User Guide

CelloMap

Yohan Lefol   Alexis Dupis   Ugo Vidal   Marie Terrien

January 22, 2020

Contents

1 Introduction

The non-canonical analysis tool (CelloApp) is a tool with several uses. It's main feature is a non-canonical gene analysis from a gene list. Yet this tool also boasts several other interesting functionalities such a differential expression analysis, customizable MA and Volcano plots which in turn allow to obtain significant gene sets according to a users custom parameters.Additionally, this tool is capable of converting gene symbols to KEGG identifier (type hsa), see section 5 for more information. This guide will go over each functionality and explain how to best use CelloApp.

As can be seen in Figure 1, the application is split into several tabs. The information tab shows two pdf files, the first being this one, and the second being a detailed explanation of the tool and it's components. The last tab 'Citation' is a simple tab which gives the citation information for the use of this tool.This guide will now go over the remaining tabs and explain their uses and how to maneuver them.
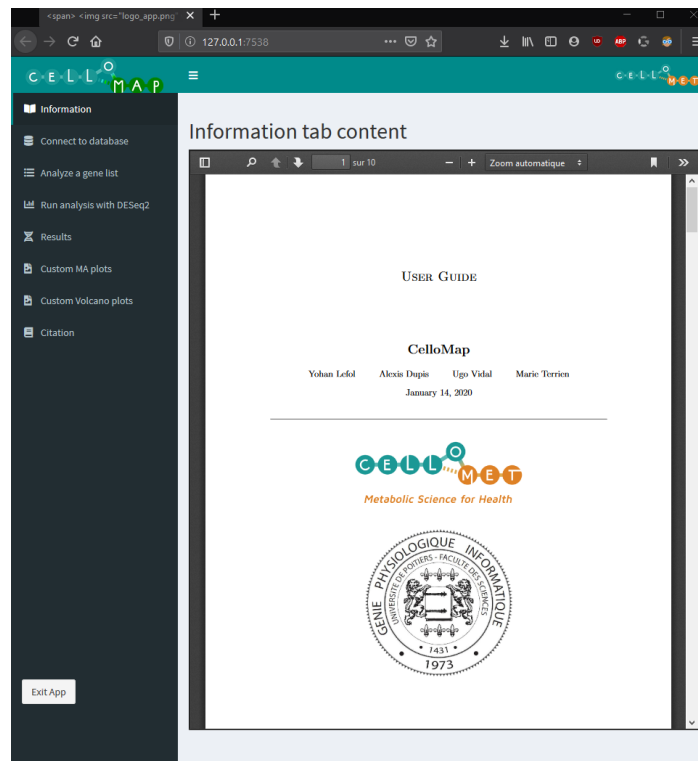

Figure 1: The main tab for the application.

2 Installation

HOW DOES THE INSTALLATION WORK.

3 Connection to database

The identification of non-canonical genes is the main aspect of this tool. As the list of non-canonical genes is increasing as discoveries are being made, we could not 'hard code' a list of genes in the program as we wanted to ensure that the search list is being updated without having users re-download the application for each update. Therefore we implemented a database that can be updated by the owners of the tool (Cellomet) and that can be accessed remotely through this application. In order to establish this connection, a user needs to fill in the small questionnaire as seen in Figure 2, and click the 'Connect to database' button. We created the small questionnaire in order to see the main users of this tool as we intend to keep developing new functionalities and we want these functionalities to be catered to the types of study that require it the most.
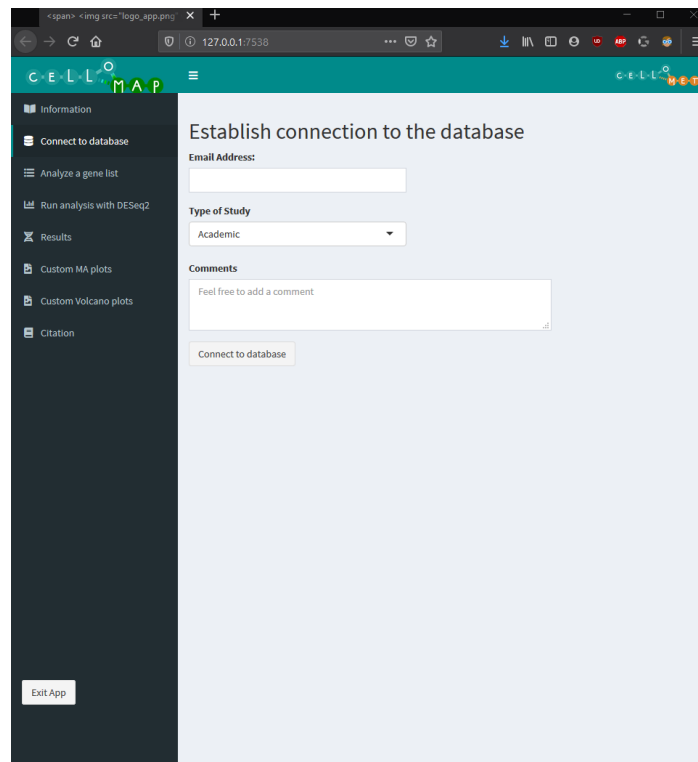

