GSEA tutorial

You can choose to do these exercises using the questions as your only guide - or see the following pages for the step-by-step checklist to finding these answers.

The data set used for this practical lab contains transcriptomics data obtained from MCF7 cells, a human breast cancer line, treated or non treated with estradiol. The cells were treated with estradiol for 12, 24 or 48 hours. Total RNA extracted from the cells was amplified, labeled and hybridized to Affymetrix GeneChip U133 Plus 2.0 microarrays. The data are available in the Gene Expression Omnibus (GEO) repository under the accession number GSE11352 (PMID: 17542648). The practical lab contains two exercises. Exercise 1 uses GSEA (http://www.broadinstitute.org/gsea/index.jsp) to perform gene-set enrichment analysis and exercise 2 uses g:Profiler (http://biit.cs.ut.ee/gprofiler/).

For this exercise, our goal is to upload the 3 required files into GSEA, set up the parameters, run GSEA, open and explore the gene-set enrichment results. We use as input file for GSEA the normalized data for all samples included in the GSE11352 dataset and formatted as a '**.gct'** file. GSEA will assess the amplitude of differential gene expression levels between the two groups of interest, in this case the treated samples and non treated samples at 12 hours using a t-test for each gene. The '**.cls'** file tells GSEA which samples correspond to our groups of interest. GSEA ranks the genes based on t values from the t-test and performs the gene-set enrichment analysis using a modified Kolmogorov-Smirnov statistics. The output result folder contains several files, and two of them are the summary tables displaying enrichment statistics for each gene-set (pathway) that has been tested and contained in the provided '**.gmt'** file. The **'.gmt**' file (gene-set file) provided for this exercise contains gene-sets obtained from KEGG, MsigDB-c2, NCI, Biocarta, IOB, Netpath, HumanCyc, Reactome and the Gene Ontology (GO) databases. (http://baderlab.org/GeneSets).

Before starting this exercise, launch GSEA using the instructions provided on the wiki and download the 3 required files:

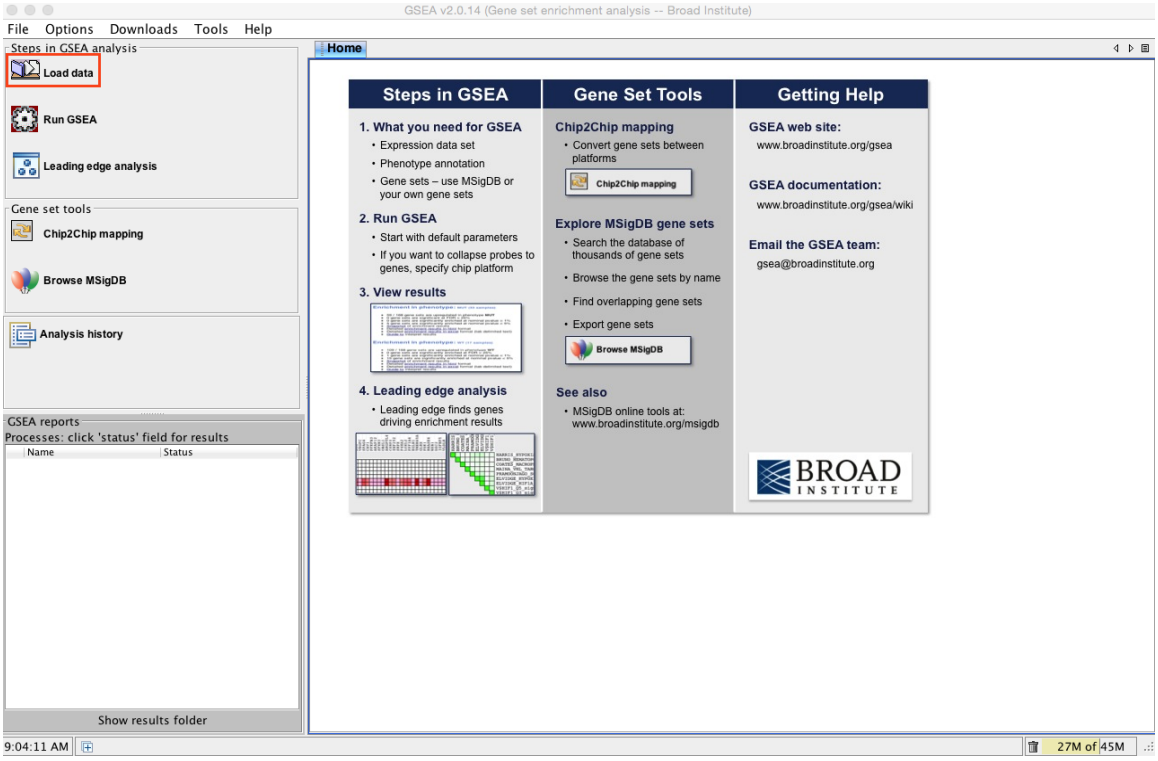
- MCF7_Expression_matrix.gct
- MCF7_groups.cls
- Human_GO_AllPathways.gmt

STEP BY STEP:

