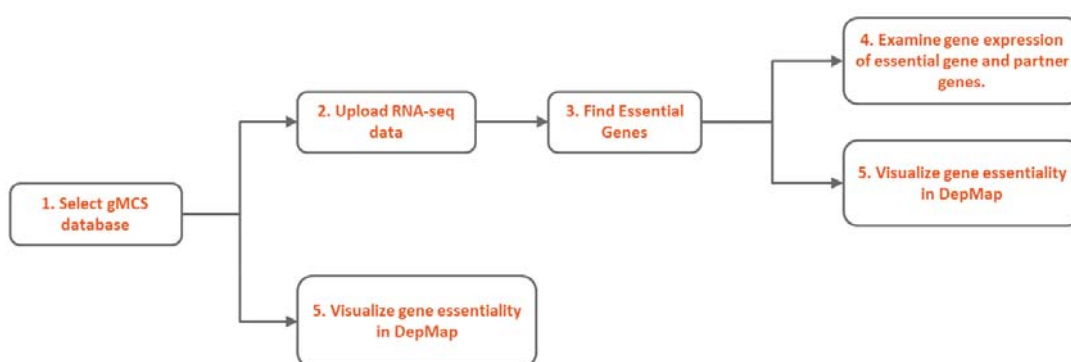


Quick start

gMCStool outputs a list of predicted essential metabolic genes and their companion biomarkers for a given cohort of RNA-seq samples.

The functionalities of gMCStool are divided into 5 different modules. Figure 1 shows the gMCStool' pipeline.



1. **Select gMCS database.** Selection of the database of gMCSs for the analysis. The user can select the metabolic tasks included in the analysis, as well as the growth medium constraints for the biomass task. Additionally, the user can examine the number gMCS associated with each task, download the entire set of selected gMCSs or generate summary figures. Note: the database of gMCSs were calculated from Human-GEM-1.4.0 (ref).
2. **Upload the RNA-seq data.** In this module, RNA-seq data and metadata for the samples can be uploaded in different formats. It includes a summary table that allows the user to check the input data.
3. **Predict Essential Genes.** In this module, gene essentiality analysis for the given samples is performed using our gMCS approach. The user needs to specify the thresholding approach to calculate highly and lowly expressed (ON-OFF) genes. We provide two options: our own

approach, *gmcsTHX*, in which the desired threshold value must be specified, or a previous approach in the literature, *localT2*, in which the user must decide to use the genes involved in the database of gMCSs or all the genes of Human1.

4. **Visualization.** This module is divided into two parts. In the left part, the user can filter genes according to its essentiality across the different sample classes analyzed. In the right part, the resulting essential genes and associated gMCSs are detailed. Selected pair gene-gMCS is plotted in a customizable heatmap and a boxplot which increase the interpretability of the results, providing information about possible biomarkers and functional explanation.
5. **DepMap Analysis.** The user can also visualize the association between the essentiality scores of predicted essential genes, available in the DepMap database (ref), and the summary expression of the partner genes involved in their associated gMCSs from Cancer Cell Line Encyclopedia (ref). Users can select different expression units and gene silencing approaches. Moreover, there are several filters that can be modified to select the subset of cell lines of interest.

Guided example

This is a tutorial to exemplify the use of gMCStool to perform gene essentiality analysis. For illustration, we analyze the metabolic dependencies of a dataset including different B cell subpopulations and the MM samples (ref).

Step 1: Access to gMCStool

We can access to the online version of gMCStool here: <https://biotecnun.unav.es/app/gmcstool>.

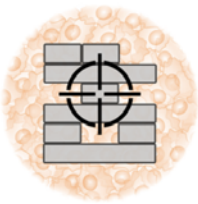
We can also download gMCStool from GitHub: <https://github.com/lvalcarcel/gMCStool>. Once the Git repository is cloned, the Shiny app can be run locally simply by typing:

```
shiny::runApp('./app.R')
```

The code is prepared to adjust the amount of available RAM and number of cores. For large datasets we strongly encourage the use of gMCStool in a local computer rather than the online version.

For any of the 2 options, the user encounters this main presentation panel:

[gMCStool](#) [Overview](#) [1. gMCS database](#) [2. Upload the RNA-seq data](#) [3. Predict Essential Genes](#) [4. Visualization](#) [5. DepMap Analysis](#) [Help](#) [About](#)



gMCStool[©]

"A tool for discovering essential genes in cancer metabolism"

Welcome!

