Manual of the `MATLAB` pipeline for the analysis of genomic time course data from the *Gene Expression Omnibus*

Juan Camilo Ramírez

December 9, 2016

Contents

List of Figures

List of Tables

1 Summary

1. The pipeline analysis is performed individually for each experimental condition.
2. An experimental condition is composed of several GEO samples, all associated to a GEO series.
3. The information of each sample must be provided in an input file located in folder Input.
4. Several condition files can be deposited in folder Input and the pipeline analysis will be performed on all the conditions.
5. The analysis results for all condition files given as input are output in folder Results.

6. Section 1 explains how to prepare the condition files that will be given as input.

7. Section 2 explains how to run the analysis on the condition files created in Section 1.

2 Theoretical basis and terminology

$$
\begin{gathered}
A = AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA \\
B = BBBBBBBBBBBBBBB \\
C = CCCCC, D = DDDDDDD \\
E = EEEEE, F = FFFFFFF
\end{gathered}
\tag{1}
$$

2.1 Experimental conditions

2.2 Experimental macroconditions

3 The pipeline analysis

4 Input files

4.1 Condition files

Condition files are the input files for the pipeline analysis of one or more experimental conditions. The information of each experimental condition must be provided in a text file. One file for each condition. The condition files must be stored in folder `Input`. Each condition file must be named with the following format: `[GEO SERIES NUMBER]_-_[NAME OF EXPERIMENTAL CONDITION]_-_[NUMBER OF TOP DRGs FOR CLUSTERING].txt`.

All the samples associated with the experimental condition must be provided along with the time points using the format described as follows. Each line of the file must consist of the time point, including the time unit (e.g., 2 hours), followed by a comma (,) and then followed by the accession number of one sample. For example, the file for condition "D10" of *GEO* series `GSE59015` must be named `GSE59015_-_D10_-_3000.csv` and the contents must be as follows.

0 hours,GSM1424453
6 hours,GSM1424454
12 hours,GSM1424455
18 hours,GSM1424456
24 hours,GSM1424457
30 hours,GSM1424458
36 hours,GSM1424459
42 hours,GSM1424460

The condition file can group different replicates that represent the same experimental condition. In this case the pipeline will run one single analysis with the average of all the replicates specified in the file. For example, the file for all the replicates in *GEO* series `GSE41067` must be named `GSE41067_-_ALL_-_10000.csv` and the contents must be as follows.

0 hours,GSM1008154,GSM1008162,GSM1008170
1 hours,GSM1008155,GSM1008163,GSM1008171
2 hours,GSM1008156,GSM1008164,GSM1008172
4 hours,GSM1008157,GSM1008165,GSM1008173
6 hours,GSM1008158,GSM1008166,GSM1008174
8 hours,GSM1008159,GSM1008167,GSM1008175
10 hours,GSM1008160,GSM1008168,GSM1008176
12 hours,GSM1008161,GSM1008169,GSM1008177

The condition files can be written manually, following strictly the format described above. Optionally, this task can be carried out more easily by using script `create_input_files.m`. In order to do this, `create_input_files.m` must be run and the instructions provided thereafter by the program must be followed. More details in Section **??**.

4.2 Macrocondition files

The conditions comprising the macrocondition must be provided in a text file that must be located in folder `Input`. This file must be named using the following format: `[GEO SERIES NUMBER]_-_[NAME OF MACRO CONDITION].txt`. For example, the following are the contents of a macrocondition of five H3N1 subjects from `GSE52428`.

H3N1_001
H3N1_002
H3N1_003
H3N1_004
H3N1_005

This file can be constructed manually, or with BASH script `prepare_input.sh`. This option requires running the script with the following syntax.

`./prepare_input.sh [GEO SERIES] [NAME OF MACRO CONDITION].txt`

Example

`./prepare_input.sh GSE52428 H1N1.txt`

The above will read ALL the conditions whose analyses have been completed for the GEO series indicated and write the file with the format described earlier.

5 create_input_files.m

The (experimental) condition files for the pipeline analysis can be created manually with the format described in Section **??** or with the help of script `create_input_files.m`. In order to do the latter, the command `create_input_files` must be executed on the `MATLAB` console from the pipeline directory, as shown in Figure **??**.