Figure 2: The connect to database tab of the application.

The database connection is only necessary if a non-canonical gene analysis is to be performed, either individually or as part of a DESeq2 analysis. If this is not the users intentions, he does not need to establish a connection to the database.

4 DESeq2 analysis

DESeq2 is a powerfull tool which allows users to perform differential analyses using RNAseq 'COUNTS' data (Love, Huber, & Anders, 2014). However this tool requires a very specific data format. This part of the guide will go over the data format necessary to run a DESeq2

analysis, the added parameters specific to this application, and the results that are generated.

4.1 Tab layout

As seen in Figure 3, there are several requirements for the the DESeq2 analysis. First we observe that two csv files are needed, these will be the two files representing the two conditions that will be analyzed. We then observe a small check box which asks if the user want to perform a non-canonical gene analysis. If this box remains checked, the user will have to be connected to the database and the non-canonical gene analysis will be run on the list of significant genes isolated by the DESeq2 analysis (explained below). A user can also select a directory, this will be the folder in which the several DESeq2 results will be saved. There are then two text boxes which can be customized to be whatever text a user whishes, the content of these text boxes is used to name the files that will be produced by the analysis. Condition 1 represents File 1 and Condition 2 represents File 2. Lastly there are two check boxes that check if the user would like to create the standard MA and Volcano plots. If checked, two types of each plot will be made, one type with text, and the other without text. The plots generated will be with standard values, however a user will be able to take the differential expression data obtained from the DESeq2 analysis and make his own customized MA and Volcano plots with the custom plot tabs, see section 7 for more details.

Figure 3: The main screen for the DESeq2 analysis tab.

4.2 Data format

As previously stated, DESeq2 requires a rather specific data format. We will be using Figure 4 to explain the format. This format is presented as a CSV file, meaning a comma separated vector, or more simply put, a file in which the different elements/values are separated by a comma. In order to better explain the data format and what a csv file really is, an excel representation of Figure 4 has been created and is shown in Figure 5

```
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
DPM1,1293 ,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
C1orf112,565 ,342 ,123 ,223 ,480 ,517 ,500 ,745 ,151 ,478 ,477 ,184 ,270 ,378 ,.
```

Figure 4: The data format for a DESeq2 analysis.

When comparing both Figure 4 and Figure 5 we can clearly observe that the commas create the 'separation' of the values. Now we will explain what these values are. The first column represents the genes, it is important to note that these genes are in gene symbols and not another type of format such as Ensembl. DESeq2 will perform the analysis regardless of the type of gene name used however **a non-canonical gene analysis can only be performed with gene symbols**. As such it is advised to use gene symbols in the data in order to fully benefit from this tool. Next we have the other columns which represents the different sequencings. In this example, the second column shows the RNAseq 'COUNTS' results of patient #1 and the second column shows the same but for patient #2.

	A	B	C
1	Gene	01daeea0-2a6d-450e-b541-5e403e998637	05a38163-a505-4f97-b2a3-7430c49b14f4
2	TSPAN6	4280	5096
3	TNMD	4	8
4	DPM1	1293	1019
5	SCYL3	728	563
6	C1orf112	565	342
7	FGR	762	1027
8	CFH	480	1037

Figure 5: A simplified view of the DESeq2 data format.

That's it, that is the data format required. As previously mentioned, two files are necessary, i.e: one file per condition. An easy example to explain this would be the comparative study of young vs old pulmonary cancer patients. In order to do such an analysis, we would have file #1 contain all of the patients in the 'young' category and file #2 would contain all the patients of the 'old' category. Of course both files must respect the data format.