We present gMCStool, a user-friendly web-tool to predict essential genes in the latest reconstruction of the human metabolism, [Human1](#), freely available in [GitHub](#).

We have chosen the game 'jenga' as the icon of our tool, because the jenga tower falls when you take out the most vulnerable part of the structure. In our case, we are searching for the most vulnerable parts of the metabolism of cancer cells in order to disrupt cellular proliferation.

In order to find metabolic vulnerabilities in cancer cells, we employ the concept of genetic *I*-minimal Cut Sets ([gMCSs](#)), a metabolic network-based approach to synthetic lethality, and RNA-seq data.

gMCStool has been developed to achieve the following analysis:

1. Find essential genes for cancer metabolism based on RNA-seq data.
2. Predict putative companion biomarkers for predicted essential genes (biomarkers is the expression of partner genes within target gMCSs).
3. Identify essential task or essential metabolites associated with predicted essential genes.

gMCStool has been developed using R and Shiny.

The databases and source code are available at [GitHub](#)

Start!

Go to gMCS database

ShinyApp created by Luis V. Valcarcel and Francisco J. Planes

There are two buttons:

- **Start!:** gMCStool automatically selects the database of gMCSs and the user directed to the panel "Upload the RNA-seq data"
- **Go to gMCS database:** the user goes to the panel "gMCS database".

Step 2 (optional): gMCS database

This optional panel allows the user to modify the database of gMCSs. We recommend to proceed in the order shown in the image below:

The screenshot shows the 'gMCS database' tab in the gMCS tool. The interface is divided into three main sections:

- Select the gMCS database:** This section allows users to select parameters for the gMCS database. It includes options for 'All metabolic essential tasks', 'Only biomass production', and 'Selected tasks'. It also includes options for 'Select biomass production restrictions': 'Growth on Ham's medium' and 'Growth on unconstrained medium (all uptakes available)'.
- Select the metabolic tasks within the gMCS database:** This section allows users to select specific metabolic tasks within the gMCS database. It includes a 'Select all' button and a 'Deselect all' button. The tasks are listed in a tree view, with checkboxes for each task.
- Summary table of gMCS database and metabolic tasks:** This section displays a table with the following columns: 'task.number', 'task.name', 'task.group', and 'num.gMCSs'. The table lists 24 tasks, including 'Aerobic rephosphorylation of ATP from glucose', 'Aerobic rephosphorylation of GTP', 'Aerobic rephosphorylation of CTP', 'Aerobic rephosphorylation of UTP', 'ATP de novo synthesis', 'CTP de novo synthesis', 'UTP de novo synthesis', 'dATP de novo synthesis', 'dCTP de novo synthesis', 'dUTP de novo synthesis', 'Isolysine uptake', 'Lysine uptake', 'Methionine uptake', 'Phenylalanine uptake', 'Threonine uptake', 'Tryptophan uptake', 'Valine uptake', 'Glycerate 3-phosphate de novo synthesis', 'Mitochondrial acetyl-CoA de novo synthesis', and 'Mitochondrial AIC de novo synthesis'.

- Select the gMCS database:** here the user selects the metabolic tasks involved in the database of gMCSs: *all metabolic essential tasks*, only those related with biomass production (*only biomass production*) or *selected tasks* (see below). Moreover, the subset of gMCSs for biomass production can be derived from Ham's growth medium (*Growth on Ham's medium*), as defined in Human1, or from an unconstrained medium including all possible input exchange fluxes from Human 1 (*Growth on unconstrained medium (all uptakes available)*).
- Select the metabolic task within the gMCS database:** if the user decides to custom the selection of gMCSs (*Selected tasks*), individual or groups of tasks can be eliminated or included through the selection tree.
- Summary table of the selected gMCS:** a table with selected metabolic essential tasks, their group from Human1 and their number of gMCSs.