6 pipeline.m

The condition files must be prepared, as described in Section **??**, and placed in folder `Input`. After this, the pipeline analysis can be started by opening the `MATLAB` console in the pipeline directory and running `pipeline`, as shown in Figure **??**

Once all condition files are located in folder Input, then the analysis can be started by running pipeline.m. The script will read ALL the condition files in folder Input and run the analysis for each one of them.

7 integrated_analisis.m

`integrated_analisis.m`: input is a macrocondition file.

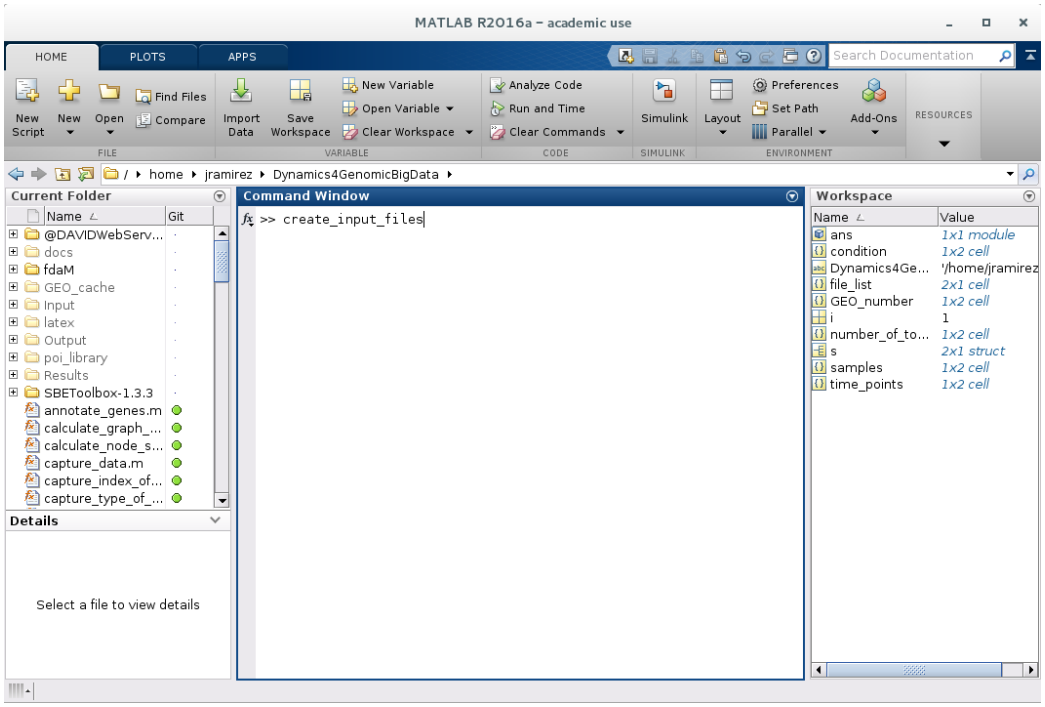


Figure 1: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

8 `compare.m`

`compare.m`: input is a macrocondition file.

9 `measure_fit_of_replicates.m`

`measure_fit_of_replicates.m`: input is a macrocondition file.

