OCT 09, 2023

OPEN ACCESS

Protocol Citation: Cameron Baker 2023. Gene Set Enrichment Analysis. **protocols.io** https://protocols.io/view/gene-set-enrichment-analysis-cw8gxhtw

Protocol status: Other

Created: Jul 13, 2023

Last Modified: Oct 09, 2023

PROTOCOL integer ID: 84968

Gene Set Enrichment Analysis

Cameron Baker[1,2,3]

[1]University of Rochester Genomics Research Center; [2]Center for Advanced Research Technologies; [3]Wilmot Cancer Center

Cameron Baker

ABSTRACT

Instructions for running https://www.gsea-msigdb.org/gsea/index.jsp

Introduction

1 Gene Set Enrichment Analysis (GSEA) is an alternative step to more general pathway enrichment (as performed by our standard RNA-Seq workflow) to determine significant expression within an experiment corresponding to predefined genesets. It is a relatively common request, and the best application for running GSEA that I've run into is the Broad institute stand alone program. This guide is intended to get users from the materials that we provide back to them to analyzed results.

Downloading and Installation

2 The standalone app, system dependent, may be downloaded from https://www.gsea-msigdb.org/gsea/downloads.jsp after providing your email.

While they provide gene sets ready to download, they are also fetched within the tool at run time.

Preparing Your Data

3 If working from our standard delivery, you will need the normalized counts for each sample located within the **deSeq2_NormCounts.txt** file.

You will need to make some minor changes to this file, detailed below.

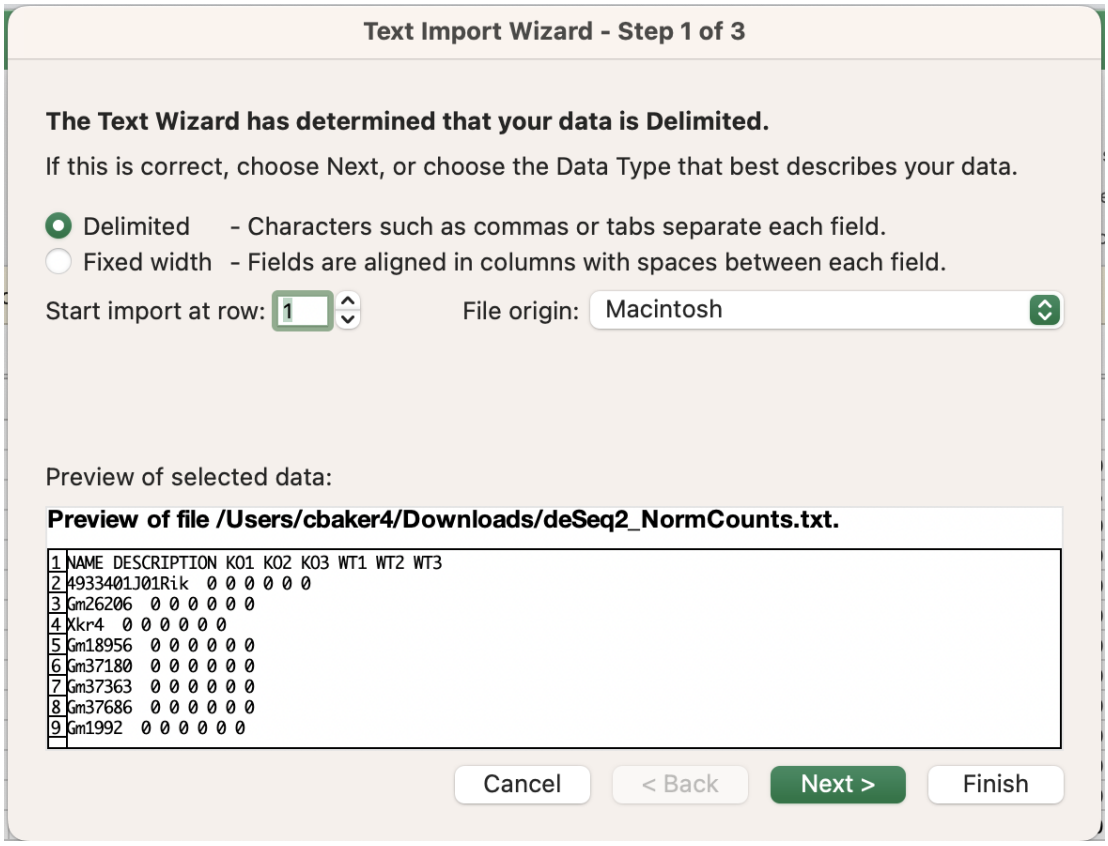
3.1 Below is a short example of normalized counts for the first 10 genes across the first 5 samples. Within the delivered files, the **genes will be the row names** and the **samples will be the column names**.