Step 3: Upload the RNA-seq data

The screenshot shows the gMCS tool interface with the following components:

- 1. Data upload instructions:** A box titled "Here you can upload the data" with options to "Select the method to update the gene information:" (Text Files, Import File, Rdata from previous sessions) and a "Show Examples" button.
- 2. File upload fields:** Two "Choose text file" sections. The first is for gene expression (file: gMCSool_gene_expression_2022_04_21_10h13m.txt) and the second is for sample information (file: gMCSool_sample_classification_2022_04_21_10h13m.txt). Both have "Upload complete" buttons.
- 3. Summary of the data input:** A section titled "Summary of the data input:" showing sample class and cohort summaries. The sample class summary lists Naive_B_cells (5), Centroblasts (7), Centrocytes (7), Memory_B_cells (8), Tumor_Plasma_Cell (5), Bone_Marrow_Plasma_Cell (3), and Multiple_Myeloma (17). The sample cohort summary shows BcellMM (72). Below are buttons for "Alphabetically sort sample class" and "Numerically sort sample cohort".
- 4. Data table view:** A table with columns "Sample ID", "sample class", and "sample cohort". It shows 8 rows of data:

Sample ID	sample class	sample cohort
1	Naive_B_cells	BcellMM
2	Naive_B_cells	BcellMM
3	Naive_B_cells	BcellMM
4	Naive_B_cells	BcellMM
5	Naive_B_cells	BcellMM
6	Centroblasts	BcellMM
7	Centroblasts	BcellMM
8	Centroblasts	BcellMM

1. Here you can upload the data: here the user can select the method to update gene expression information. In particular, we have 3 options:

- Text files:** gene expression is provided in a text file, in which the first column stores the genes in Ensemble ID and the rest of columns store their expression for the different samples, being the column name the ID of each sample. The text file can be *.txt, *.tsv, *.csv or any file that can be automatically read with `data.table::fread` in R. Optionally, the user can upload the sample metadata in another text file with contains three columns: sample ID (same as the column names in the gene expression file), sample class and sample cohort. This last column is optional and it is only used to study and visualize separately essential genes per cohort.
- Tximport files:** This option allows the user to load an RDS file with the direct output from `tximport`, a popular R package to read results from pseudo aligners, such as *Kallisto* or *Salmon*.

- c. *Rdata from previous session*: this option allows the user to download the gene expression data and the metadata of the samples for future use. In particular, the user can load previous sessions, including gene expression data, metadata and final results.

In this box, the user can finally click *Show examples*, which hides other input data options and shows the three example datasets:

- i. Example dataset of B-cell subpopulations (35 samples) and MM samples (37 samples) in TPM;
 - ii. Example dataset of B-cell subpopulations (35 samples) and MM samples (37 samples) in $\log_2(\text{TPM}+1)$;
 - iii. The DepMap-CCLE dataset of 621 cell lines used in the reference article.
- 2. Data upload:** depending on the selected option in **step 3.1**, there will appear a different number of inputs. By clicking in the “*Browser...*” option, the user can select the desired file from the local PC.
- 3. Summary table of the input data:** here two tables are shown. The first one summarizes the number of samples and their associated class. The second one summarizes the number of samples for each cohort. The 4 buttons allow the user to arrange the labels according to different criteria.
- 4. Sample metadata:** here we have the complete information of all the samples. If there is no information for the cohort, it will be replaced by “---”. The table is automatically colored for easier inspection. Moreover, it is possible to manually change the values of the metadata. Manual changes are recommended to be saved by clicking the download buttons that are below this table.

Step 4: Predict Essential Genes

1. Select the gMCS database:

Select one:

- ☒ All metabolic essential tasks
- ☐ Only biomass production
- ☐ Selected tasks

Select biomass production restrictions:

- ☒ Growth on Ham's medium
- ☐ Growth on unconstrained medium (all uptakes available)

Gene expression thresholding method

Select one:

- ☒ gmcsTH
- ☐ localT2

quantile [%] of expression threshold

0.05

2. Load Examples or previously calculated results:

Load previously calculated results

Load Example Results

3. Load Example Data:

Load Example Data precomputed Results for gmcsTH [%] (TPM)

Load Example Data precomputed Results for localT2 (all genes in gMCS) (TPM)

Load Example Data precomputed Results for localT2 (all genes in gMCS) (log2(TPM+1))

Load Results of essentiality analysis for all cell lines in DepMap: gmcsTH [%] (TPM)

4. Results of Gene Essentiality Analysis:

Select:

- ☐ number
- ☒ percentage

Show 10 entries

ENSEMBL	SYMBOL	ENTREZID	IsEssential	num.gMCS	Naive_B_cells	Centroblasts	Centrocites	Memory_B_cells	Tonsil_Plasma_Cell	Bone_Marrow_Plasma_Cell	Multi
ENSG00000011884	CCLC	2729	true	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000003137	CYP28B1	54603	true	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000004779	ACUFAB1	4706	false	1632	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000005075	POLR2J	5439	true	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000010256	UQCRC1	7384	true	125	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000013375	PCJ3	5238	true	56	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000013303	POLR3B	55703	true	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000013320	NPCL1	29681	true	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
ENSG00000015186	CYP24A1	15341	true	1	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%

1. Options for the gene essentiality analysis prediction:

1a. Select the gMCS database:

Select one:

- ☒ All metabolic essential tasks
- ☐ Only biomass production
- ☐ Selected tasks

Select biomass production restrictions:

- ☒ Growth on Ham's medium
- ☐ Growth on unconstrained medium (all uptakes available)

1b. Gene expression thresholding method

Select one:

- ☒ gmcsTH
- ☐ localT2

1c. quantile [%] of expression threshold

0.05

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

0.05

1d. Calculate!

a. **Select the gMCS database:** Summarize the options selected in Step 2.

b. **Gene expression thresholding method.** The user can select *gmcsTH* and quantile of expression threshold for X%, which is by default 5%; or *localT2*, where the user can select all genes in Human1 or all genes involved in the database of gMCSs.

- c. **Calculate.** Click here to perform the gene essentiality analysis (GEA) using the gMCS approach. This is the most time-consuming step.

2. Load examples or previously calculated results. In the figure above we clicked in the button “Load Example results” and all the examples included in the tool are shown:

- a. Example results of B-cell subpopulations and MM samples [TPM], *gmcsTH5*.
- b. Example results of B-cell subpopulations and MM samples [log2(TPM+1)], *gmcsTH5*.
- c. Example results of B-cell subpopulations and MM samples [TPM], *localT2*.
- d. Example results of B-cell subpopulations and MM samples [log2(TPM+1)], *localT2*.
- e. Results for the 621 cell lines of DepMap used in the main article with *gmcsTH5*.

We can also click in the button “Load previously calculated results” and go to “Browse...”.

This will show an explorer to select a *.Rdata file with previously saved results, together with sample metadata and gene expression data.

3. Results of Gene Essentiality Analysis. Here there is a summary table that shows the predicted percentage of samples for which a gene is predicted as essential. The user can change from percentages to number of samples. The columns are:

- a. *ENSEMBL*, *SYMBOL*, *ENTREZID*: identifiers of the gene.
- b. *isEssential*: this indicates whether this gene constitute a gMCSs of order 1 (essential for all samples).
- c. *num.gMCS*: number of gMCS in which the gene is involved.
- d. *NB*, *CB*, *CC*, *MEM*, *TPC*, *BMPC*, *MM*: number/percentage of samples of Bcell subpopulations and MM in which this gene is considered essential.