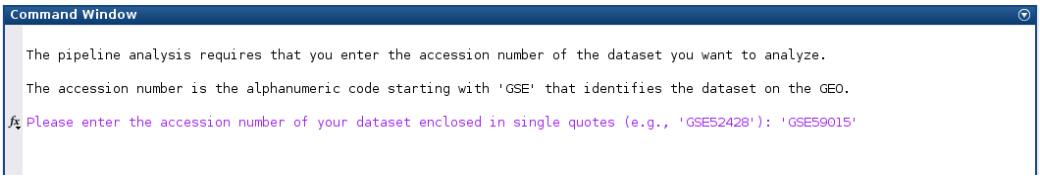


Figure 2: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

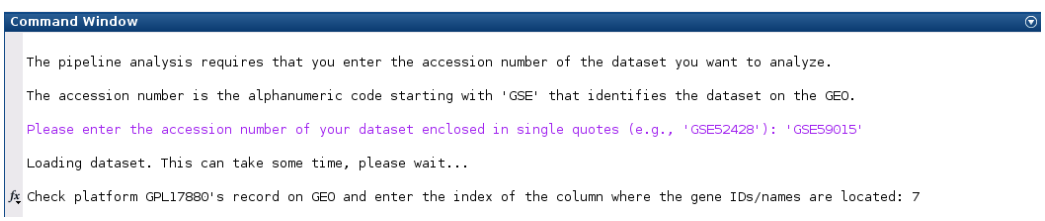


Figure 3: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

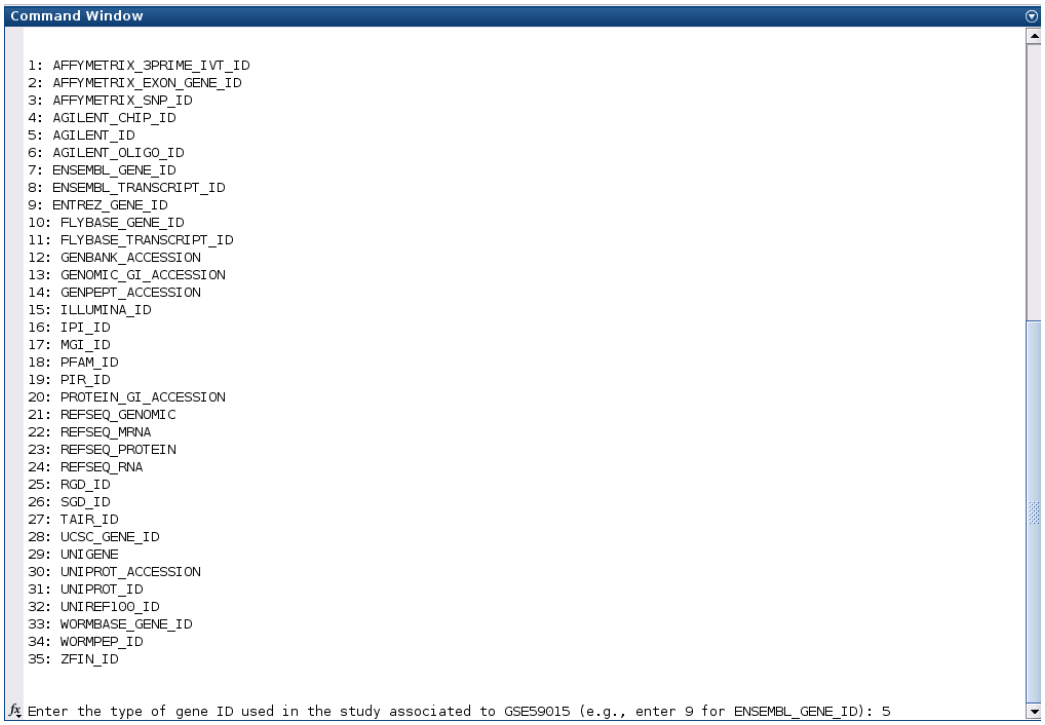
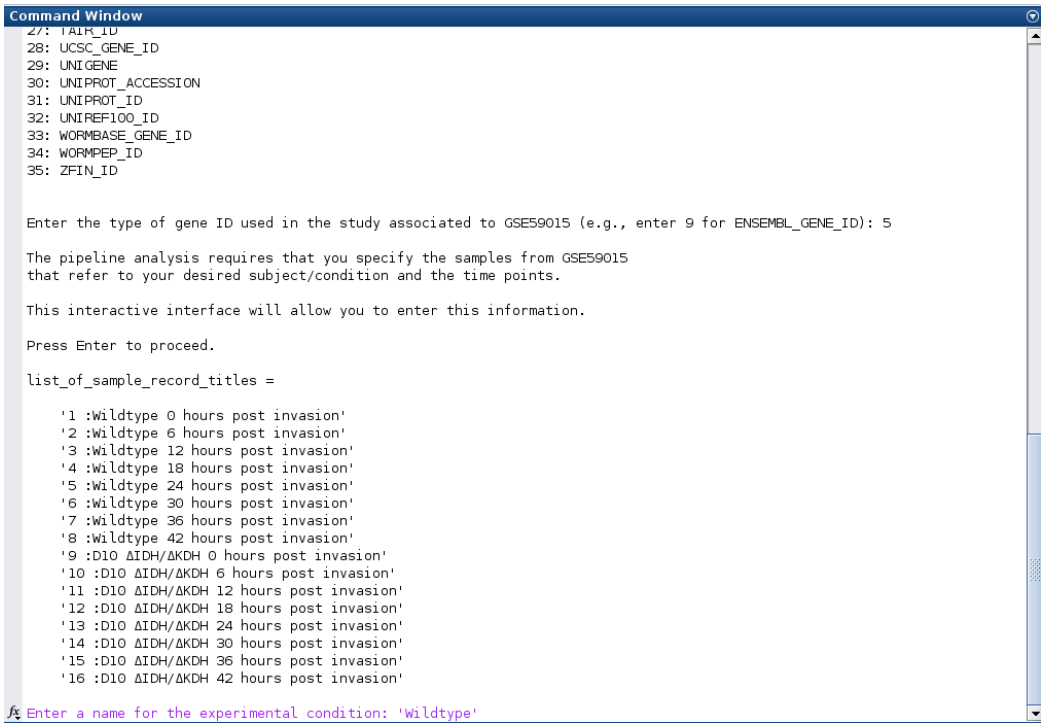