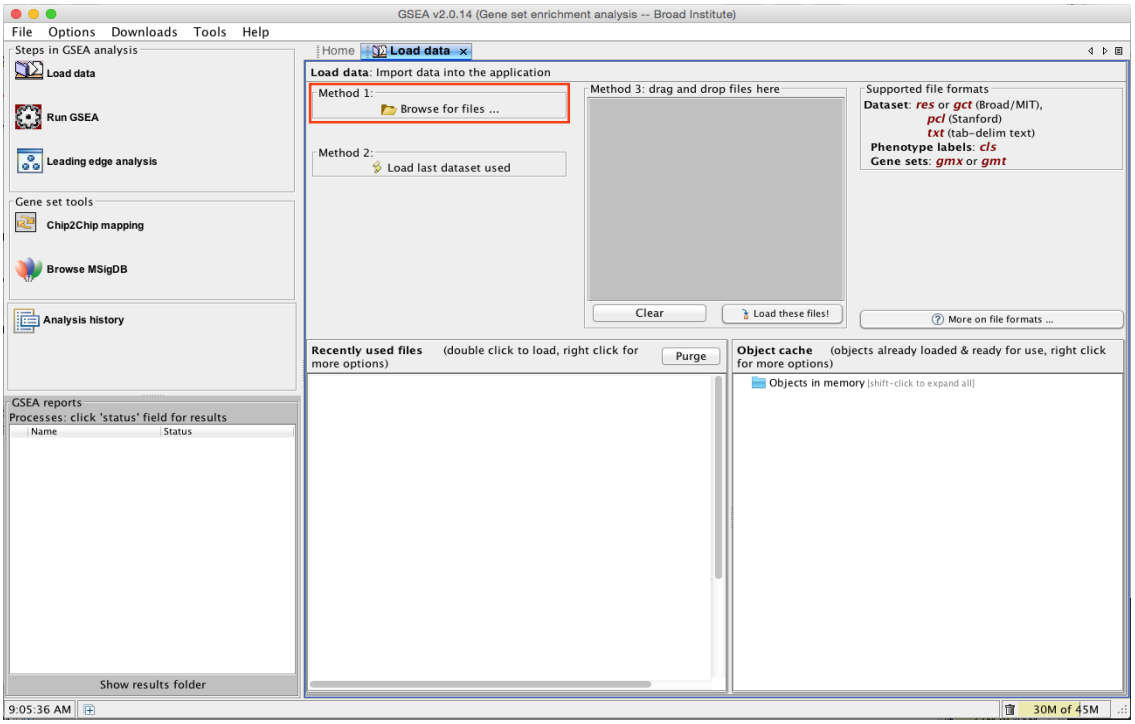
Step	Action	Check
1	Launch GSEA.	
2	Locate the '**Load data'** icon at the upper left corner of the window and **click on it**.	
3	In the central panel, select '**Method 1'** and '**Browse for files'**. A new window pops up. Browse your computer to locate the 3 files : Import the MCF7_Expression_matrix.gct, MCF7_groups.cls and Human_GO_AllPathways.gmt. Click on '**Choose'.** A message pops us when the	

	files are loaded successfully. Click on '**OK'**. Alternatively, you can choose '**Method 3'** to '**drag and drop files here'**; you need to click on the '**Load these files!'** button in this case.	
4	Locate the '**Run GSEA'** icon at the left side of the main window located below the '**Load data button'** and **click on it.**	
5	In the central window called **'GSEA: Set parameters and run enrichment tests'**, fill the first field called '**Expression dataset'** by clicking on the up and down arrows. **Choose MCF7_Expression_matrix**. Tip: Mousing over the parameters fields will highlight a short description.	
6	Click on the 3 dots [...] of the radio button corresponding to the '**Gene sets database'** field. A new window will pop up after approximately 10 seconds. Using the right arrow in the menu bar of this window, locate the '**Gene Matrix (local gmx/gmt)'** tab and select the file **Human_GO_AllPathways.gmt.** Click on '**OK'**.	
7	In the field '**Number of permutations'** enter the number 100. Note: for this exercise and purpose of demonstration, please use 100. For real life data analysis, 2000 permutations is recommended and it will require about 1 hour to run using a complete set of gene-sets.	
8	In the '**Phenotype labels'** field, click on the 3 dots [...] of the radio button. A window **"Select a phenotype"** will pop up. Make sure that the file **MCF7_groups.cls** appears as selected source file; locate and select the comparison **ES12_versus_NT12**. Click on '**OK'**.	
9	Set the '**Collapse dataset to gene symbols'** field to '**false'**.	
10	Set '**Permutation'** type to '**gene-set'**.	
11	Leave the '**Chip platform(s)'** empty.	
12	In '**Basic Fields'**, choose an informative name for your analysis in the '**Analysis name field'.** Tip: name of the comparison that you are making and date (e.g ES12_versus_NT12_date).	
13	Set the '**Metric for ranking genes'** to **'tTest'.**	
14	**'In the Save results in this folder'**, click on the 3 dots [...] and browse your computer to select a folder.	
15	Locate and click the '**Run'** button at the bottom right corner of the window. Tip: you may need to expand the window to see the Run button. Note: it takes **5 min** to run using a maximum of 1.4Gb of memory. GSEA has finished to run when a message **'Success 5'** appears in the Status field of the GSEA reports box.	
16	In the **'GSEA reports'** box, click on **'Success 5'** to see the results. Open the links and explore the results.	