4. Saving the results. there are three buttons to download the results.

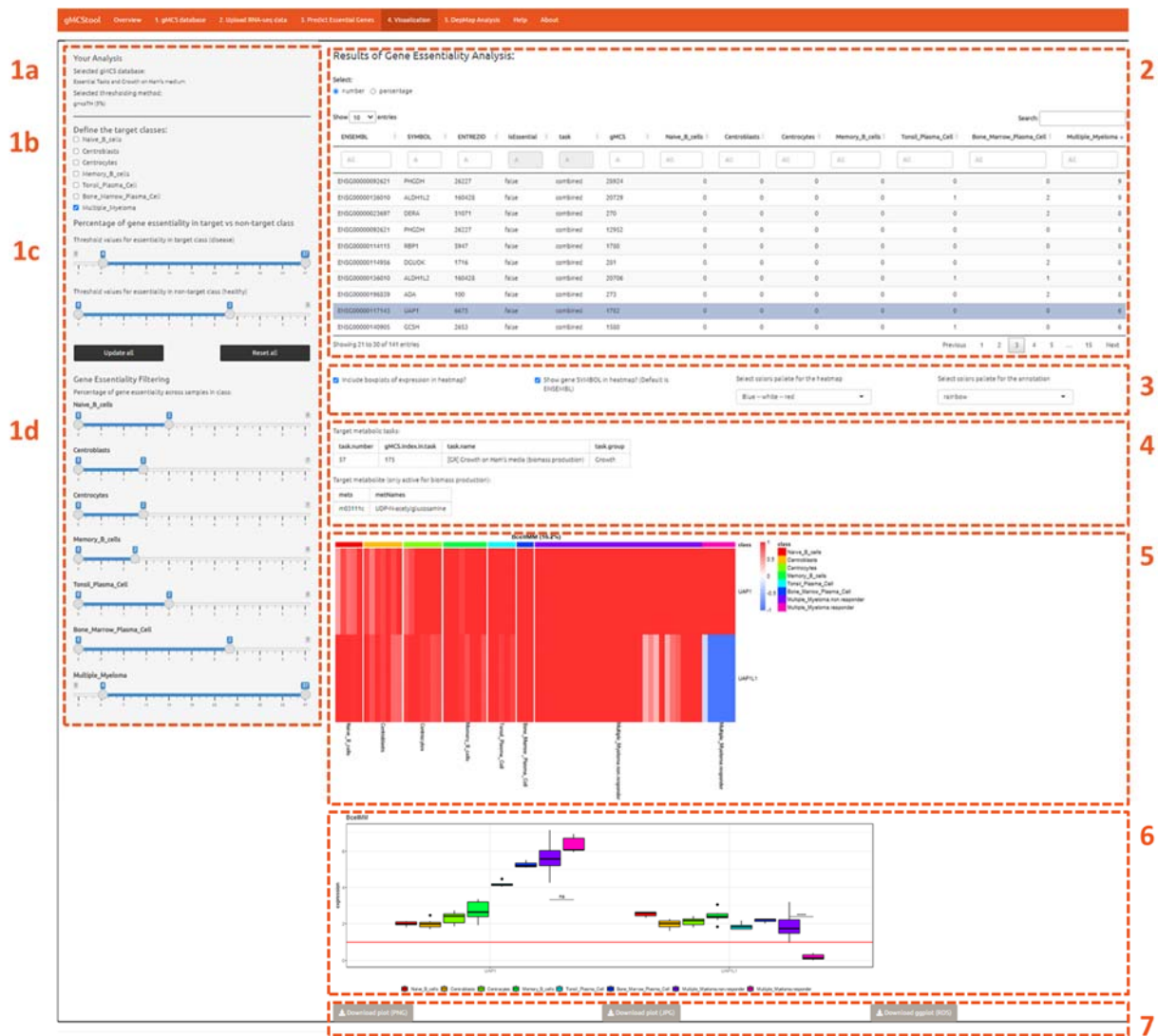
- a. *Download table*: download in a *.txt file the summary table in the step 4.3.
- b. *Download Excel (all)*: Download an Excel file that contains the following information:

- **single met tasks:**
 - i. *ENSEMBL, SYMBOL, ENTREZID*: identifiers of the gene.
 - ii. *isEssential*: this indicates whether this gene constitute a gMCSs of order 1 (essential for all samples).
 - iii. *num.gMCS*: number of gMCS in which the gene is involved.
 - iv. *NB, CB, CC, MEM, TPC, BMPC, MM*: number of samples of Bcell subpopulations and MM in which this gene is considered essential.
- **ratio met tasks:**
 - v. *ENSEMBL, SYMBOL, ENTREZID*: identifiers of the gene.
 - vi. *isEssential*: this indicates whether this gene constitute a gMCSs of order 1 (essential for all samples).
 - vii. *num.gMCS*: number of gMCS in which the gene is involved.
 - viii. *NB, CB, CC, MEM, TPC, BMPC*: percentage of samples of Bcell subpopulations and MM in which this gene is considered essential.
- **gmcs single:**
 - ix. *ENSEMBL, SYMBOL, ENTREZID*: identifiers of the gene.
 - x. *isEssential*: this indicates whether this gene constitute a gMCSs of order 1 (essential for all samples).
 - xi. *task*: name of the task in which it is involved. 'all_57_met_task_combined' means the set of all the gMCS combined for all tasks.
 - xii. *gMCS*: ID of the gMCS.
 - xiii. *NB, CB, CC, MEM, TPC, BMPC*: percentage of samples of Bcell subpopulations and MM in which this gene is considered essential.