4.3 launching the analysis

Once all necessary files have been uploaded, a user can click the 'launch analysis' button, at which point a pop-up will appear indicating the the analysis is being performed. This pop-up will lock the application preventing any further use, this is done to avoid the application from crashing if too many actions are performed at once. It is important to note that a DESeq2 analysis can be very quick just as it may take several hours. The length of the analysis depends on the size of the files used as well as the computing strength of the users computer. If a user is using a old computer, it may be advised to let the computer run the analysis without doing anything else on the side. If an error occurs during the analysis, another pop-up will appear stating that there was an error and that the user can read about the error in the error_log.txt that was created. Using the information from the error log, the user should consult section 8 in order to solve the issue. Once the analysis is done and if no error has occurred, the pop-up message will be replaced by a different pop-up message indicating that the analysis is done, the pop-up message will also remind the user in which files he has stored the results of the analysis.

4.4 Results obtained

Several results can be obtained from this analysis, some are obtained regardless of user input, others will only appear if a user has asked for those results to be produced.

4.4.1 Main results

A DESeq2 analysis generates three standard results:

- diffexpr_results condition_1 vs condition_1 .csv
  A differential expression file

- condition_1vscondition_1_RESULTS_VOLCANO.csv
  A differential expression file for significant genes

- gene_list_ condition_1 vs condition_1 _Most_Significant.txt
  A significant gene list

It is important to note that the significant genes were obtained by filtering the main differential expression file for standard significance parameters while using a volcano plot to visualize the significant genes. The standard significance algorithm is the following:

$$padj < 0.05 \ \& \ abs(log2FoldChange) > 2$$

This equation indicates that genes which have a Padj value below 0.05 and a log2FoldChange above 2 or below -2 are considered to be significant. A user can also generate his own list of significant genes by modifying these parameters. This is further explained in section 7 of this guide.

4.4.2 Figures generated

If a user has selected that he wanted standard MA and Volcano plots, there will be four png files in the results folder; Two MA plots, one with text and the other without text, the same

applies for the volcano plot. The plots with texts are much larger in order to accommodate for the additional text besides the points representing significant genes. If a user wishes to modify these plots and create his own set, it can be easily be done in the custom plot tabs, further explained in section 7 of this guide.

4.4.3 Non-canonical gene analysis

If a user had selected that he wanted a non-canonical gene analysis to be performed, there will three extra text files in the results folder. That is, if any non-canonical genes were found within the significant genes list. The files are the following:

- non_canonic_results.txt
  This file will contain the gene symbols and gene names that were found to be non-canonic genes within the significant gene list that was analyzed. The file also contains the non-canonical action and localization of the genes identified

- canonic_results.txt
  Similarly to non-canonic results, however this once contains the canonic action and localization.

- references.txt
  Any and all references that were used to identify non-canonical genes and their actions.

The results were saved as text files that can easily be read by microsoft excel, or any similar program. Additionally, results can be viewed immediately in the results tab of the application, see section 6 of this guide for more information.

5 Gene list analysis

This tab allows users to perform a non-canonical gene analysis and a gene symbol to hsa identifier conversion with a simple gene list file. The only obligation that this file must respect is that the genes must be in a column with a column header names 'Genes'. If the file entered is correct, a table will appear showing that specific column as seen in Figure 6. If this does not occur, a text will appear, stating that a different file should be chosen.

Considering that the requirement for this analysis is a single column carrying the name of 'Genes', possible files can be the significant gene list obtained from the DESeq2 analysis, or the significant differential expression results obtained from either the DESeq2 analysis or the downloading of the data from custom plots in section 7.