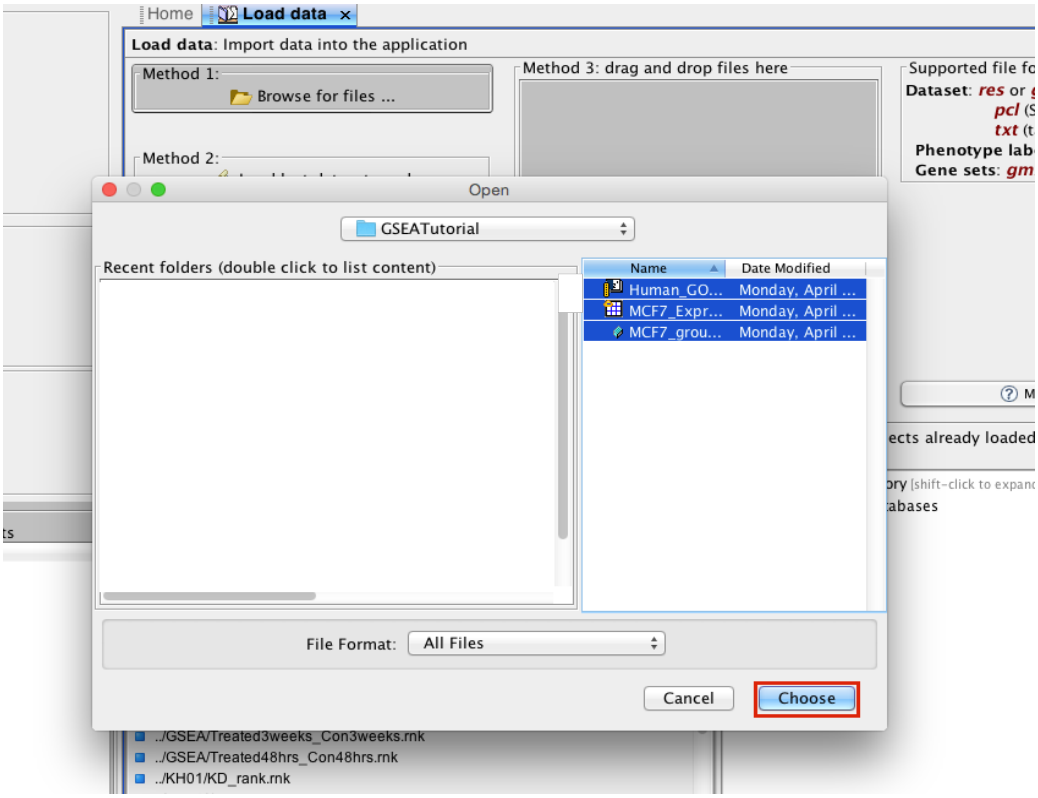
Steps 1-2



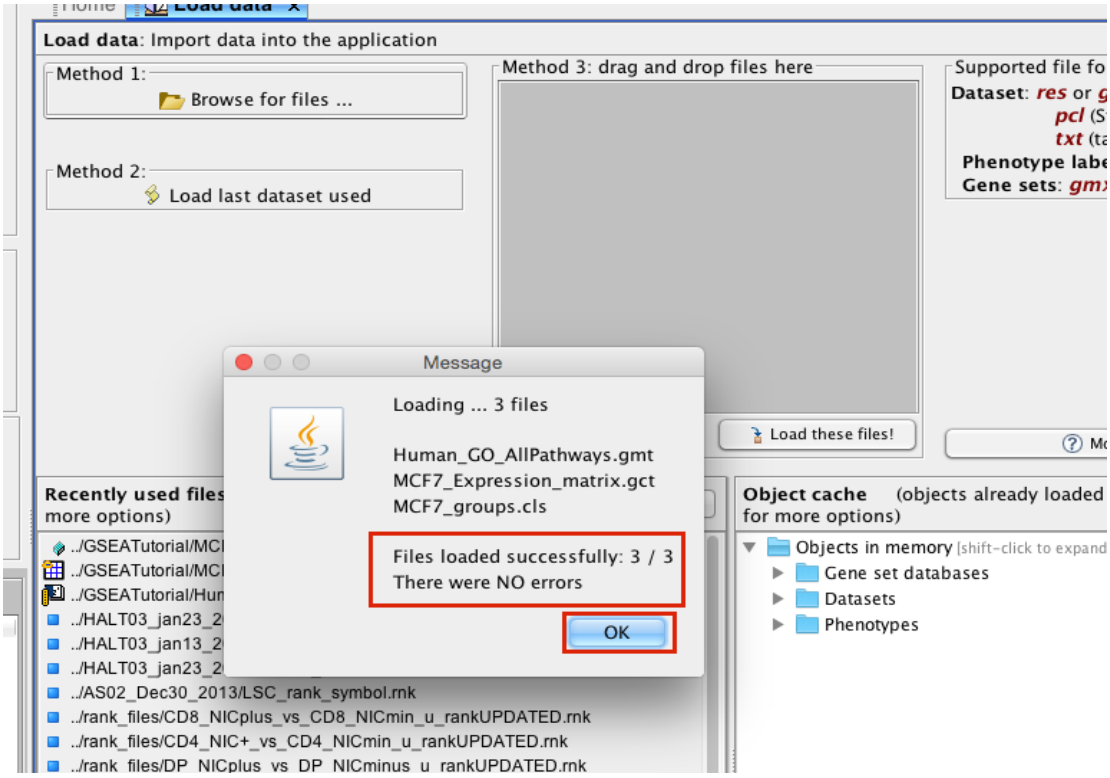
Step 3



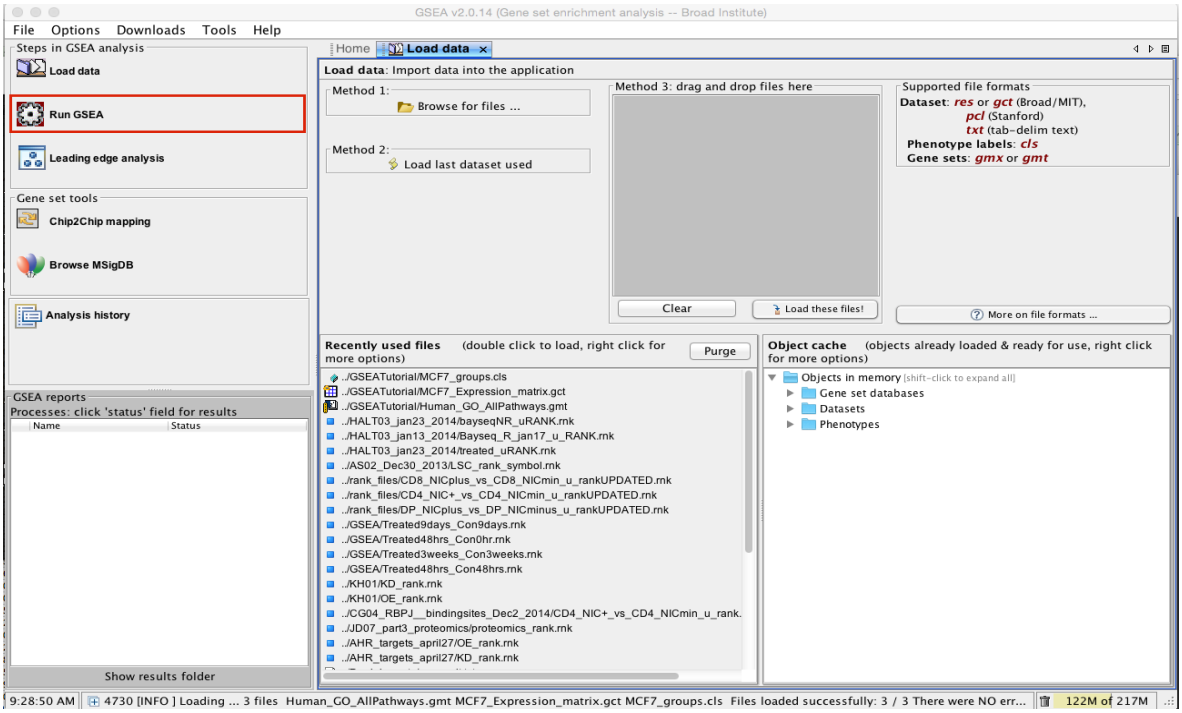
Step 3 (continued)