- c. *Download Rdata*. We highly recommend this option, which allows the user to save an *.Rdata file that can be posteriorly loaded, as it is indicated in Step 4.2.

All tables will be downloaded according to the sample class that is in the current study, using the names given by the user. Moreover, we can find in the *.Rdata file the variable 'mat.essential.gene', which stores a binary matrix of genes by samples, being 1 it is essential, 0 otherwise.

Step 5: Visualization



1. Options in the selection of predicted essential genes:
 - a. **Your analysis.** Summary of the options used in the gMCS approach.
 - b. **Define the target class.** Select the (disease) sample class we are interested in targeting. In this case, we selected Multiple Myeloma samples.
 - c. **Percentage of gene essentiality in target vs non-target class.** Allowed essentiality percentage intervals for target class (here greater than 10%) and non-target class (here smaller than 2 samples). These values are automatically fixed in the filters below once we press *update all* or *reset all*.

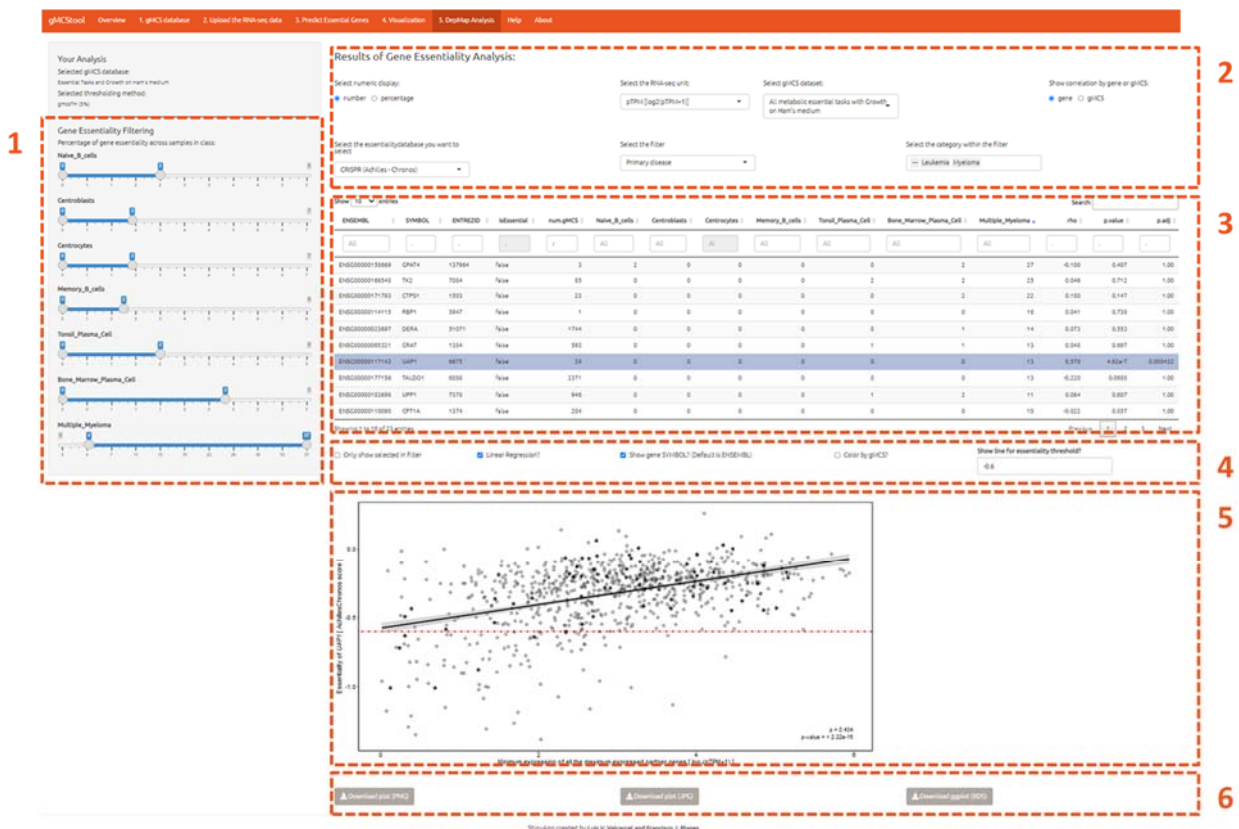
- d. **Gene essentiality filtering.** Here we can change one by one the filters fixed above for each of the sample classes in the study.
2. **Results of gene essentiality analysis.** The summary table has been modified according to the filters fixed above. The options and information are the same as in the previous figure. With a single click in of the rows, the essential gene and its associated gMCS are visualized using a heatmap of expression (see below Step 5. 5). Here, we selected the essential gene UAP1, which is involved in the gMCS {UAP1, UAP1L1}.
3. **Visualization options:**
- a. *Include boxplot of expression in heatmap?:* checkbox to display or not an expression boxplot for genes implied in the gMCS (see below in step 5.6).
 - b. *Show gene SYMBOL in Heapmap?:* checkbox to display SYMBOL or ENSEMBL ID. By default, we have ENSEMBL ID.
 - c. *Select colors palette for heatmap:* In the 'gene' mode, color the dot by the gMCS which explains essentiality (the one with minimum expression of maximum expressed partner).
 - d. *Show line for essentiality threshold:* draw an auxiliary line to visualize essentiality.
4. **Target metabolic task:** the first table indicates the metabolic tasks associated for the gMCS that is visualized, and, in the case this task is the biomass production, the second table indicate the metabolite of the biomass that is blocked by it.
5. **Expression Heatmap:** columns are samples and row are genes involved in the gMCS of interest. Color intensity for each gene and sample is the log2 gene expression relative to the sample expression threshold (gMCSTH5 in this case). Target classes are divided in two sections, according to the samples that are predicted or not as essential

