

GSEA hands-on

by Ester Feldmesser, March 2016

The data for this exercise was taken from the paper by Umansky, Groner et al called “Runx1 Transcription Factor Is Required for Myoblasts Proliferation during Muscle Regeneration” (<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005457#abstract1>). In this paper the function of Runx1 in muscle regeneration is investigated. One of the analysis performed was RNA-seq comparing mice lacking dystrophin and muscle Runx1 (*mdx/Runx1^{ff}*) to mice lacking dystrophin and having muscle Runx1 (*mdx/Runx1^{L/L}*). The mdx mice are a model of Duchenne muscular dystrophy. In contrast to human, the mdx mice model (with wt Runx1) regenerates constantly the muscle.

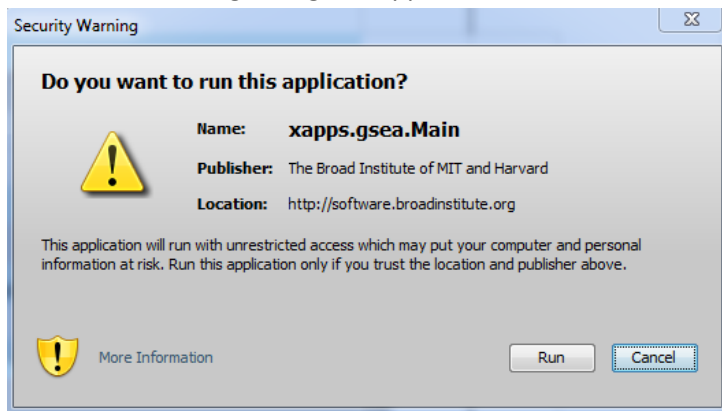
The RNA-seq data was analyzed and an input file was prepared. The input file is called “MdxVsMdxKO_Capital.rnk” (the KO is for Runx1) and it is located in your computer under ngs001\course2016\Functional Analysis. This is a text file and can be opened with Excel. The suffix in the name is part of the file format required by GSEA for the analysis we are performing. The file includes all the genes that had more than 10 reads in at least one sample and it is sorted according to their fold changes. There are two columns in the file: the gene name and the fold change. Note that the gene name is in capitals. The GSEA software was developed for human data and in this way we convert the mouse gene names into human.

Our workflow:

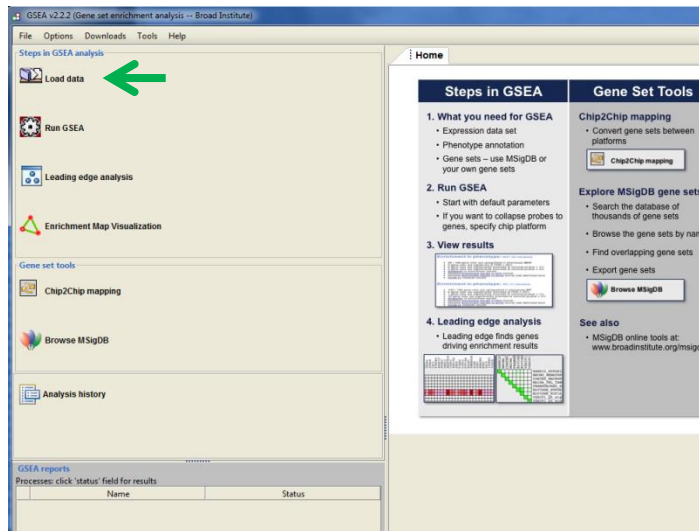
1. Open GSEA by double clicking in the icon at the desktop:



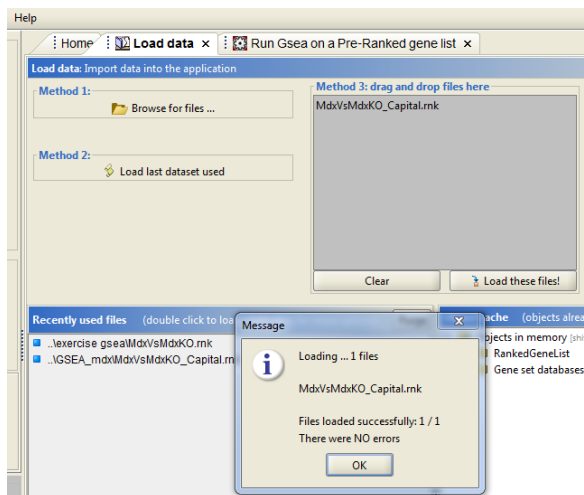
2. When the following dialog box appears, click on Run.