5.1 KEGG Mapper

By checking the 'Find hsa conversions' checkbox in the CelloMap tool, a conversion script will automatically be run along with a non-canonical gene analysis. This conversion will allow users to run a KEGG mapping analysis using the file generated from the conversion. KEGG Mapper is a mapping tool that will search and find any pathways associated to a list of identifiers provided (Kanehisa & Sato, 2019). It will then allow the user to view every pathway found while highlighting the gene that was in the original list. To perform the analysis, follow these steps, also seen in Figure 7
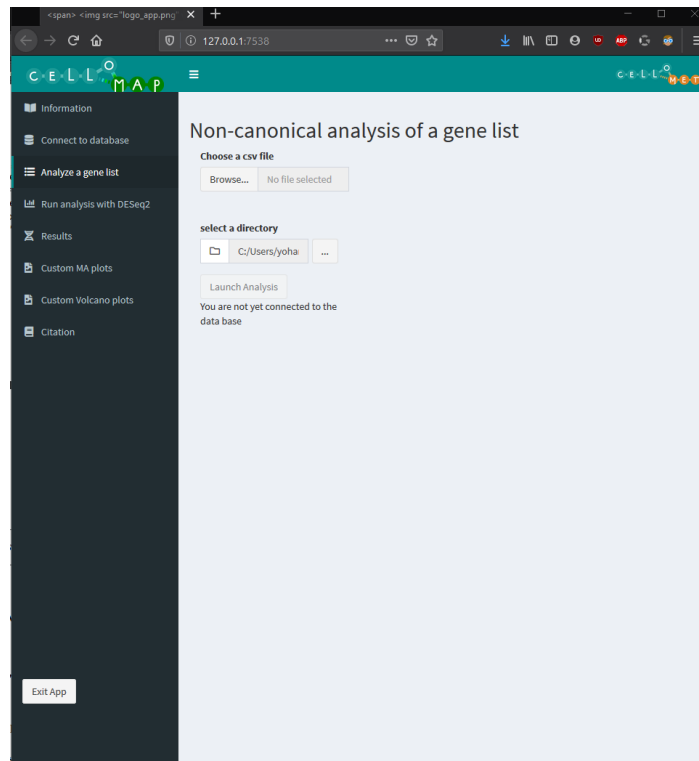
Figure 6: The gene list analysis tab once a correct file has been uploaded.

1. Go to the website: `https://www.kegg.jp/kegg/tool/map_pathway2.html`

2. Make sure that the search mode is **Organism-specific** with the letter **hsa** in the text box

3. Upload the file created by CelloMap, the file will be called **hsa_ID.txt**

4. Launch the analysis and wait a few seconds/minutes

5.2 Metabolic GEne RApid Visualizer (MERAV)

An alternative to the KEGG Mapper is Metabolic GEne RApid Visualizer (MERAV), this tool was designed to analyze **human** gene expression across a large variety of arrays. All of the arrays were normalized together to generate a gene expression database composed of several types of human tissue (Shaul et al., 2015). This tool offers two types of searches:

1. Search expression levels in one or several genes
   This search will provide the user with the ability to search the database for the expression of a given gene(s).

2. Search all genes expression levels in one or more cell types
   This search will provide the user with the ability to search the database for the expression of a given array.
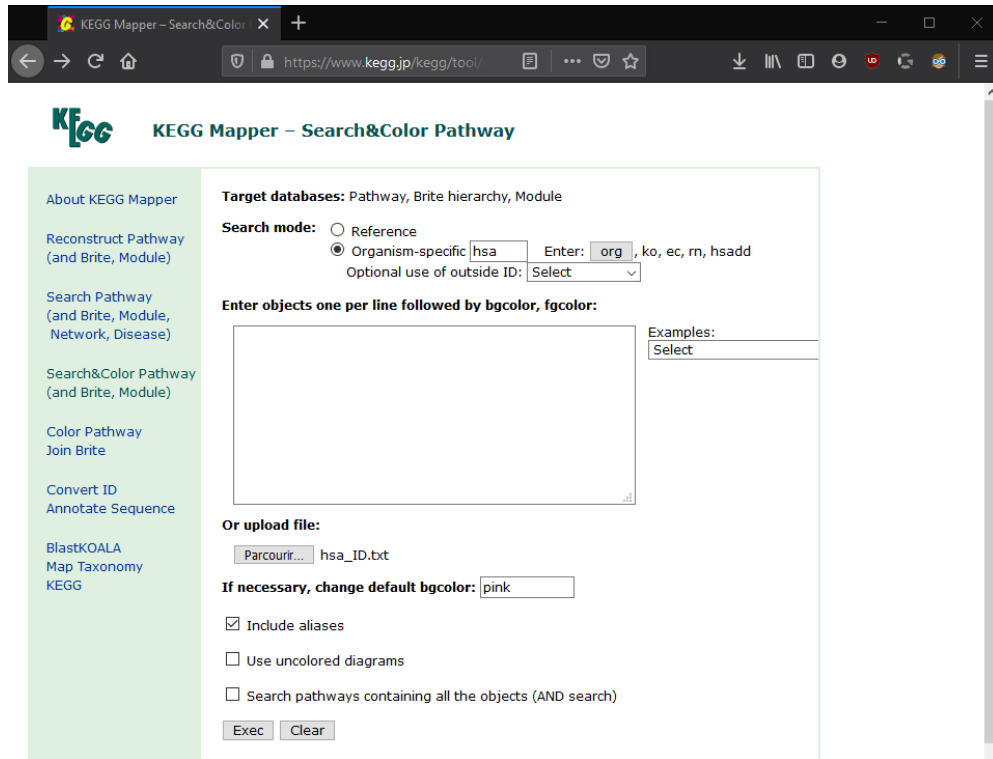
Figure 7: The set-up for the KEGG Mapper tool