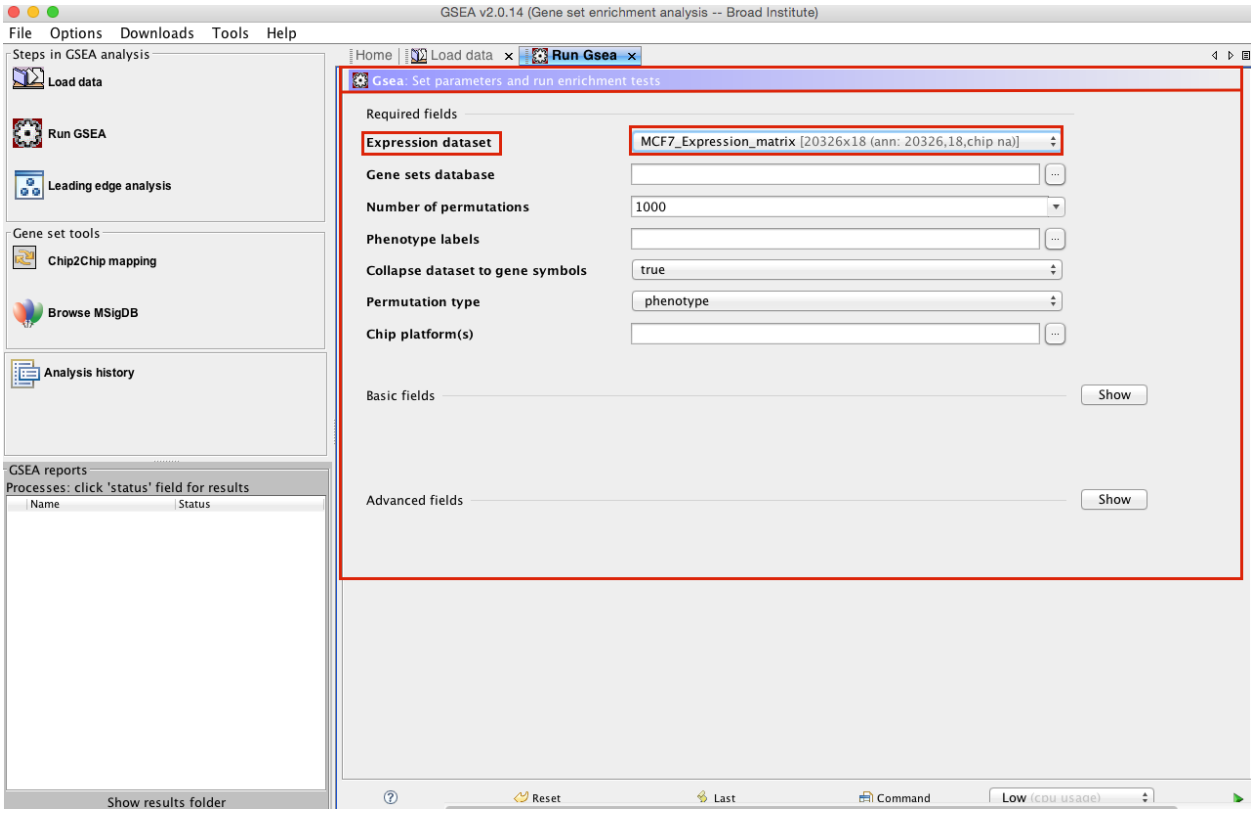
Step 3 (continued)