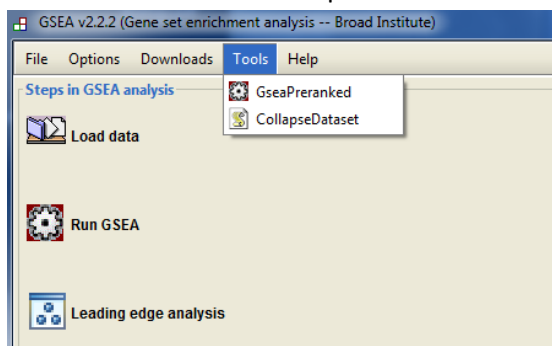
3. GSEA will open. Click on Load data (green arrow).



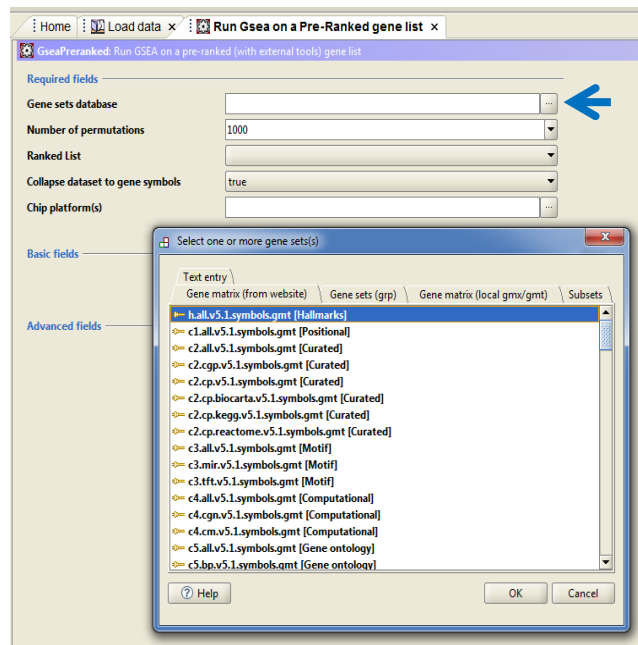
4. The easiest way to load the file is to drag it into the gray box, then click on Load these files! A message indicating that the file was successfully loaded should appear. Click OK.



5. Go to Tools menu at the top bar and choose GseaPreranked.



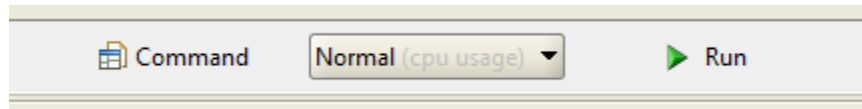
6. We will start filling the form with the Gene sets database. Click on the three points (blue arrow), a dialog box will appear, chose the first one (Hallmarks) and click OK.



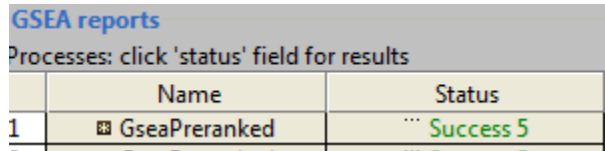
7. Click on the two “Show” buttons. Be sure that all the parameters are defined as shown here. Change the analysis name and folder for saving the results as you wish.

The screenshot shows the 'GSEAPreRanked: Run GSEA on a pre-ranked (with external tools) gene list' window. The 'Required fields' section is filled with the following values: 'Gene sets database' is 'eaftp.broadinstitute.org/pub/gsea/gene_sets/h.all.v5.1.symbols.gmt', 'Number of permutations' is '1000', 'Ranked List' is 'MdxVsMdxKO_Capital [16165 names]', 'Collapse dataset to gene symbols' is 'false', and 'Chip platform(s)' is empty. The 'Basic fields' section is expanded, showing: 'Analysis name' is 'mdx_h', 'Enrichment statistic' is 'weighted', 'Max size: exclude larger sets' is '500', 'Min size: exclude smaller sets' is '15', and 'Save results in this folder' is 'D:\course2016\exercise gsea'. The 'Advanced fields' section is also expanded, showing: 'Collapsing mode for probe sets => 1 gene' is 'Max_probe', 'Normalization mode' is 'meandiv', 'Omit features with no symbol match' is 'true', 'Make detailed gene set report' is 'true', 'Plot graphs for the top sets of each phenotype' is '20', 'Seed for permutation' is 'timestamp', and 'Make a zipped file with all reports' is 'false'. 'Hide' buttons are present next to the 'Basic fields' and 'Advanced fields' sections.

