-24

Download

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

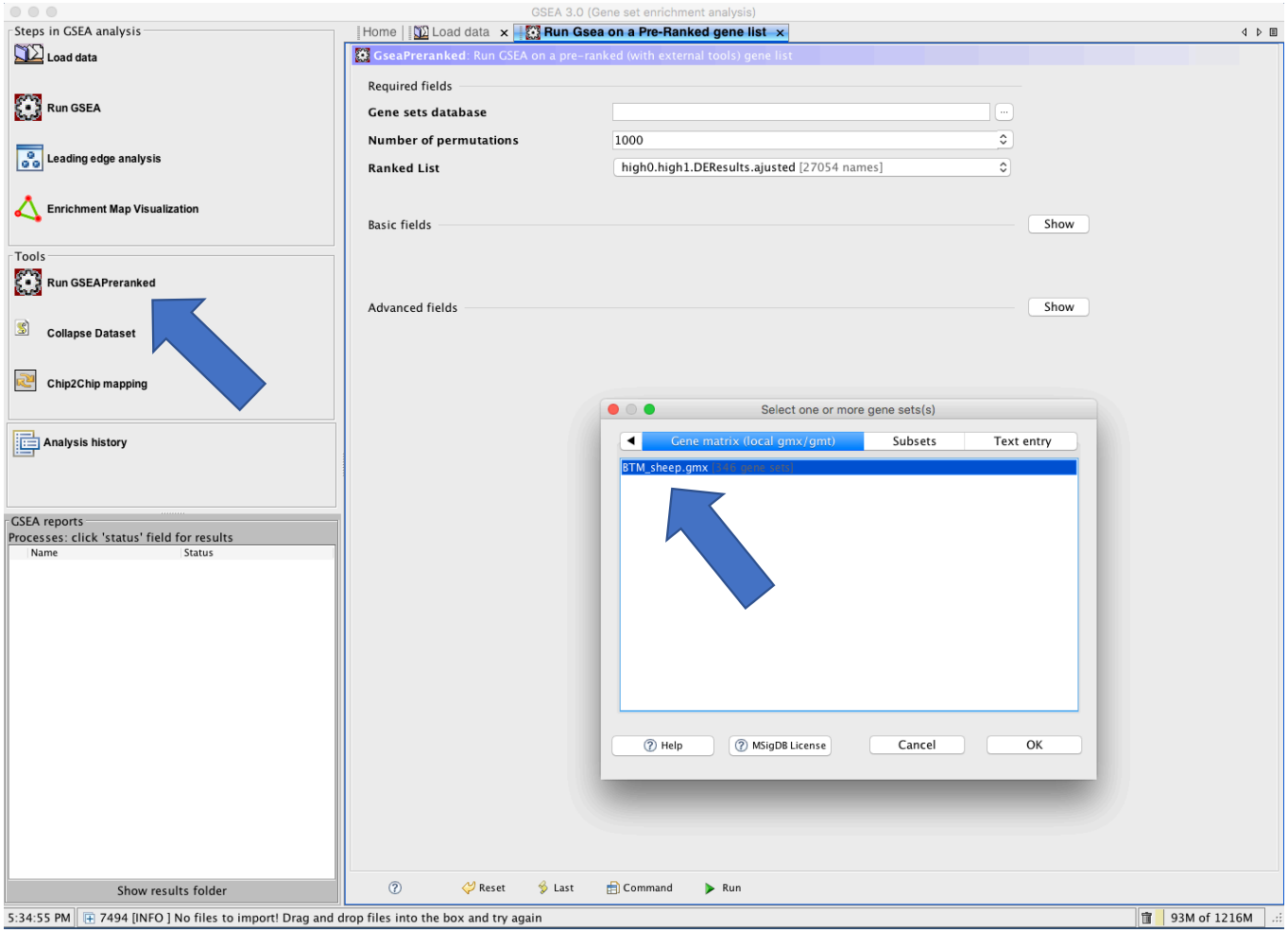


1. Load data

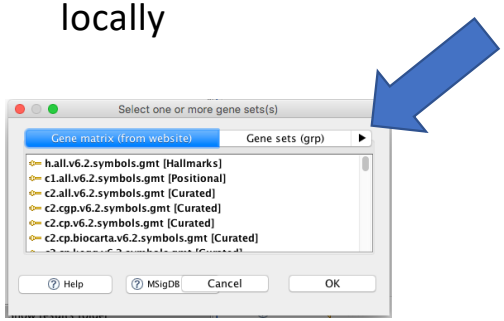


- click “load data”
- import your files using methods 1,2 or 3. These files are:
  - ranked gene list (required)
  - custom gene set (optional)
- a message will tell you if upload was successful. It should say “there were NO errors”.
- you can now see the uploaded files in the bottom left field