CelloMap creates a significant gene file that can be used with the first type of search, the one relating to finding gene expressions. To perform the search, follow these steps, also seen in Figure 8.

1. Go to the website: `http://merav.wi.mit.edu/`

2. Click the 'Search expression levels...' link.

3. Load the gene search box
   To do this, copy paste all the genes from the significant gene file generated from DESeq2 or the one inputted in the gene list analysis. The main thing to watch out for is to not include the column name 'Genes'. If it is included, it's not a problem, it will simply be considered a gene symbol and will not be found by the search, this will not prevent the tool from searching for the other genes.

4. Set the parameters to your liking

5. If help is required, follow this link: `http://merav.wi.mit.edu/help/help.html`

5.3   Non-canonical gene analysis

The non-canonical analysis will cross check the provided gene list with our database, and depending on if it finds non-canonical genes within the gene list, three text files will be created.
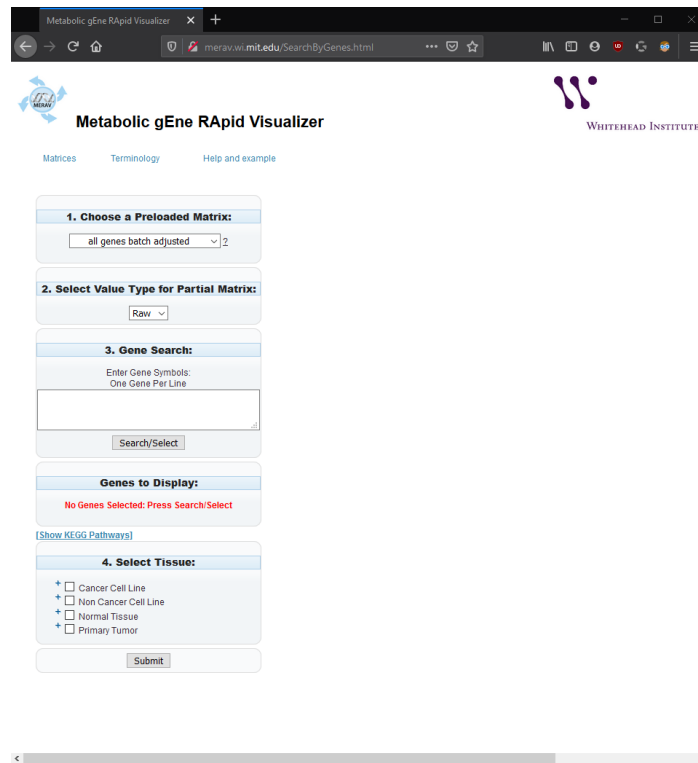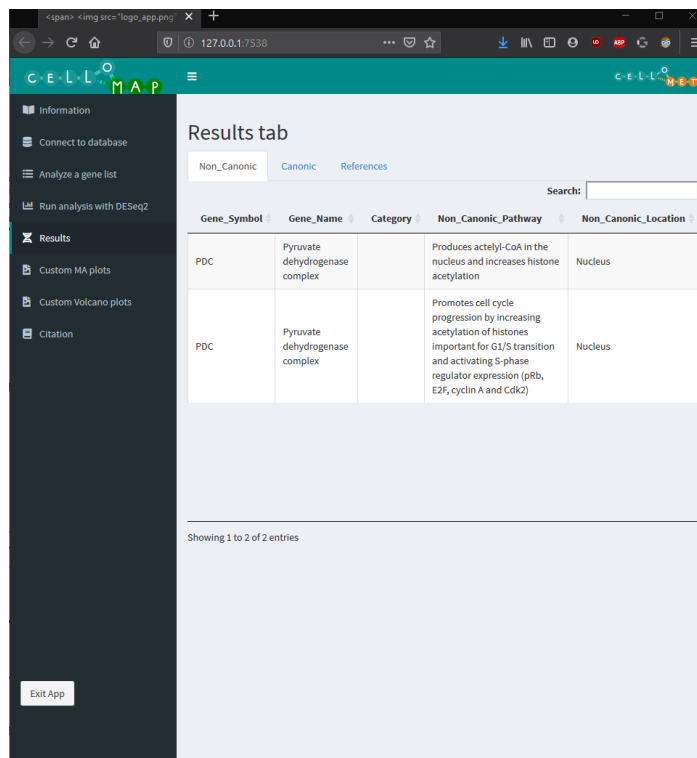
Figure 8: The set-up for the MERAV search tool