8. On the lower part of the window change the Low (cpu usage) to Normal and click on Run.



9. Look at the lower right panel.

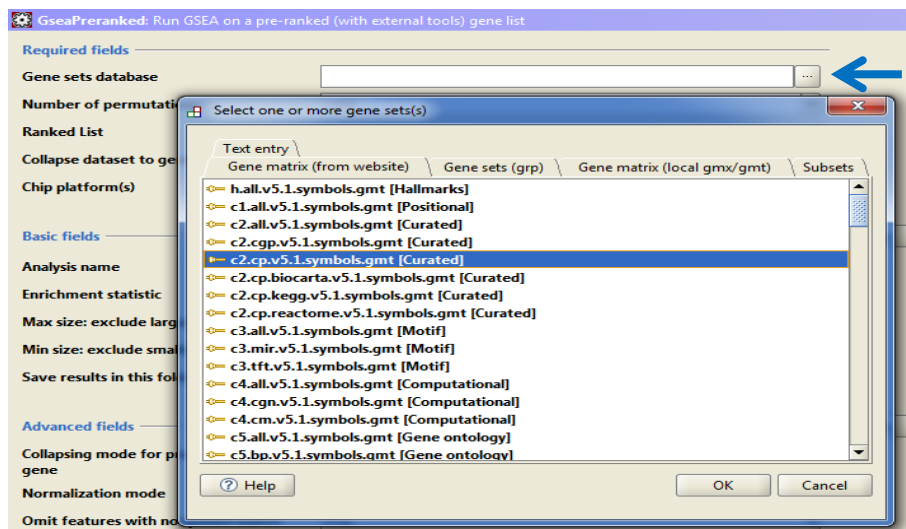


GSEA reports

Processes: click 'status' field for results

	Name	Status
1	GseaPreranked	Success 5

10. When the analysis finishes, click on the word Success. A web page will open showing the results. An additional way to get to the results (after you close the application) is by double clicking on the file called index.html inside the gsea results folder.
11. In the newly opened web page of the results the title **Enrichment in phenotype : na appear twice**. The first one is for the results of the upregulated genes and the second one for the down regulated genes. First take a look at the Guide to interpret results.
12. Open the Snapshot link to see the enrichment score graphs of the most significant gene sets. Some of them look very convincing and some of them are noisy. Not all of them are significant.
13. To see more details, return to the results page and click on “Detailed enrichment results in html format” link. You will see a table with the results. Compare the FDRs to the snapshots that you saw before. To see the details of the analysis and the genes that contribute to the score go to Details... To see the gene set definition, click on the name. You may be requested to register. Do it. Let’s look what are the genes included in the gene set HALLMARK_MYOGENESIS.
14. Do you think the enriched gene sets are related to the biology? Find one that makes sense and explain how it is related to the experiment.
15. We will perform an additional analysis in GSEA to get more detailed information. This time we will choose to run our genes against canonical pathways from several databases (the Gene set that starts with c2.cp.v5.1...). Remember to change the name of your analysis.



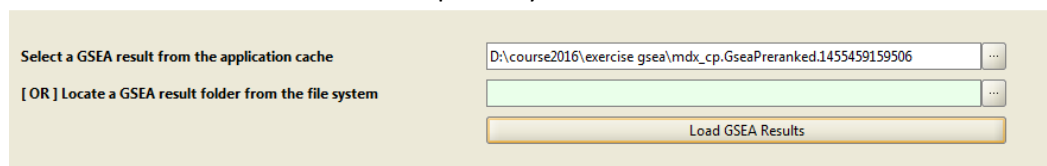
16. This analysis will take longer than before because there are more gene sets in the canonical pathways. Look at the table of results when the analysis finishes.

17. Since there are much more significant gene sets in this analysis, it is difficult to look at long tables and there is overlap between the pathways, we will look at the graphic summaries of the results using two different tools.

18. Return to the GSEA software and choose the Leading edge analysis (red arrow).

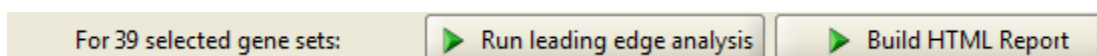


19. Select the folder with the canonical pathways result and click on Load GSEA Results.



20. Select the rows of the Gene Sets in the table with FDR less than 0.03 and click on Run leading edge analysis.

We select a number of gene sets that enables visualization. You could choose a different cutoff or use other criteria.

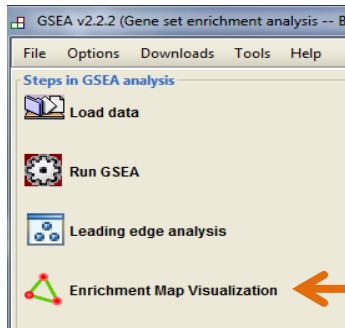


To understand the results go to:

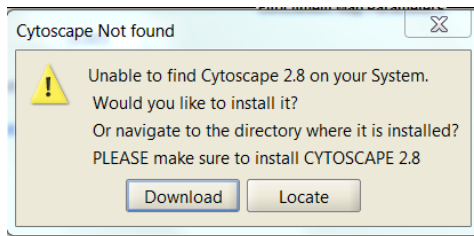
[http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html? Interpreting GSEA Results](http://software.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html?_Interpreting_GSEA_Results). Look for [Interpreting Leading Edge Analysis Results](#) in the left panel.