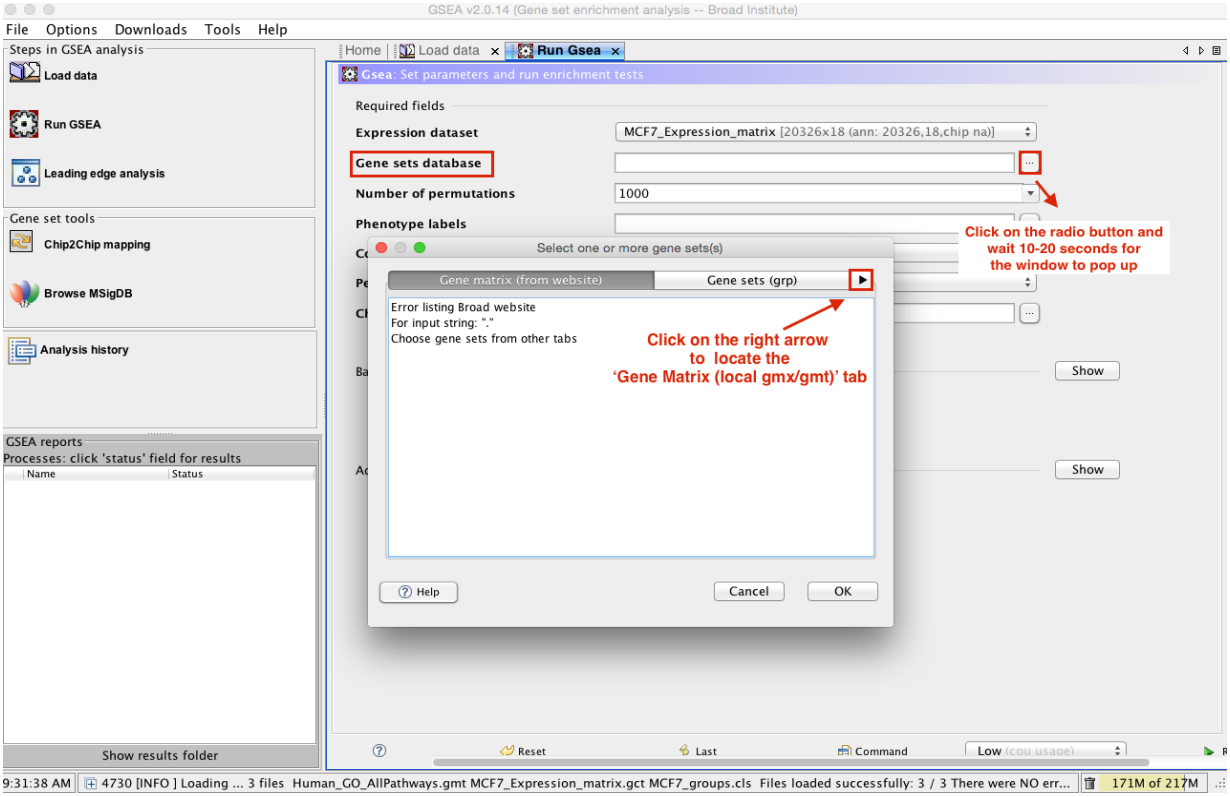
Step 4


GSEA v2.0.14 (Gene set enrichment analysis -- Broad Institute)
File Options Downloads Tools Help
Steps in GSEA analysis
Load data
Run GSEA
Leading edge analysis
Gene set tools
Chip2Chip mapping
Browse MSigDB
Analysis history
GSEA reports
Processes: click 'status' field for results
Name Status
Show results folder
Home Load data
Load data: Import data into the application
Method 1:
Browse for files ...
Method 2:
Load last dataset used
Method 3: drag and drop files here
Clear
Load these files!
Supported file formats
Dataset: res or gct (Broad/MIT),
pcl (Stanford)
txt (tab-delim text)
Phenotype labels: cls
Gene sets: gmx or gmt
More on file formats ...
Recently used files (double click to load, right click for more options)
Purge
../GSEATutorial/MCF7_groups.cls
../GSEATutorial/MCF7_Expression_matrix.gct
../GSEATutorial/Human_GO_AllPathways.gmt
../HALT03_jan23_2014/bayseqNR_uRANK.rnk
../HALT03_jan13_2014/Bayseq_R_jan17_u_RANK.rnk
../HALT03_jan23_2014/treated_uRANK.rnk
../AS02_Dec30_2013/LSC_rank_symbol.rnk
../rank_files/CD8_NICplus_vs_CD8_NICmin_u_rankUPDATED.rnk
../rank_files/CD4_NIC+_vs_CD4_NICmin_u_rankUPDATED.rnk
../rank_files/DP_NICplus_vs_DP_NICminus_u_rankUPDATED.rnk
../GSEA/Treated9days_Con9days.rnk
../GSEA/Treated48hrs_Con0hr.rnk
../GSEA/Treated3weeks_Con3weeks.rnk
../GSEA/Treated48hrs_Con48hrs.rnk
../KH01/KD_rank.rnk
../KH01/OE_rank.rnk
../CG04_RBPJ__bindingsites_Dec2_2014/CD4_NIC+_vs_CD4_NICmin_u_rank.
../JD07_part3_proteomics/proteomics_rank.rnk
../AHR_targets_april27/OE_rank.rnk
../AHR_targets_april27/KD_rank.rnk
Object cache (objects already loaded & ready for use, right click for more options)
Objects in memory [shift-click to expand all]
Gene set databases
Datasets
Phenotypes
9:28:50 AM 4730 [INFO] Loading ... 3 files Human_GO_AllPathways.gmt MCF7_Expression_matrix.gct MCF7_groups.cls Files loaded successfully: 3 / 3 There were NO err... 122M of 217M