Figure 4: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.



```
Command Window
27: TAIR_ID
28: UCSC_GENE_ID
29: UNIGENE
30: UNIPROT_ACCESSION
31: UNIPROT_ID
32: UNIREF100_ID
33: WORMBASE_GENE_ID
34: WORMPEP_ID
35: ZFIN_ID


Enter the type of gene ID used in the study associated to GSE59015 (e.g., enter 9 for ENSEMBL_GENE_ID): 5

The pipeline analysis requires that you specify the samples from GSE59015
that refer to your desired subject/condition and the time points.

This interactive interface will allow you to enter this information.

Press Enter to proceed.

list_of_sample_record_titles =

    '1 :Wildtype 0 hours post invasion'
    '2 :Wildtype 6 hours post invasion'
    '3 :Wildtype 12 hours post invasion'
    '4 :Wildtype 18 hours post invasion'
    '5 :Wildtype 24 hours post invasion'
    '6 :Wildtype 30 hours post invasion'
    '7 :Wildtype 36 hours post invasion'
    '8 :Wildtype 42 hours post invasion'
    '9 :D10 ΔIDH/ΔKDH 0 hours post invasion'
    '10 :D10 ΔIDH/ΔKDH 6 hours post invasion'
    '11 :D10 ΔIDH/ΔKDH 12 hours post invasion'
    '12 :D10 ΔIDH/ΔKDH 18 hours post invasion'
    '13 :D10 ΔIDH/ΔKDH 24 hours post invasion'
    '14 :D10 ΔIDH/ΔKDH 30 hours post invasion'
    '15 :D10 ΔIDH/ΔKDH 36 hours post invasion'
    '16 :D10 ΔIDH/ΔKDH 42 hours post invasion'

fx Enter a name for the experimental condition: 'Wildtype'
```