Back to the results, look at top right graph and find a function that is supported by more than one term or pathway.

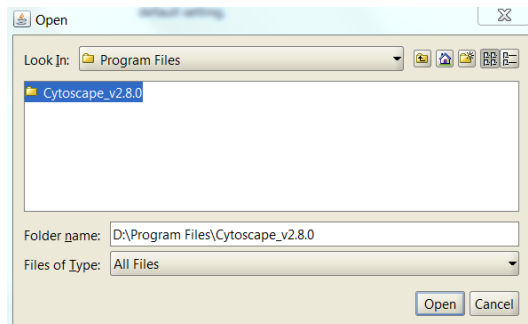
21. A different way of visualizing the data is by using networks. GSEA has a plug-in for the Cytoscape software. Cytoscape is an open source software platform for visualizing complex networks and integrating these with any type of attribute data, including expression data (<http://www.cytoscape.org/>). In the left panel, click on Enrichment Map Visualization (orange arrow).



22. Choose the same results file as before and load the file. We will leave the default options as is and click on Build Enrichment Map. If GSEA does not find the Cytoscape software, it will ask to locate it.



23. Click on locate and go to D:\Program files\Cytoscape_v2.8.0



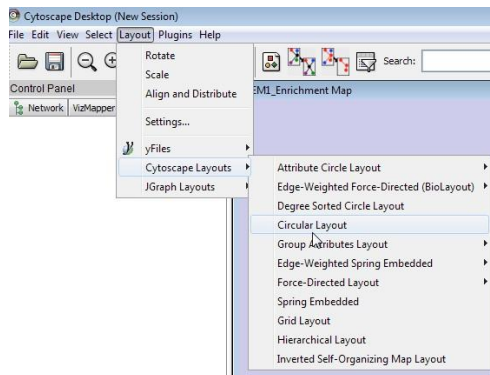
24. Cytoscape will open and will show a network of the pathways according to the gene overlap between them.

Node size corresponds to the number of genes within the gene set. Color of the node corresponds to positive or negative enrichment score. Edge width corresponds to the number of genes that overlap between the two connected gene sets.

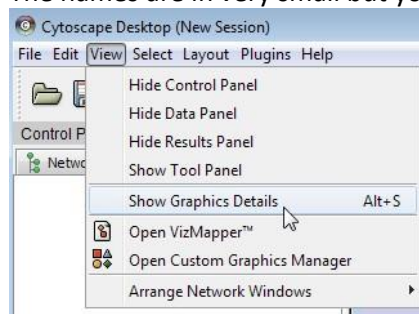
Right click on a node shows its name and left click on the node shows the list of genes in this pathway on the Data Panel. Left clicking on an edge shows the genes overlapping between the two pathways

25. To see the genes that contributed to the positive score in the GSEA analysis: Click on a node (gene set) in the Enrichment map. In the Data Panel, the fold changes (Rank/Score) of all genes included in the selected gene-set should appear. If they do not, click on Show values. Change the Sorting method to GSEARanking. Genes that are part of the leading edge are highlighted in yellow.

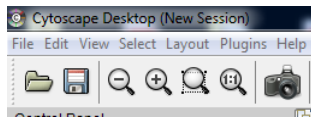
26. The layout of the network can be changed to display the graph as circular. Open the Layout menu and chose Cytoscape layouts and then Circular layout.



27. To display the pathway names open the View menu and chose Show Graphics details. The names are in very small but you can zoom in to see them.



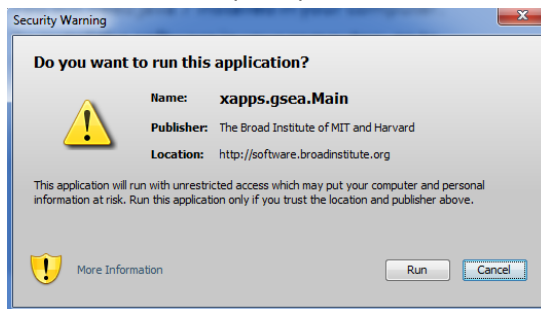
28. You can save the network image by clicking on the camera at the top bar



Installation of GSEA and Cytoscape

In case you want to install the software in your computer for further use.

1. You will need java 7 installed in your computer.
2. To install the GSEA software in your computer, go to <http://software.broadinstitute.org/gsea/downloads.jsp>, click on the orange button Launch to install the software. Save the gsea.jnlp file and double click on it. You may be requested to register during the process.
3. Click on Run when prompted.



4. To install Cytoscape, go to http://www.cytoscape.org/download_old_versions.html and download the 2.80 version. Click on Cytoscape_2_8_0_windows_64bit.exe. Save the file and run the Setup Wizard, or run it without saving. Open the software to check that the installation went OK.