1. canonic_results.txt

   This file will contain the non-canonical gene symbols and names along with their **canonical** function and location.

2. non_canonical_results.txt

   This file will contain the non-canonical gene symbols and names along with their **non-canonical** function and location.

3. references.txt

   This file will contain the gene symbol and name of non-canonical genes and the references that were used to determine their canonical and non-canonical functions within this database.

Each text file was created with the intention to be read by Microsoft excel. The text files need to be opened in excel and the separator/deliminator must be declared as 'Tab'. Once opened, it is recommended to use the 'wrap text' and 'AutoFit Row Height' in the alignment and cell format sections of excel. This will ensure that the table is legible in excel.

6 Results

The results tab, seen in Figure 9, shows the results of a non-canonical gene analysis within the application itself. There isn't much to be said about this tab, it is only for visualization. Every registered non-canonical genes are registered in a database hosted on Cellomet.com which can be accessed, updated and generally modified by the owner of Cellomet. For any questions regarding the non-canonical genes, their function, localization etc. Please contact Cellomet via their website.

Figure 9: The results tab after a non canonical gene list analysis.

7   Custom MA and Volcano plots

This application allows users to create their own custom figures, and from these figures, significant gene data can be downloaded. There are two possible figures, MA plots and Volcano plots, each plot type has it's own tab, however they behave very similarly, the volcano plot tab can be seen in Figure 10. The main thing to know about these tabs is that a differential expression file is required, like the one obtained from DESeq2 analyses. The columns required are the 'Genes' column, the 'baseMean' column, the log2FoldChange' column, the 'pvalue' column, and the 'padj' column, each written the same way as they are written in the guide. The file must also be a csv file. Aside that, each element of the different tabs is self explanatory. After each parameter modification, the create plot button needs to be clicked again. When a plot is loading, the rest of the application is locked to prevent the application from crashing due to a potential large number of buffered actions. On every button click the plot must be re loaded from scratch, this factor should be taken into consideration if a user is using a large differential expression files.

Figure 10: The tab to create a custom volcano plot.

8 Common errors

1. **Failed to connect to database**
   This may occur from time to time, the ideal is to wait and try again later. The database we are using can have a maximum of four connections at a time, as such, CelloMap was created in a way that it utilizes the connection as little as possible. As in, it only applies the connection right before it needs it and removes the connection once it is done. This time frame lasts no more than ten seconds. That said, if retrying several times has not solved the issue, it is recommended to contact one CelloMap's developer at the following email address: yohan.lefol@gmail.com.

2. **Error in DESeqDataSet(se, design = design, ignoreRank): design has a single variable, with all samples having the same value. use instead a design of ' 1'. estimateSizeFactors, rlog and the VST can then be used**
   This error may occur when only one condition is named, as well as randomly. It is a simple bug where DESeq2 finds only one variable and thus cannot proceed with its intended process. To correct this issue, simply try again, make sure that two conditions are names and that both csv files are properly uploaded to the application.

References

Kanehisa, M., & Sato, Y. (2019). Kegg mapper for inferring cellular functions from protein sequences. *Protein Science*.

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.

Shaul, Y. D., Yuan, B., Thiru, P., Nutter-Upham, A., McCallum, S., Lanzkron, C., ... Sabatini, D. M. (2015). Merav: a tool for comparing gene expression across human tissues and cell types. *Nucleic acids research*, *44*(D1), D560–D566.