Figure 5: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

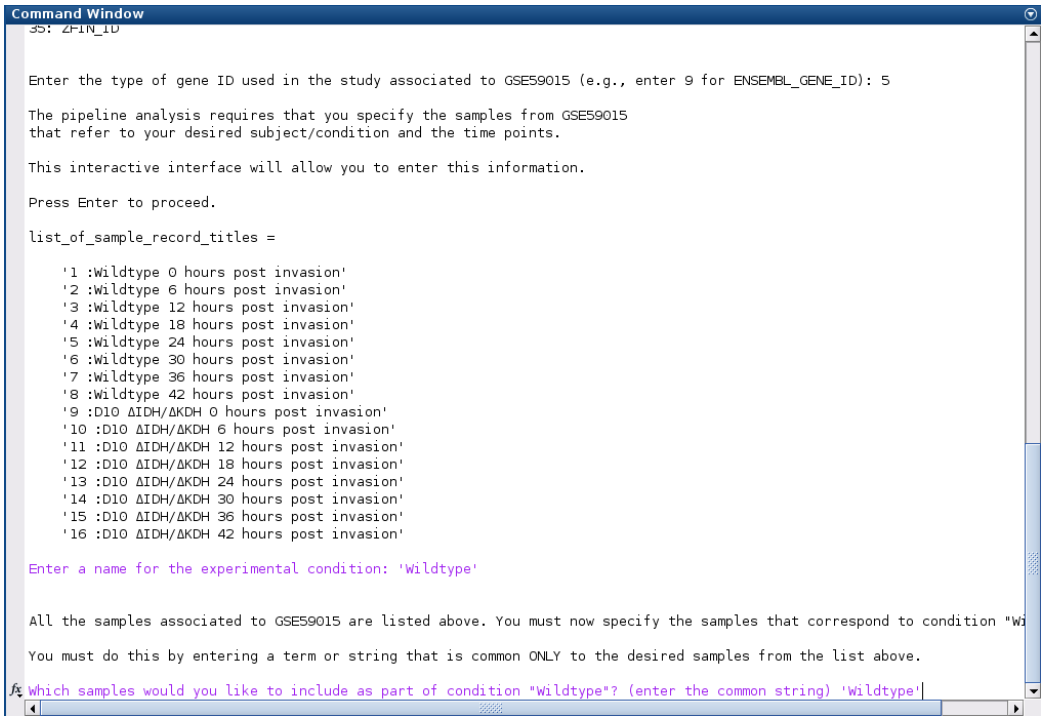


Figure 6: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