	A	B	C	D	E	F
		AML10	AML11	AML1	AML2	AML3
	DDX11L1	0	0	0	0	0
	WASH7P	0	1.877518671	0	0.997665923	5.48419191
	MIR6859-1	0	0.938759336	0.811303251	0.997665923	2.056571966
	MIR1302-2HG	0	0	0	0	0
	MIR1302-2	0	0	0	0	0
	FAM138A	0	0	0	0	0
	OR4G4P	0	0	0	0	0
	OR4G11P	0	0	0	0	0
	OR4F5	0	0	0	0	0
	AL627309.1	0	0	0.811303251	0.997665923	0

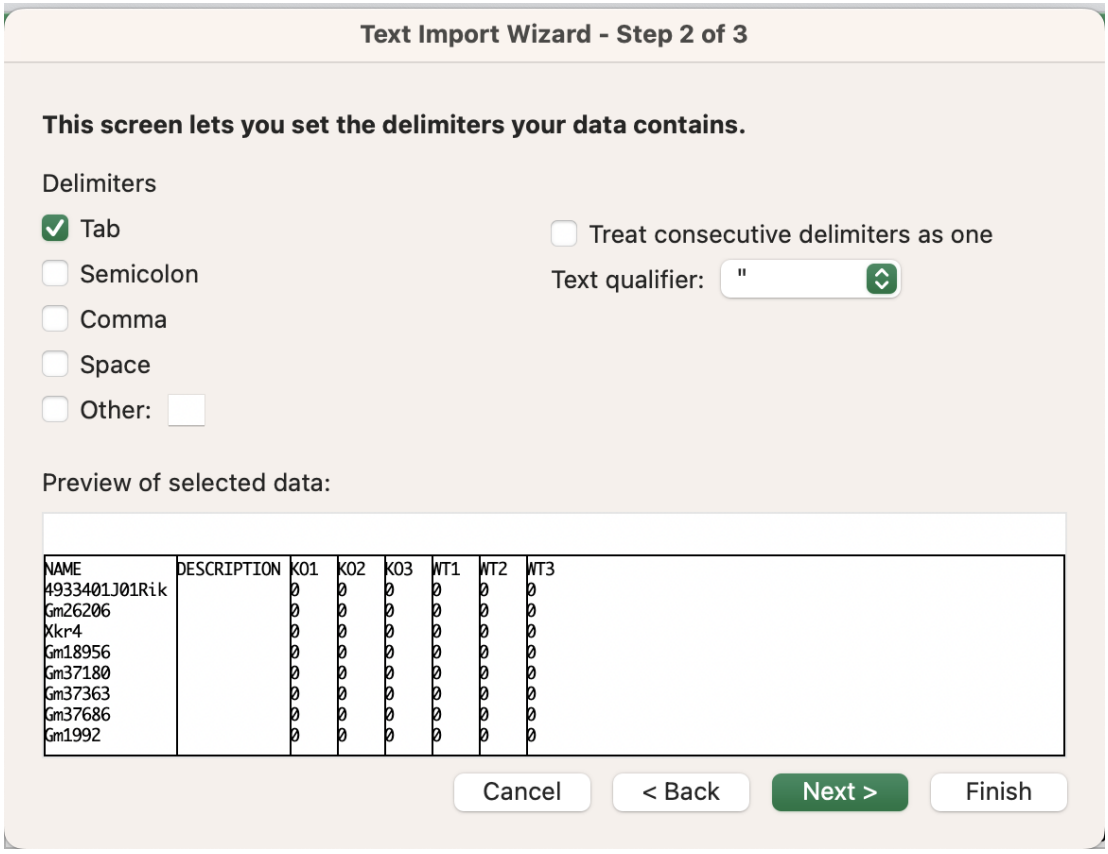
We will need to add in a column name to indicate that the gene names are the genes and an empty column named description. You can do this by