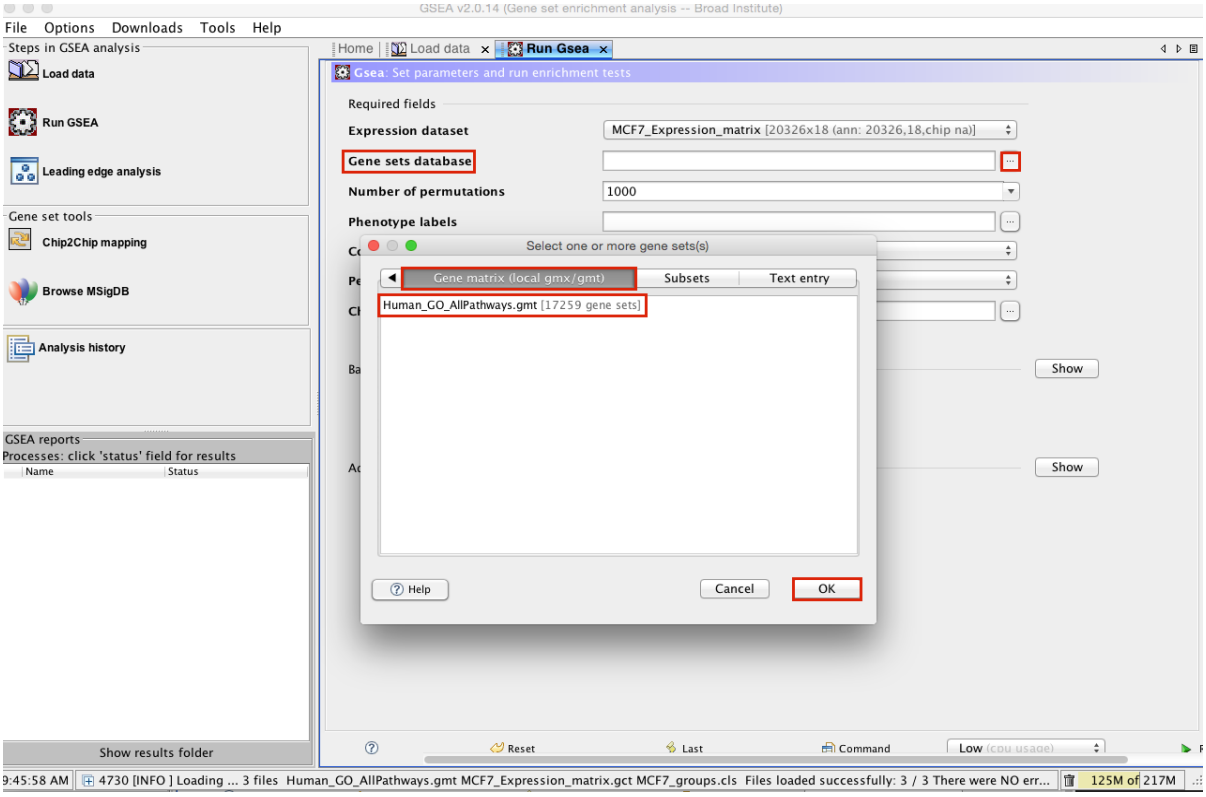
Step 5


GSEA v2.0.14 (Gene set enrichment analysis -- Broad Institute)
File Options Downloads Tools Help
Steps in GSEA analysis
Load data
Run GSEA
Leading edge analysis
Gene set tools
Chip2Chip mapping
Browse MSigDB
Analysis history
GSEA reports
Processes: click 'status' field for results
Name Status
Show results folder
Home Load data Run Gsea
Gsea: Set parameters and run enrichment tests
Required fields
Expression dataset
MCF7_Expression_matrix [20326x18 (ann: 20326,18,chip na)]
Gene sets database
Number of permutations
1000
Phenotype labels
Collapse dataset to gene symbols
true
Permutation type
phenotype
Chip platform(s)
Basic fields
Show
Advanced fields
Show
Reset
Last
Command
Low (cpu usage)