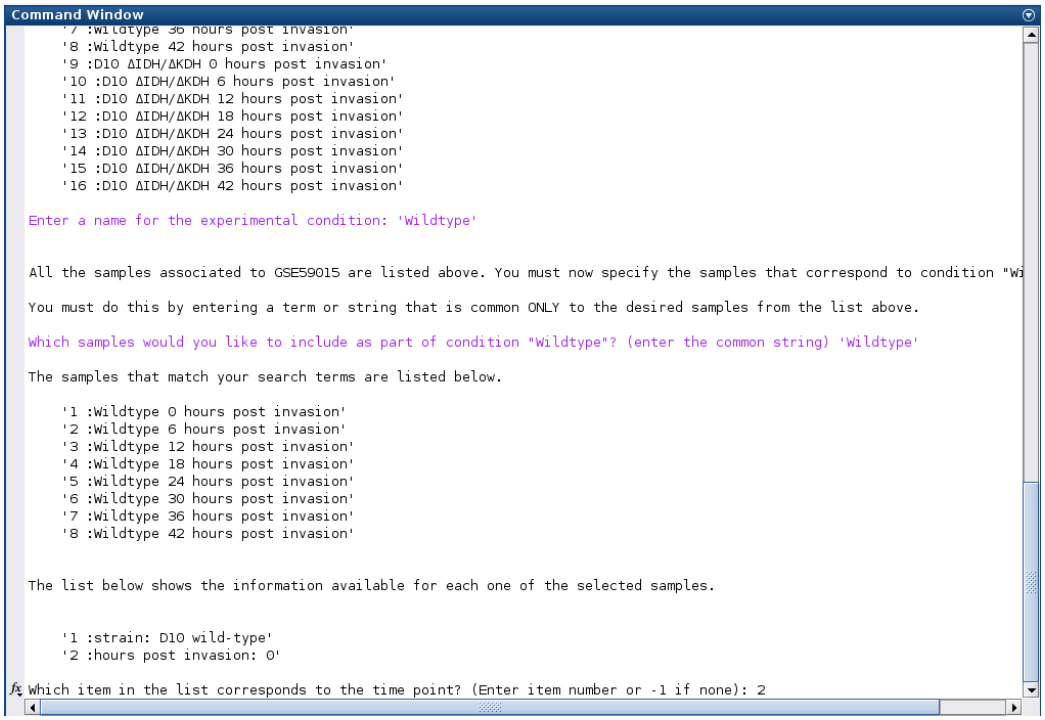
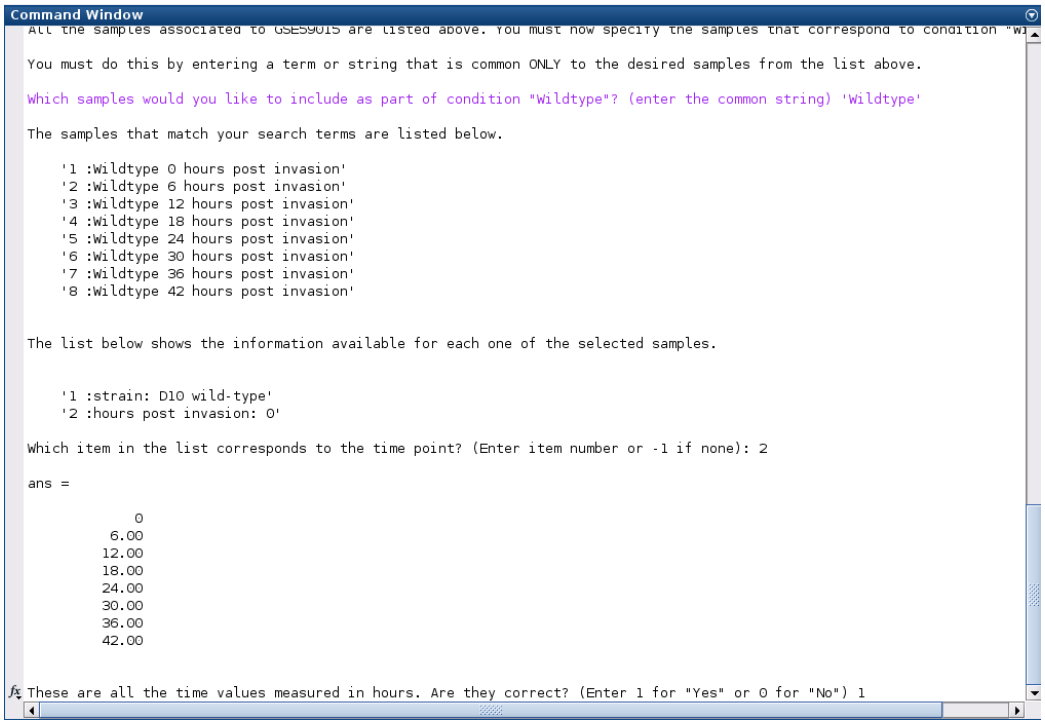
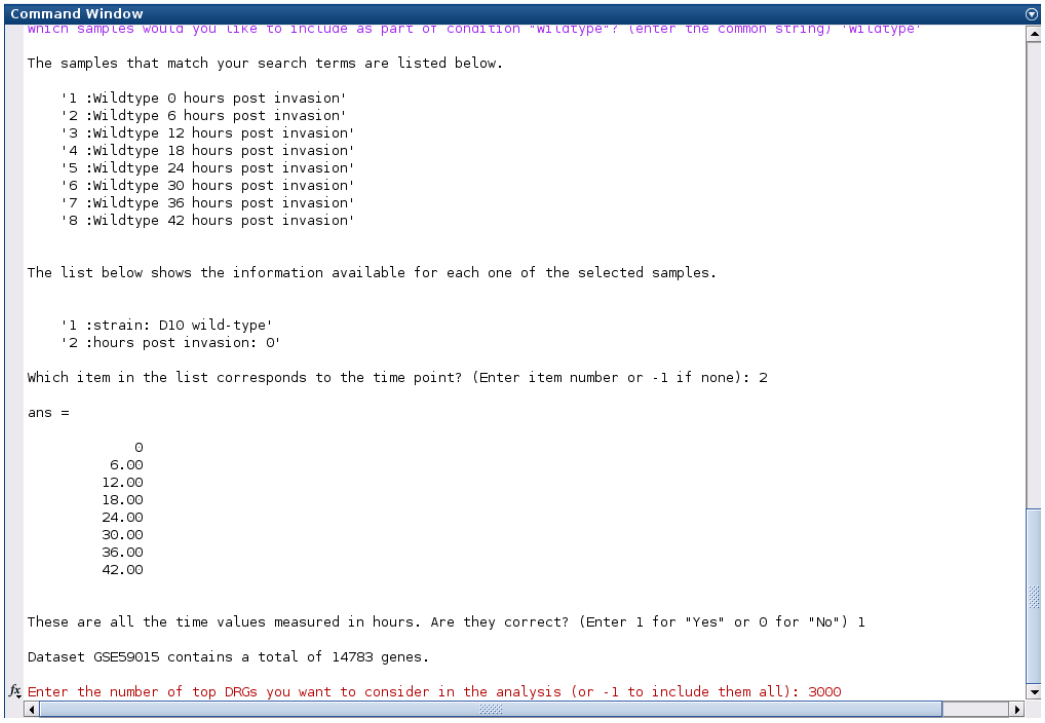


Figure 7: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.