1. Right click on **deSeq2_NormCounts.txt**, select **Open With**, and then select **Excel**. If Open With / Excel is not an option, you can go into Excel, select **File**, **Open,** find **deSeq2_NormCounts.txt**, select **Next** within the import wizard, and then **Finish**

Option 1 for importing txt file in Excel

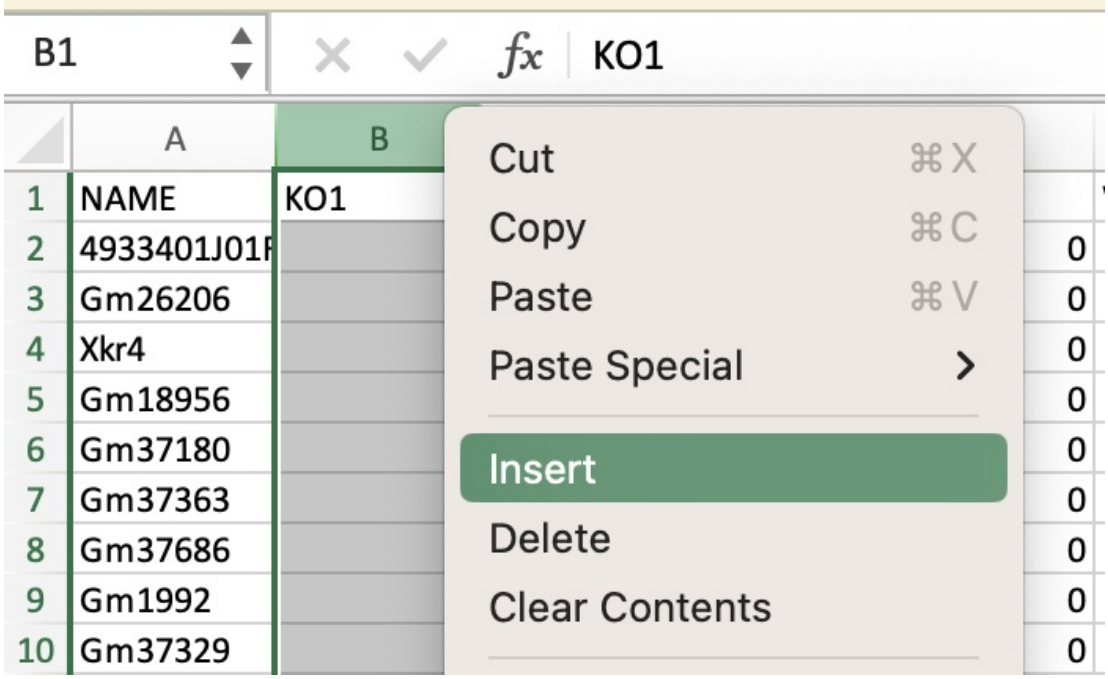


Option 2a for importing txt file in Excel, click **Next**



Option 2b for importing txt file in Excel, click **Finish**

2. Edit the name of the first column, where the rows are gene names, to be called **NAME**
3. Insert an empty column between the first and second columns. This can be done by right clicking the top of the second column and clicking **Insert**.



4. Name this column **DESCRIPTION**

This txt file should now have a **NAME** and **DESCRIPTION** within the first two column names. You may save the file and exit. **Make sure it is saved as a txt file**. Below is the basic format from above, filled in correctly.