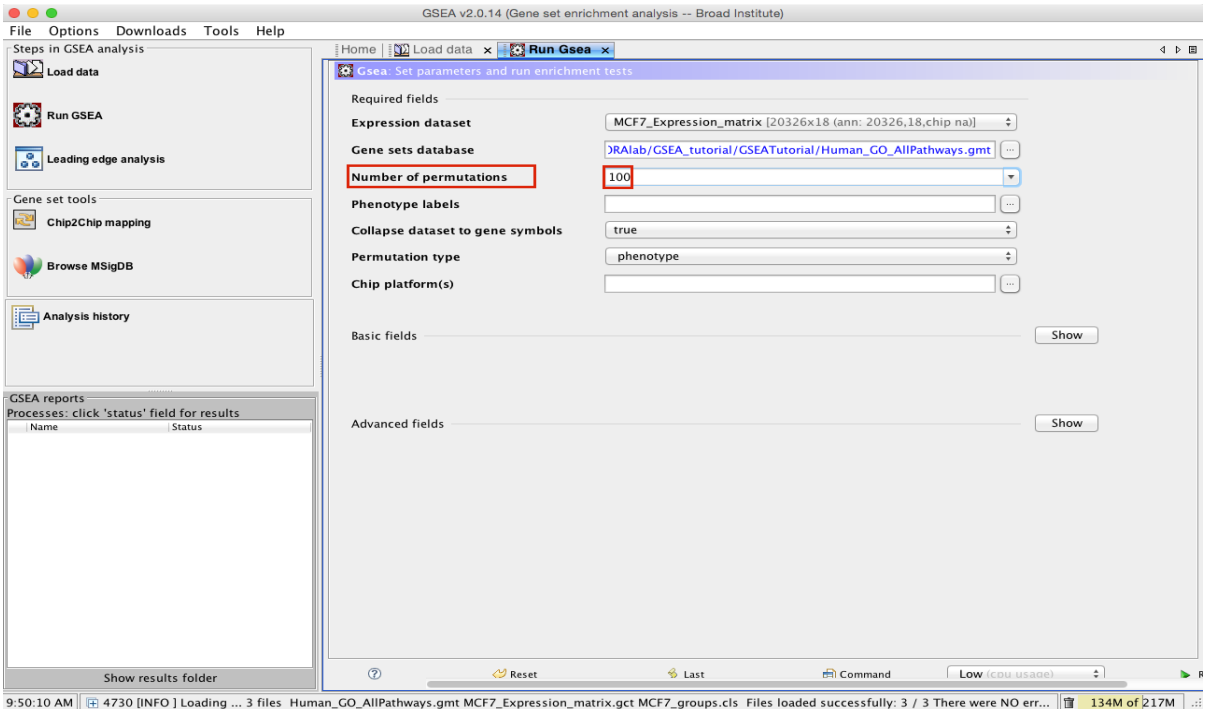
Step 6



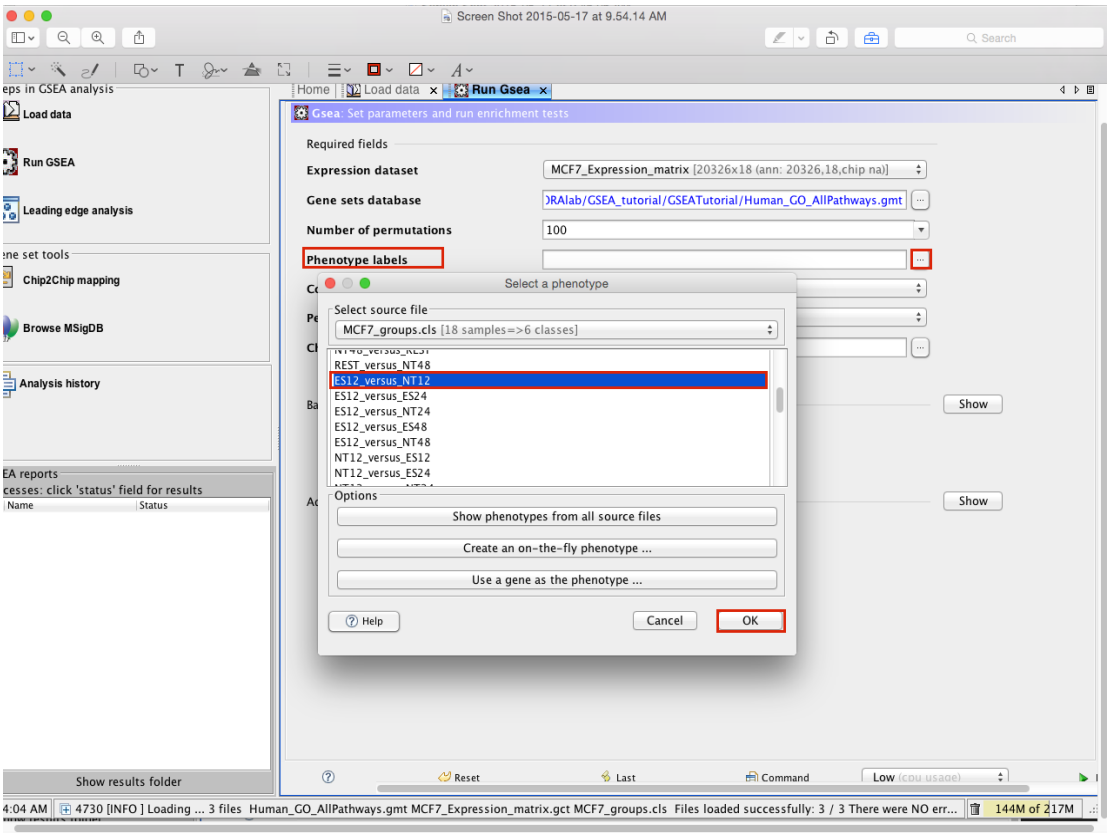
Step 6 (continued)



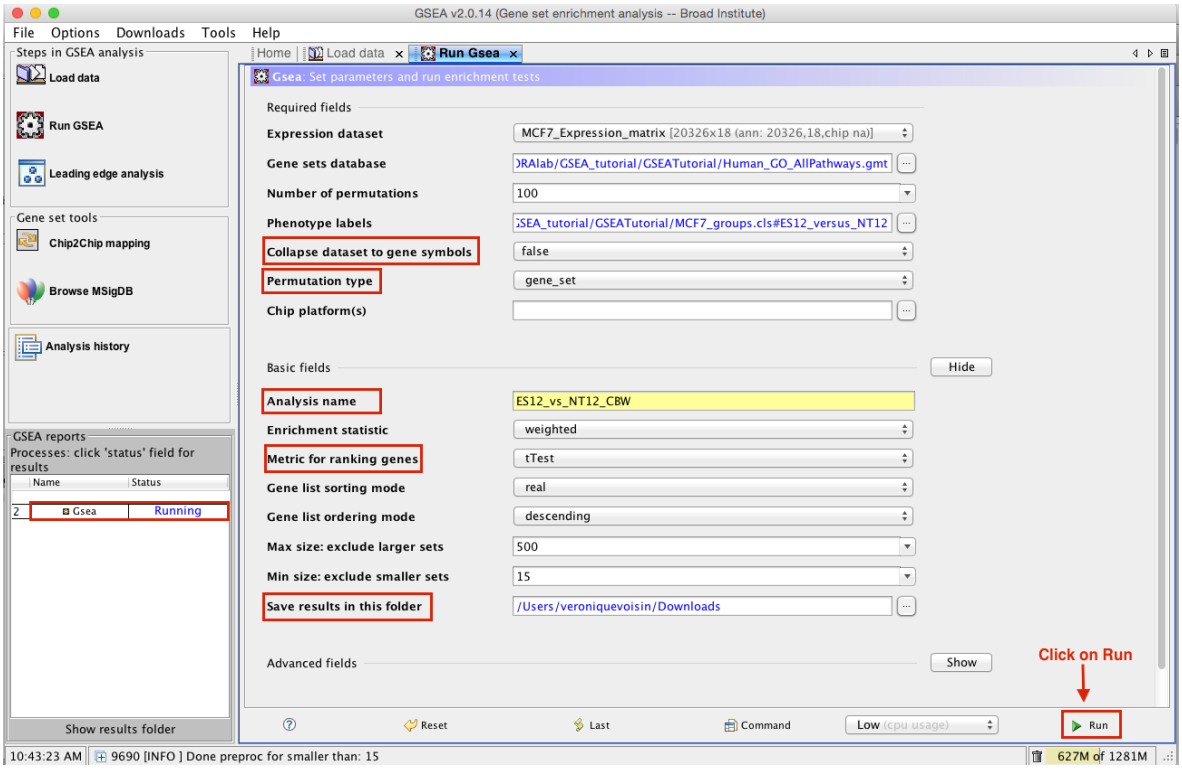
Step 7


GSEA v2.0.14 (Gene set enrichment analysis -- Broad Institute)
File Options Downloads Tools Help
Steps in GSEA analysis
Load data
Run GSEA
Leading edge analysis
Gene set tools
Chip2Chip mapping
Browse MSigDB
Analysis history
GSEA reports
Processes: click 'status' field for results
Name Status
Show results folder
Home Load data Run Gsea
Gsea: Set parameters and run enrichment tests
Required fields
Expression dataset
MCF7_Expression_matrix [20326x18 (ann: 20326,18,chip na)]
Gene sets database
)RAlab/GSEA_tutorial/GSEATutorial/Human_GO_AllPathways.gmt
Number of permutations
100
Phenotype labels
Collapse dataset to gene symbols
true
Permutation type
phenotype
Chip platform(s)
Basic fields
Show
Advanced fields
Show
Reset
Last
Command
Low (cpu usage)
9:50:10 AM
4730 [INFO] Loading ... 3 files Human_GO_AllPathways.gmt MCF7_Expression_matrix.gct MCF7_groups.cls Files loaded successfully: 3 / 3 There were NO err...
134M of 217M