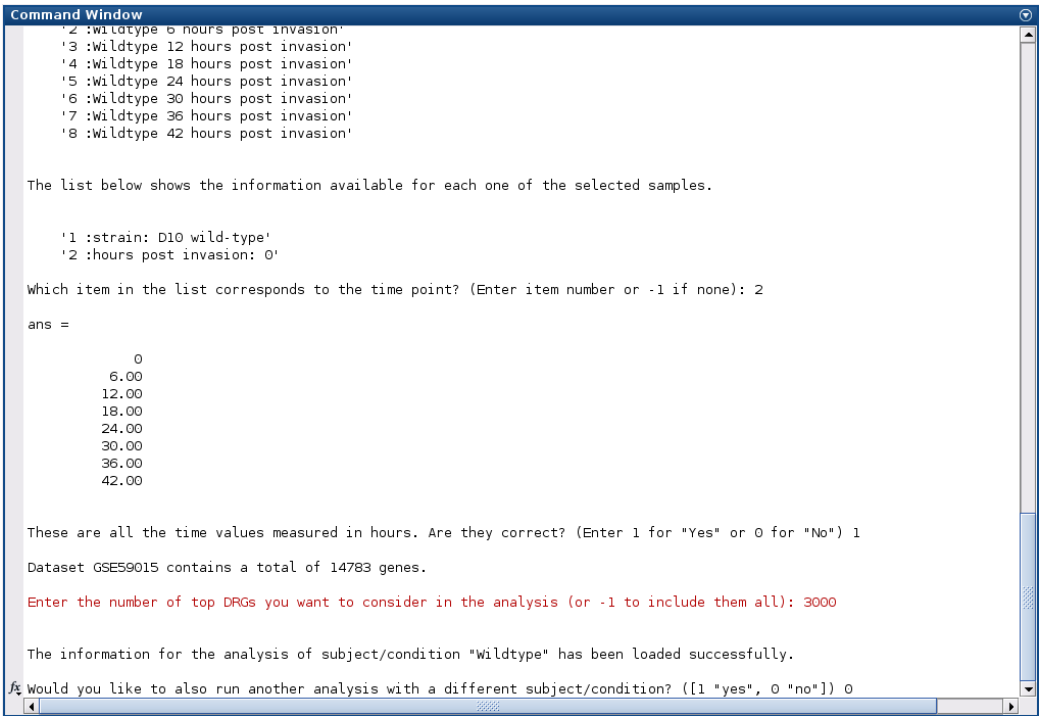
```
Command Window
All the samples associated to GSE59015 are listed above. You must now specify the samples that correspond to condition "wi

You must do this by entering a term or string that is common ONLY to the desired samples from the list above.

Which samples would you like to include as part of condition "Wildtype"? (enter the common string) 'Wildtype'

The samples that match your search terms are listed below.

    '1 :Wildtype 0 hours post invasion'
    '2 :Wildtype 6 hours post invasion'
    '3 :Wildtype 12 hours post invasion'
    '4 :Wildtype 18 hours post invasion'
    '5 :Wildtype 24 hours post invasion'
    '6 :Wildtype 30 hours post invasion'
    '7 :Wildtype 36 hours post invasion'
    '8 :Wildtype 42 hours post invasion'


The list below shows the information available for each one of the selected samples.


    '1 :strain: D10 wild-type'
    '2 :hours post invasion: 0'

Which item in the list corresponds to the time point? (Enter item number or -1 if none): 2

ans =

             0
          6.00
         12.00
         18.00
         24.00
         30.00
         36.00
         42.00


These are all the time values measured in hours. Are they correct? (Enter 1 for "Yes" or 0 for "No") 1
```

Figure 8: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.



```
Command Window
Which samples would you like to include as part of condition "Wildtype"? (enter the common string) 'Wildtype'

The samples that match your search terms are listed below.

    '1 :Wildtype 0 hours post invasion'
    '2 :Wildtype 6 hours post invasion'
    '3 :Wildtype 12 hours post invasion'
    '4 :Wildtype 18 hours post invasion'
    '5 :Wildtype 24 hours post invasion'
    '6 :Wildtype 30 hours post invasion'
    '7 :Wildtype 36 hours post invasion'
    '8 :Wildtype 42 hours post invasion'


The list below shows the information available for each one of the selected samples.


    '1 :strain: D10 wild-type'
    '2 :hours post invasion: 0'

Which item in the list corresponds to the time point? (Enter item number or -1 if none): 2

ans =

             0
          6.00
         12.00
         18.00
         24.00
         30.00
         36.00
         42.00


These are all the time values measured in hours. Are they correct? (Enter 1 for "Yes" or 0 for "No") 1

Dataset GSE59015 contains a total of 14783 genes.

Enter the number of top DRGs you want to consider in the analysis (or -1 to include them all): 3000
```