	A	B	C	D	E	F	G	H	I
	NAME	DESCRIPTION	AML10	AML11	AML1	AML2	AML3	AML4	AML5
	DDX11L1		0	0	0	0	0	0	0.967626664
	WASH7P		0	1.877518671	0	0.997665923	5.48419191	0.997805736	6.773386649
	MIR6859-1		0	0.938759336	0.811303251	0.997665923	2.056571966	2.993417207	1.935253328
	MIR1302-2HG		0	0	0	0	0	0	0
	MIR1302-2		0	0	0	0	0	0	0
	FAM138A		0	0	0	0	0	0	0
	OR4G4P		0	0	0	0	0	0	0
	OR4G11P		0	0	0	0	0	0	0
	OR4F5		0	0	0	0	0	0	0

3.2 In addition to editing the normalized counts file, we will also need to create a metadata file that will tell the software which samples belong to which groups. To do this, we will need to create a separate text file.

1. Windows users can use notepad and Mac users can download an external text editor like sublime text (https://www.sublimetext.com)

2. Create new file
3. Within the first line, you will put 3 numbers. The first number is the number of samples within your normalized counts file. The second number is the number of groups. The third number is just 1. These are separated by spaces.

Example first line of metadata file. This indicates an experiment with 28 samples and 3 sample groups.

```
28  3  1
```

4. The second line of the file indicates the names the sample groups. Starting with a **#**, enter in the sample groups separated by a space

Example second line of the metadata file. This indicates the names of the 3 sample groups.

```
# AML bcCML BM
```

5. The last line of the file will be the space separated order of the samples, as they exist within the normalized counts file.

Example third line of the metadata file. This indicates the order of the samples within the normalized counts and the groups they belong to.

```
AML AML AML AML AML AML AML AML AML AML AML bcCML bcCML bcCML
bcCML bcCML bcCML bcCML bcCML bcCML bcCML BM BM BM BM BM BM BM
```

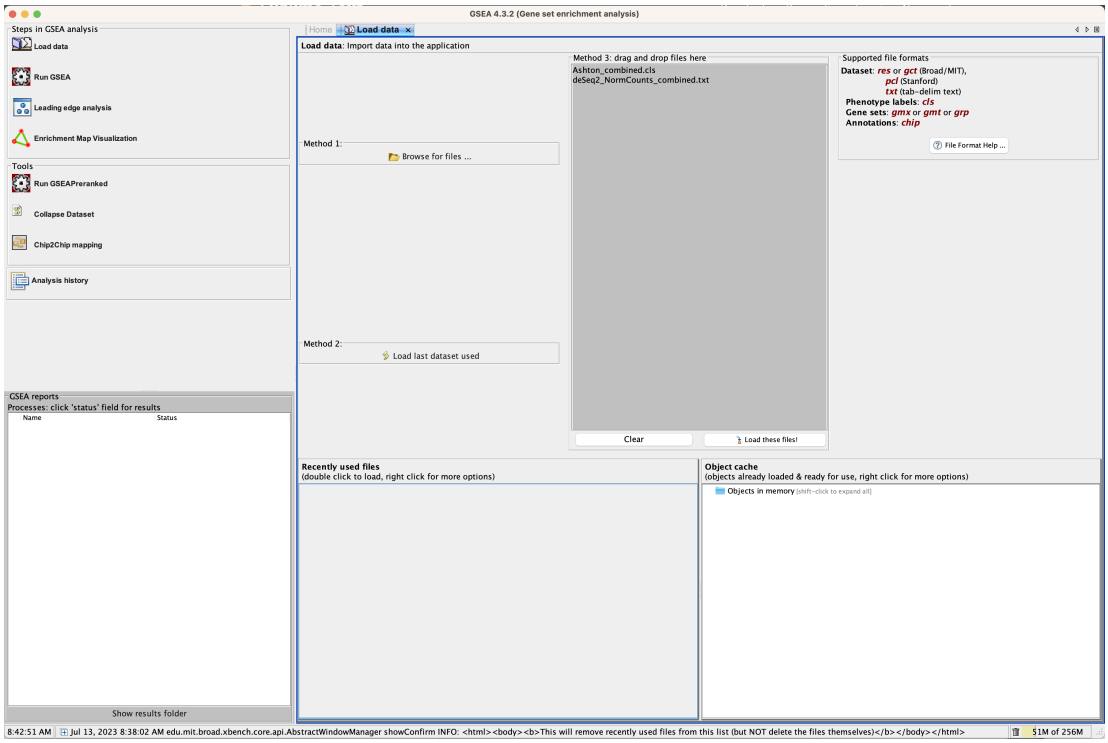
6. save the file, ending the file extension in **.cls**

The resultant **.cls** file, created within Sublime text editor.