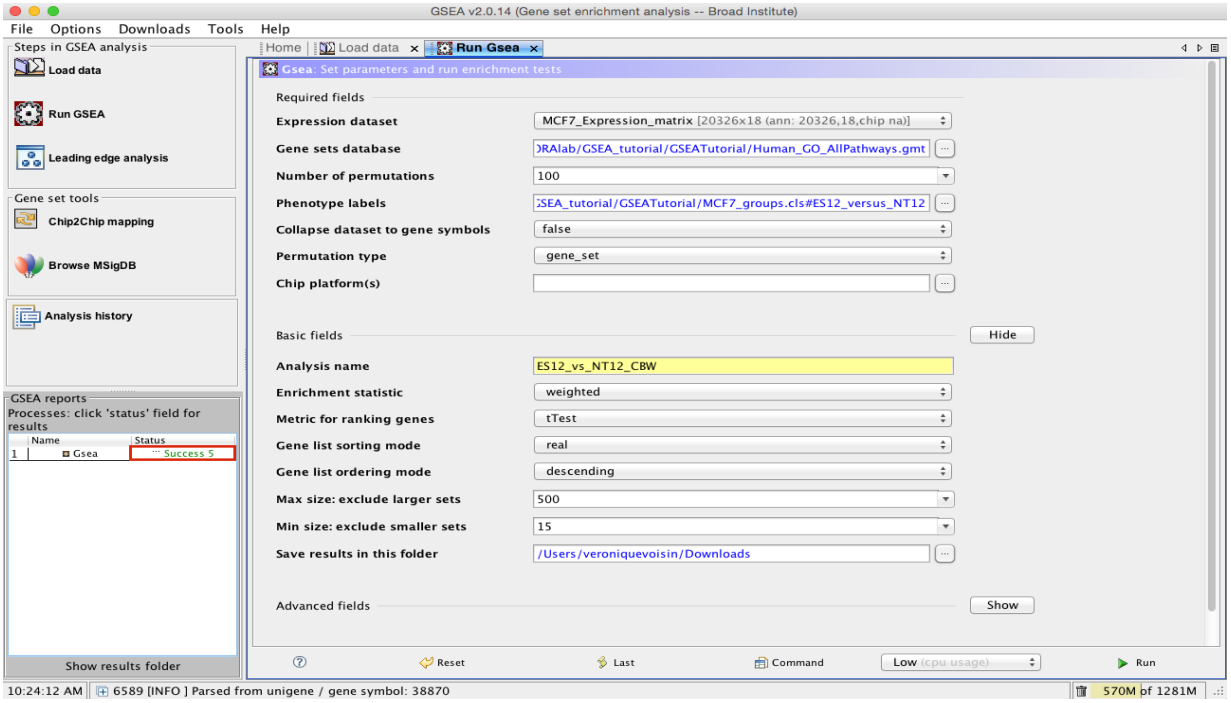
Step 8


Screen Shot 2015-05-17 at 9.54.14 AM
Search
Load data
Run GSEA
Chip2Chip mapping
Browse MSigDB
Analysis history
Show results folder
Home Load data Run Gsea
Gsea: Set parameters and run enrichment tests
Required fields
Expression dataset
MCF7_Expression_matrix [20326x18 (ann: 20326,18,chip na)]
Gene sets database
)RAlab/GSEA_tutorial/GSEATutorial/Human_GO_AllPathways.gmt
Number of permutations
100
Phenotype labels
Select a phenotype
Select source file
MCF7_groups.cls [18 samples=>6 classes]
REST_versus_NT48
ES12_versus_NT12
ES12_versus_ES24
ES12_versus_NT24
ES12_versus_ES48
ES12_versus_NT48
NT12_versus_ES12
NT12_versus_ES24
Options
Show phenotypes from all source files
Create an on-the-fly phenotype ...
Use a gene as the phenotype ...
Help
Cancel
OK
Show
Show
Reset
Last
Command
Low (cpu usage)
4730 [INFO] Loading ... 3 files Human_GO_AllPathways.gmt MCF7_Expression_matrix.gct MCF7_groups.cls Files loaded successfully: 3 / 3 There were NO err...
144M of 217M

Steps 9-15



Step 16



Step 17

GSEA Report for Dataset MCF7_Expression_matrix

gene-sets enriched in genes up-regulated in treated cells compared to non-treated samples

Enrichment in phenotype: ES12 (3 samples)

- 2120 / 4756 gene sets are upregulated in phenotype **ES12**
- 665 gene sets are significant at FDR < 25%
- 422 gene sets are significantly enriched at nominal pvalue < 1%
- 612 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

gene-sets enriched in genes down-regulated in treated cells compared to non-treated samples

Enrichment in phenotype: NT12 (3 samples)

- 2636 / 4756 gene sets are upregulated in phenotype **NT12**
- 445 gene sets are significantly enriched at FDR < 25%
- 337 gene sets are significantly enriched at nominal pvalue < 1%
- 601 gene sets are significantly enriched at nominal pvalue < 5%
- Snapshot of enrichment results
- Detailed enrichment results in html format
- Detailed enrichment results in excel format (tab delimited text)
- Guide to interpret results

Dataset details

- The dataset has 20323 features (genes)
- No probe set => gene symbol collapsing was requested, so all 20323 features were used

Gene set details

- Gene set size filters (min=15, max=500) resulted in filtering out 12503 / 17259 gene sets
- The remaining 4756 gene sets were used in the analysis
- List of gene sets used and their sizes (restricted to features in the specified dataset)

Gene markers for the ES12 *versus* NT12 comparison

- The dataset has 20323 features (genes)
- # of markers for phenotype **ES12**: 9758 (48.0%) with correlation area 49.7%
- # of markers for phenotype **NT12**: 10565 (52.0%) with correlation area 50.3%
- Detailed rank ordered gene list for all features in the dataset
- Heat map and gene list correlation profile for all features in the dataset