Figure 9: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.



```
Command Window
    '2 :Wildtype 6 hours post invasion'
    '3 :Wildtype 12 hours post invasion'
    '4 :Wildtype 18 hours post invasion'
    '5 :Wildtype 24 hours post invasion'
    '6 :Wildtype 30 hours post invasion'
    '7 :Wildtype 36 hours post invasion'
    '8 :Wildtype 42 hours post invasion'


The list below shows the information available for each one of the selected samples.


    '1 :strain: D10 wild-type'
    '2 :hours post invasion: 0'

Which item in the list corresponds to the time point? (Enter item number or -1 if none): 2

ans =

             0
          6.00
         12.00
         18.00
         24.00
         30.00
         36.00
         42.00


These are all the time values measured in hours. Are they correct? (Enter 1 for "Yes" or 0 for "No") 1

Dataset GSE59015 contains a total of 14783 genes.

Enter the number of top DRGs you want to consider in the analysis (or -1 to include them all): 3000


The information for the analysis of subject/condition "Wildtype" has been loaded successfully.

Would you like to also run another analysis with a different subject/condition? ([1 "yes", 0 "no"]) 0
```

Figure 10: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

```
0 hours,GSM1424445
6 hours,GSM1424446
12 hours,GSM1424447
18 hours,GSM1424448
24 hours,GSM1424449
30 hours,GSM1424450
36 hours,GSM1424451
42 hours,GSM1424452
```

Figure 11: Executing the `create_input_files.m` script in order to create the condition file(s) to be used in the pipeline analysis.

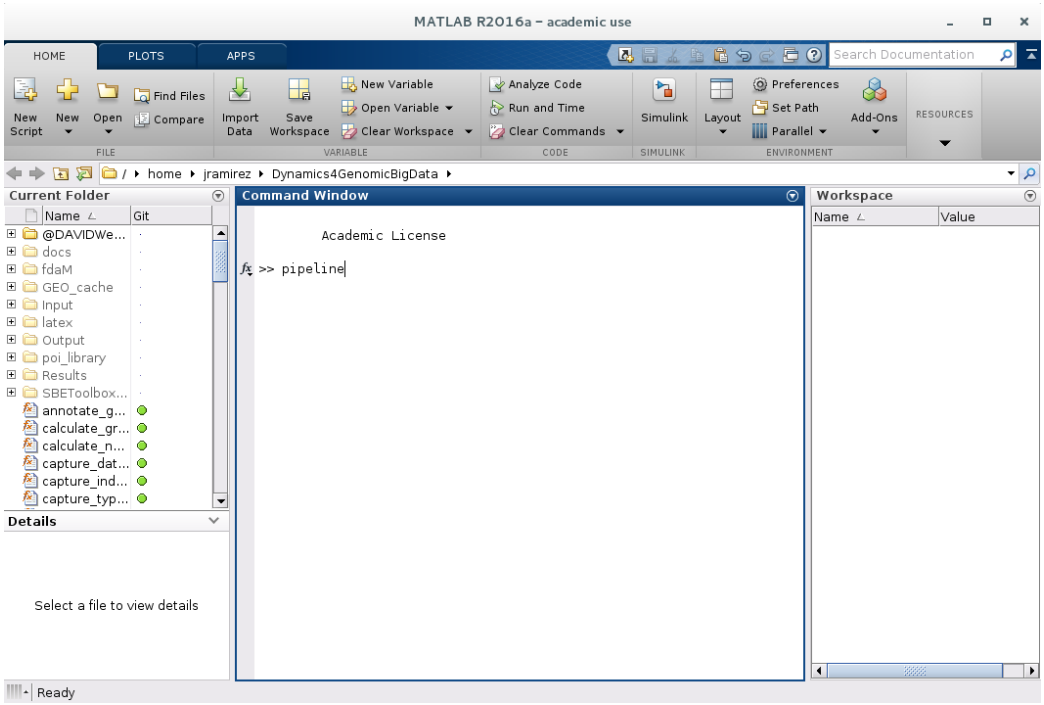


Figure 12: Executing the `pipeline.m` script in order to run the pipeline analysis on a set of one or more experimental conditions.