More information related to setting up this file, or if you have continuous data, can be found at the following link: https://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats#Phenotype_Data_Formats

Loading Your Data

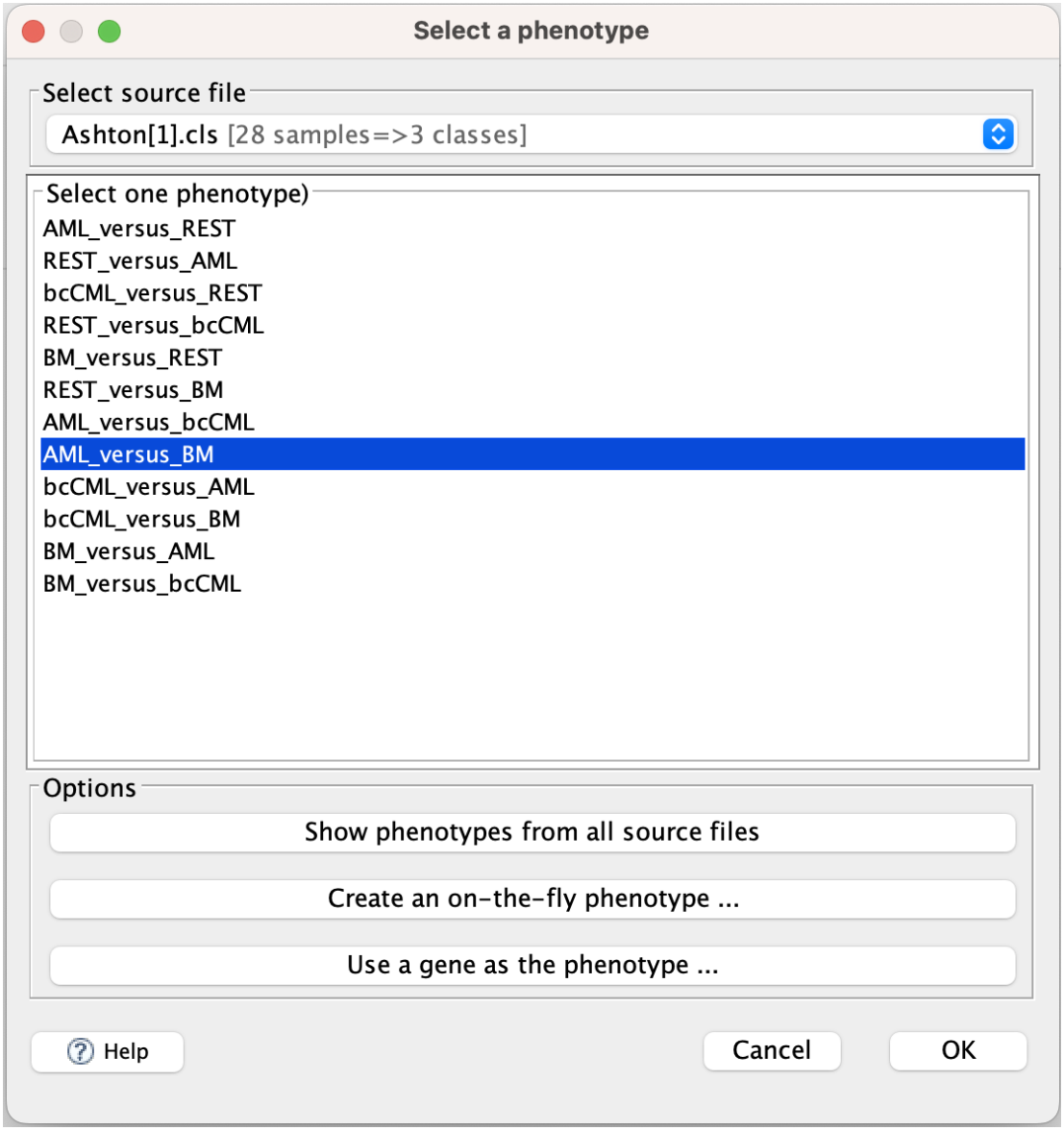
4 Within the GSEA app, you can drag and drop your normalized counts and meta data files into the **Method 3** box.



