

Working Title

Yohan Lefol, Marie Terrien, Alexis Dupis, and Ugo Vidal

December 18, 2019

1 Introduction

The definition between canonical and non-canonical pathways is not very clear, not does it have a clear cut definition. We describe canonical pathways as pathways that are well defined in the science field, such as glycolysis. In contrast to this, we define non-canonical pathways as being one or several genes that have an alternate (non-canonical) function which causes them to be a part of a non-standard pathway.

Our team has thus set out to find these non-canonical genes through bibliographic research and create a tool that will allow users to identify which genes of a human gene set may be genes that have non-canonical functions.

Additionally, this tool has several other applications. The tool allows users to perform a different gene expression analysis using a negative binomial distribution. In order to do so, the tool implements the DESeq2 R package (Love, Huber, & Anders, 2014). This aspect of the tool can only be done using COUNT data from RNA sequencing. Initially this method was selected as it allows users to exploit data from the TCGA (The Cancer Genome Atlas) database, which amasses large quantities of RNAseq files for patients of various cancers.

By pairing this package with our tool, we allow users to identify significant genes from differential analyses and we then look for potential non-canonical genes from the list of isolated significant genes. Additionally, users will have the option to customize MA plots and volcano plots to best suit their publishing needs.

When the tool identifies potential non-canonical genes, the user will be provided with a brief explanation of the canonical pathway that this gene affects along with it's location in the human body, and the same information is given for non-canonical pathways. The tool will also provide any references that have been associated to these genes within our database.

2 Installation

2.1 quick description

This tool is split into two components, one is the application and processing power, this functional aspect of the tool is downloaded on the users computer and run as a program. The program's interface will then allow users to connect to a database, this database is hosted on the Cellomet website. The database contains all of the information pertaining to non-canonical genes that we have acquired. Once connected to the database, the tool will interact with the users files as indicated by the user and utilize the processing power of the

users computer to run the various analyses. This choice was made to prevent a surcharge of a servers processing power, however it is not without consequence. The DESeq2 package can be heavy on processing power, this is of course dependent on the size of the files given to it. The processing power of a given computer will affect the speed at which analyses are performed. Lastly, as a database connection is required, users must be connected to the internet in order to use the non-canonical analysis aspect of this tool.

2.2 Downloading the tool

Difference MAC and Windows

2.3 Connecting to the database

Just a quick explanation

3 How to use tool

Will be a quick overview of the tool without going into details, simply explain how to navigate it and what not.

4 Non-canonical analysis

This simple analysis is the innovative side of this tool, however it is very simple to use. If a user runs a DESeq2 analysis, a non-canonical analysis will automatically be performed on the isolated significant genes. Alternatively, a user can input a gene list file which will be analyzed. The gene list file is very simple in format, it can be seen in Figure 1. This file can be created as a text file, but must be converted into a csv file before using it for the analysis. This is very simple to do, one must just rename the file from .txt to .csv.

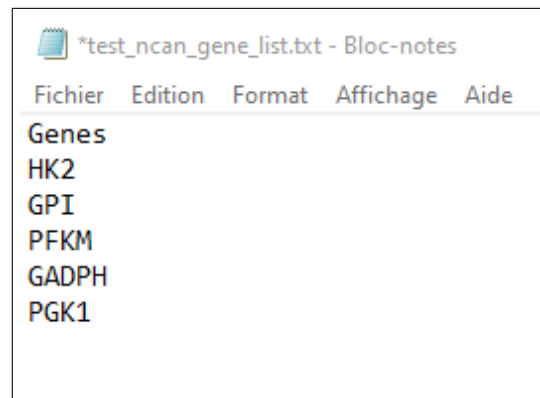


Figure 1: An example of a custom gene list

The results for this analysis, be it via a custom list or a DESeq2 analysis, is always presented as three separate text files.

1. non_canonic_results.txt

2. canonic_results.txt

3. references.txt

These files each contain the gene name and gene symbol of identified non-canonical genes in the dataset provided by the user along with information regarding those genes. The information varies based on which file it is contained in.

These text files have been designed to be able to be read by microsoft excel. To do so, a user should open the text file with microsoft excel and select tabs (\t) as the separator. This will ensure proper delimitation of the various information contained in the text files. Also, these results can be viewed within the tool in the results tab. These results are only p69*esent after and analysis is done and will not be present if the application is closed or if another analysis is run afterwards. However, files are automatically saved on the users computer, it is not necessary to go through the results tab to save this information.

5 DESeq2

The DESeq2 analysis is a powerfull tool allowing a user to perform differential analyses on large quantities of patient files all while customizing the two conditions of the differential analysis. This type of analysis is very picky about the data type and format used. These are the requirements of the data.

- The data must be in COUNTS format
- The data format must be as seen in Figure 2. This means that the data must be in csv format, have the gene names as the row headers and patient or sample names as column headers. CSV files can sometimes be difficult to understand, thus a brief explanation will be given. CSV stands for comma separated vector, this means that a vector (list of elements/values) are separated by commas. One can see a comma as a column delimitation, in the case of Figure 2, if we look at the TSPAN6 row, we see numerous values separated by commas, above this row we see two patient IDs as only two commas are present. This means that the first patient ID would have a value of 4280 for the TSPAN6 gene and the second patient would have a value of 5096 for the same gene. The values following these are values for other patients that are not seen in the figure.

```
Gene,01daeea0-2a6d-450e-b541-5e403e998637,05a38163-a505-4f97-b2a3-7430c49b14f4,
TSPAN6,4280,5096,1949,2711,1669,3992,5499,5770,794,4039,4943,4236,2
TNMD,4,8,2,3,2,7,14,5,2,26,17,58,4,8,3,1933,2,27,31,2,10,1
DEML1,1293,1019,440,1088,844,1271,1137,1546,440,1549,1319,1374,753,
SCYL3,728,563,127,458,650,542,651,742,161,596,811,453,547,551,519
Clorf112,565,342,123,223,480,517,500,745,151,478,477,184,270,378,
```

Figure 2: Data format for a DESeq2 file.

- Gene names must be in gene symbol format If this section is not respected, a complete DESeq2 analysis will still be performed, however no non-canonic genes will be identified as the data base finds non-canonic genes via their gene symbols. As an example, the data

base will not recognize Hexokinase2 as being a non-canonic gene, but it will recognize HK2.

- Two files must be used, each representing a different condition This is the basis of the differential analysis, DESeq2 will take in two files with n number of patient genomes. However these files must be split into their respective categories before being inputted into the analysis tool. As an example, one would have a file with all cancer patients between the age of 10 and 50, and a second file for patients of the same cancer but that are between the ages of 70 and 99, This would result in a differential analysis of young vs old.

With the two files in the proper format, the user must only select the folder in which he will want to store the results of the analysis. The user will also be able to name his two conditions which will be automatically added to the names of the results created by the analysis. Additionally, a user can choose to create standard MA and Volcano plots, this said, a user will also be able to create custom MA and Volcano plots after a DESeq2 analysis has been finished see section 6 of this guide. Before a user can launch an analysis, he will have to connect to the database, this can be done by following the simple steps in the 'Connect to Database' tab of the tool, also talked about in section 2.3 of this guide. Once the analysis has been launched, the user will be prompted to wait until the analysis is over. A notification in the form of a pop up message will be given to user user once the analysis has either finished or if an error has occurred. While the analysis is running, a user should not use the tool and let it work. The user can however continue to use the computer provided the processing power of the computer can handle several things at once.

6 Customizable figures

References

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12), 550.