(*Multiple_Myeloma_non responders* and *Multiple_Myeloma_responders*). Separate heatmaps are done for each cohort of samples.

6. **Expression Boxplot:** the same results are plotted in absolute expression, being the x-axis distributed by gene and colored by sample class. A t-test is performed between the two populations of the target classes.
7. **Save the resulting plots.** We can use different image formats: *.png, *.jpeg or a *.rds, which contains a ggplot object that can be modified in R (advanced users).

Step 6: DepMap Analysis

Finally, the user can inspect the resulting gMCS to see if there is a correlation with the DepMap database:



1. **Gene Essentiality Filtering.** Analogously to the analysis conducted in the previous step, we can filter the subset of essential genes to be analyzed. (Step 5.1.d)

2. Visualization Options.

- a. *Select numeric display*: the same as in the previous steps.
- b. *Select the RNA-seq unit*: select the expression unit for involved genes. The options are: $\log_2(\text{TPM}+1)$, z-scores (TPM) or z-scores ($\log_2(\text{TPM})+1$).
- c. *Summary table by single or aggregated gMCSs*: By default we provide the summary table by single gMCS as in Step 8. Aggregating gMCSs for each essential gene implies identifying the most limiting partner gene for each sample across all gMCSs. This is done by extracting the minimum expression over the maximum expressed partner gene for each gMCS. This second option provides the correlation with DepMap data (ρ) and its associated p-value (p.value and adjusted p.value).
- d. *Select the essentiality database you want to select*: The user can choose the different datasets available in DepMap 21Q2.
- e. *Select filter*: the user can select the between the following categories: none (default), primary disease, subtype or lineage.
- f. *Select category within the filter*: This is a multiple selection panel that is activated whenever a filter category is selected.

3. Options for the generated image:

- a. *Only show selected in filter*: checkbox to discard filtered cell lines.
- b. *Linear Regression?*: checkbox to plot the linear regression line.
- c. *Show gene SYMBOL?*: checkbox to display SYMBOL or ENSEMBL ID..
- d. *Color by gMCS?*: In the 'gene' mode, color the dot by the gMCS which explains essentiality (the one with minimum expression of maximum expressed partner).
- e. *Show line for essentiality threshold*: draw an auxiliary line to visualize essentiality.

4. *Correlation plot*: each dot is a cell line, the y-axis represents the DepMap essentiality score, being the more negative, the more essential. The x-axis represents the summary expression of the aggregated gMCSs.
5. Save the resulting plot. Options are: *.png, *.jpeg or a *.rds, which contains a ggplot object that can be modified in R (advanced users).