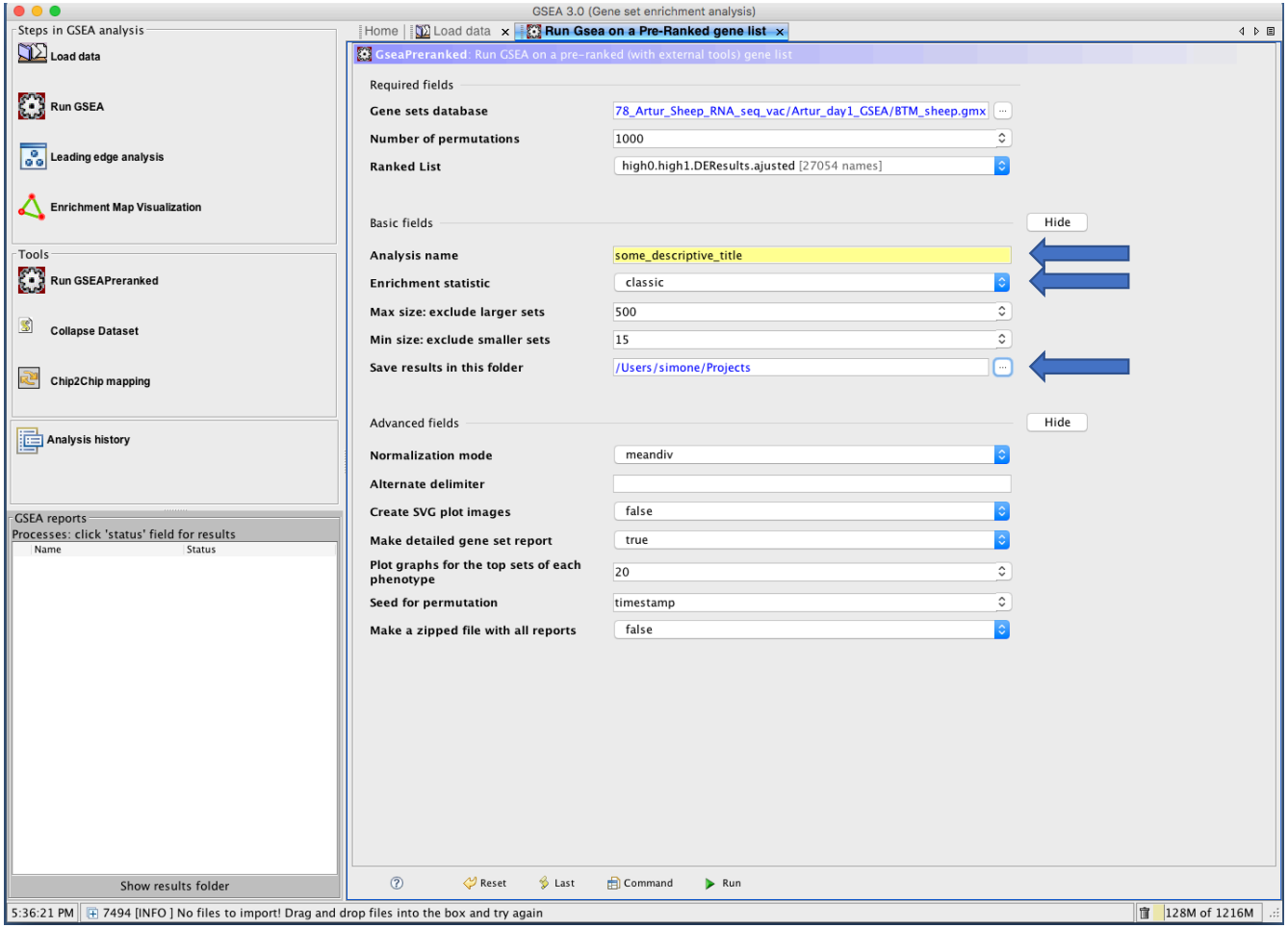
2. Choose gene set



- 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

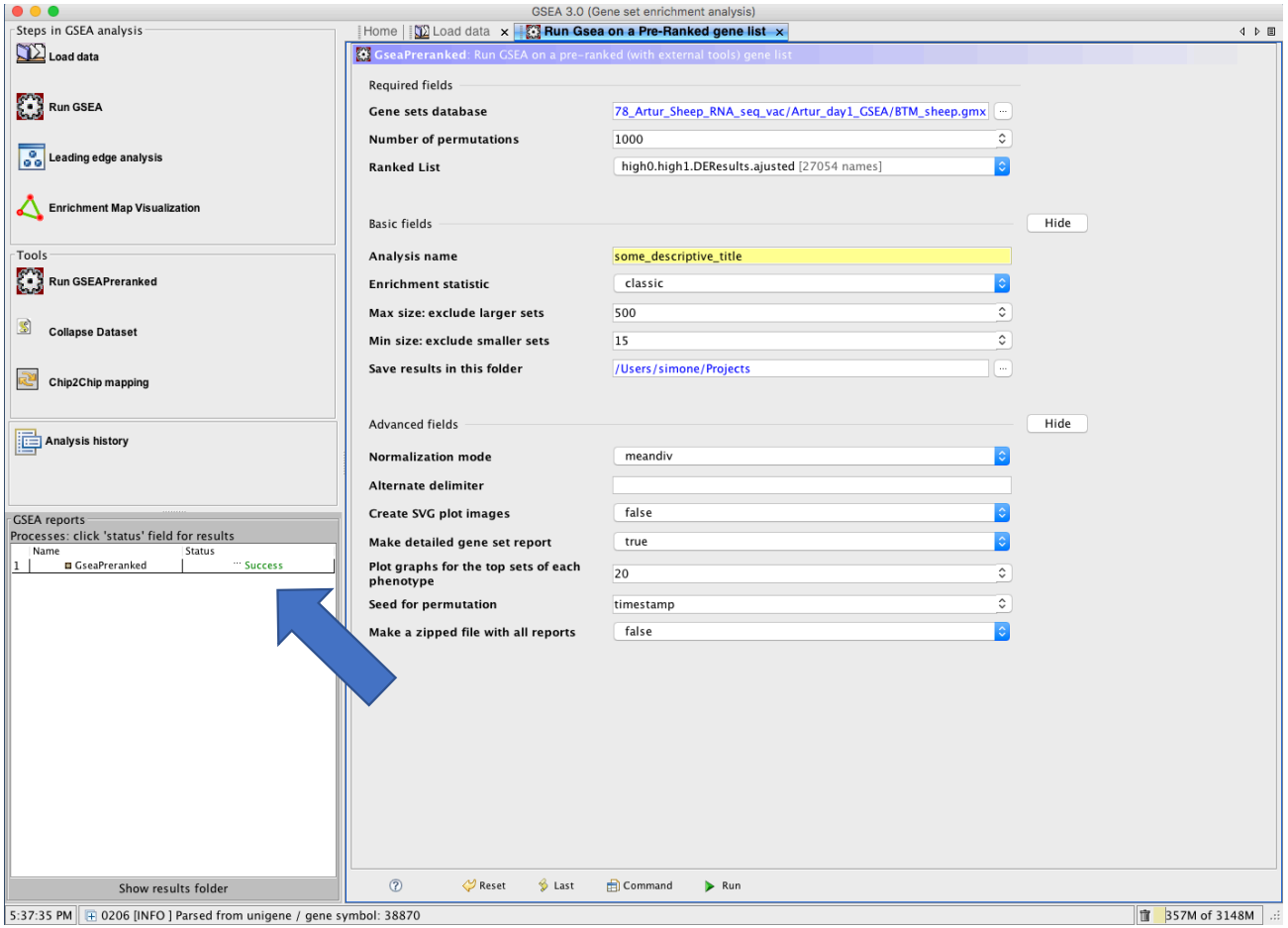


3. Run analysis



- 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



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

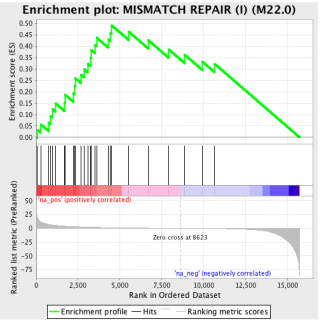
	ranking metric
gene 3	...
gene 1	...
gene 1000	...
...	...
gene 2	...

high value

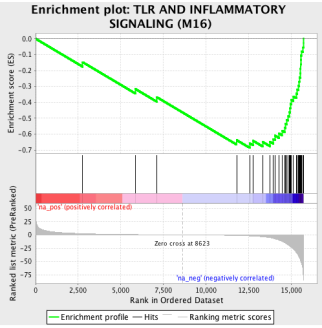
low value

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



enriched in differentially expressed genes with positive fold change



enriched in differentially expressed genes with negative fold change

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)

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).

indicates direction of fold change

the lower the more significant

the higher the more significant

	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

	-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