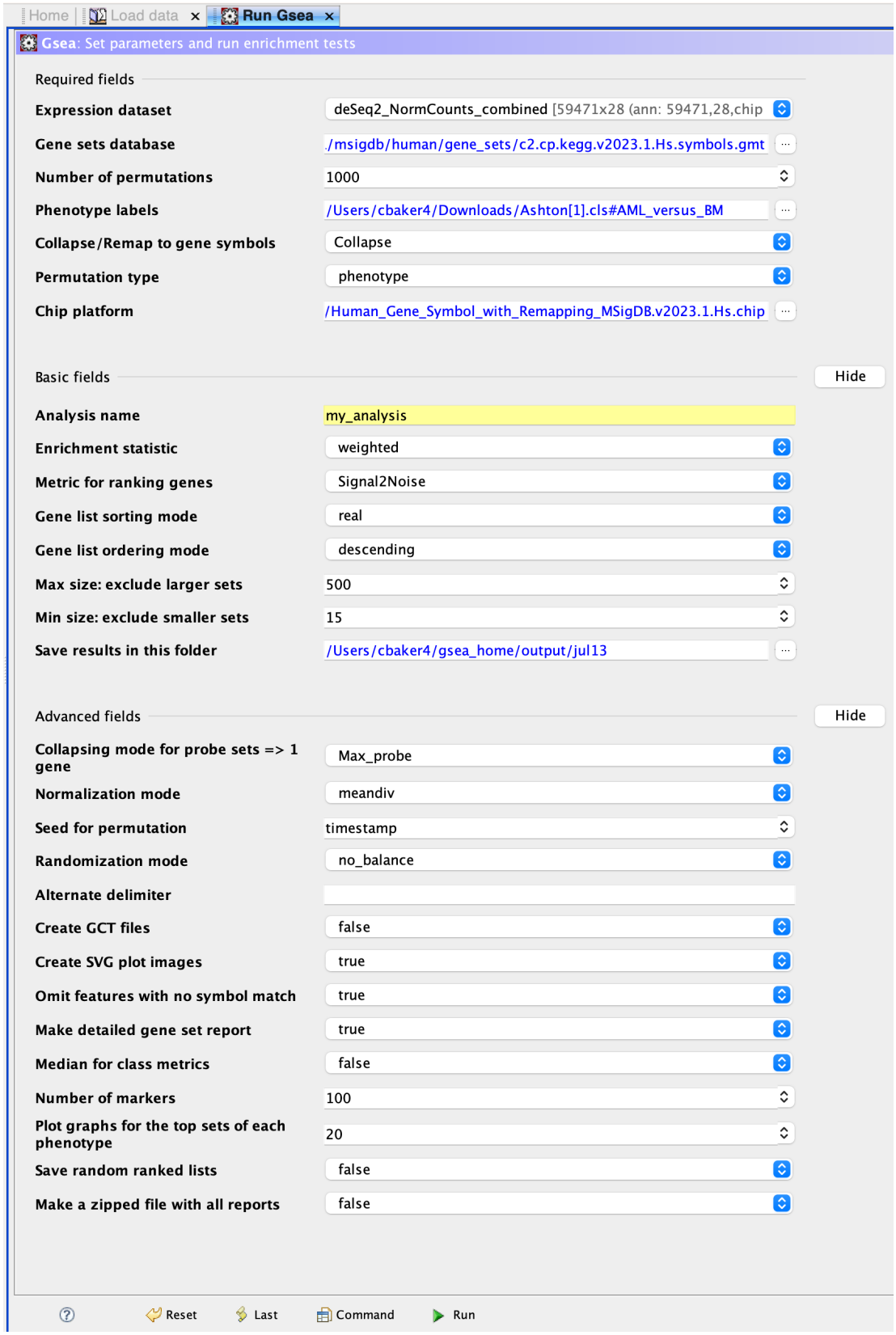
Running GSEA

5 Click on the **Run GSEA** button on the left hand side of the screen shot above.

1. Select the normalized counts file within the drop down menu of **Expression dataset**
2. Select your Gene set database of interest. I usually like **c2.cp.kegg.v2023.1.Hs.symbols.gmt**. This may be slightly different if you are using a mouse dataset. You may also explore other databases
3. Select your phenotype file and comparison of interest within the **Phenotype labels** pop up menu. Below is an example.



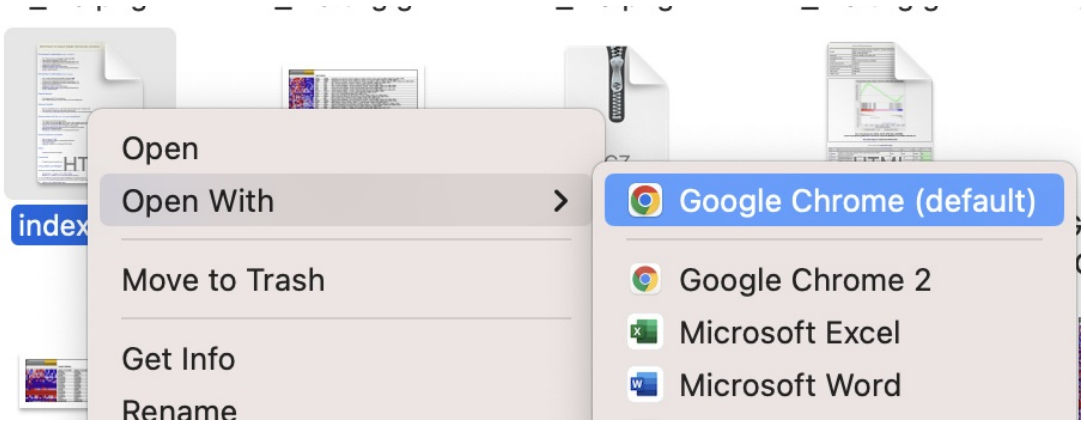
4. For **Permutation type**, Broad recommends **Phenotype** if you have **at least seven** samples within each phenotype. Otherwise, choose **gene_set**. More information on permutation type can be found here https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html?Run_GSEA_Page
5. The **Chip platform** indicates the format of our gene ID's. In this case, it is **Human_Gene_Symbol_with_Remapping_MSigDB.v2023.1.Hs.Chip**. In our deliveries, we return results in either human gene symbol of mouse gene symbol.
6. From there, I usually leave most of the fields as is. You may wish to export images as **svg** if they will be used for a publication. You may also wish to adjust the number of enrichments plotted, via **Plot graphs for the top sets of each phenotype**, if you see some significant hits you would like to include outside of the default 20.
7. Click **Run.** Below are our final list of parameters.



8. Once the analysis finishes running, you can click the green **Success** square within the **GSEA Reports** window to arrive at the landing page for your results, **index.html**

Perusing your results

6 The path to the output folder should be located within the **Save results to this folder** path indicated in the **Basic fields** section (see the screenshot above). The most important file is **index.html** as this contains experiment level information and serves as a map to traverse the large number of plots written by the workflow. Below is an example screenshot of index.html. If it does not open in a browser automatically, you can **right click -> Open with -> (your browser of choice)**.



You should be all set. For more information related to interpretation, including the tantalizing *Why does GSEA use a FDR cutoff of 0.25 instead of 0.05?*, the writers at Broad do a better job then I could explaining at https://www.gsea-msigdb.org/gsea/doc/GSEAUserGuideFrame.html?Interpreting_GSEA