MAPPI-DAT: analysis and management tool for high-throughput protein-protein interaction data generated from MAPPIT cell microarray system

Surya Gupta[1,2,3], Veronic De Puysseleyr[1,2], José Van der Heyden[1,2], Davy Maddelein[1,2,3], Irma Lemmens[1,2], Sam Lievens[1,2], Jan Tavernier[1,2], Lennart Martens[1,2,3]

[1] Medical Biotechnology Center, VIB, Ghent, Belgium, [2] Department of Biochemistry, Ghent University, Ghent, Belgium, [3] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium

MAPPI-DAT (MAPPIT cell microArray Protein Protein Interaction- Data management & Analysis Tool) is an automated high-throughput data management and analysis system for MAPPIT cell microarray system(Lievens et al., 2009). MAPPI-DAT is capable of processing many thousand data points for each experiment, and comprising a data storage system that stores the experimental data in a structured way for meta-analysis. To extend and ease the usage of the analysis pipeline and database system, a graphical user interface has been developed for the MAPPI-DAT tool.

Contents

Tool download and usage

Standalone program is currently only available for windows. For other operating system you need to start with a python project using a command line.

Pre-requisite

- MySQL platform: to create the database schema, more information about the download can be found at this link: http://dev.mysql.com/downloads/windows/
  (**optional**: Only if you want to use database)
- R version: 3.3.1 with given packages installed:
    - qvalue (https://bioconductor.org/packages/release/bioc/html/qvalue.html)
    - matrixStats (https://cran.r-project.org/web/packages/matrixStats/)
    - BioPhysConnectoR (https://cran.r-project.org/web/packages/BioPhysConnectoR/)
    - car (https://cran.r-project.org/web/packages/car/)
    - docopt (https://cran.r-project.org/web/packages/docopt/)

  you can download these packages by using R command *install.packages("packageName").* After download of the packages check if all the packages are installed.

  After installation of R change the location of R in the "**Parameter.txt**" file given in the MAPPIT folder.

Running Tool

You can access the MAPPI-DAT project from https://github.com/compomics/MAPPI-DAT.git

1. Using command line:
   You can download the MAPPIT project from http://genesis.ugent.be/uvpublicdata/MAPPI-DAT/MAPPI-DAT.zip
   - **Pre-requisite:**

     Python 2.7 with given packages installed:

     1. **mysql connector**: can be downloaded from the link https://dev.mysql.com/downloads/connector/python/)
     2. **matplotlib:** more information: http://matplotlib.org/users/installing.html

        It is suggested to download one of the scientific python distribution like Anaconda from http://www.scipy.org/install.html which include all the packages needed for matplotlib

   - Once you are done fulfill the requirements, download the project from the http://genesis.ugent.be/uvpublicdata/MAPPI-DAT/MAPPI-DAT.zip and unzip the folder.
   - Change the database and R setting in the "**Parameter.txt**" file
   - Tool can be access by running the script inside the ***MAPPI-DAT*** folder as:
     "*python* ***MAPPI-DAT_GUI.py***" or "*python* ***Art_Database_InputGui.py***"
     from command line which should open a user interface for corresponding methods.

2. Directly run tool from executable file for windows:
   - Download **MAPPI-DAT-MainGui.zip** and **MAPPI-DAT_SmallGui.zip** from http://genesis.ugent.be/uvpublicdata/MAPPI-DAT
   - Unzip both folders and double click on ***MAPPI-DAT_GUI.exe*** (*Figure 2*) from folder **MAPPI-DAT-MainGui** and ***Art_Database_InputGui.exe (****Figure 6****)*** from folder **MAPPI-DAT-SmallGui** which will open a user interface for corresponding methods.
   - Change the database and R setting in the "**Parameter.txt**" file.

Connection to database

1. Download MAPPI-DAT schema from http://genesis.ugent.be/uvpublicdata/MAPPI-DAT/Mappi_datScript.sql (*Figure 1*)
2. Connect to a MySQL server (e.g. with MySQL workbench, see http://www.mysql.com/products/workbench/)
3. Create a new schema in the connected server and set it as the default schema.
4. Create a database by running sql script ***Mappi_datVer4_Script.sql*** downloaded in step 1. This will create all tables and relationship between the tables for the MAPPI-DAT database.
5. Use the username, password, server name and database name to modify "***Parameter.txt***" which can be found in:
   - **MAPPI-DAT** folder for command line version, and
   - **MAPPI-DAT-MainGui** and **MAPPI-DAT-SmallGui** folders for the executable versions

   This will allow MAPPI-DAT tool to access the database.

Sample Data

You can download sample data for primary and retest screens from http://genesis.ugent.be/uvpublicdata/MAPPI-DAT/SampleFolder.zip with respective primary **PrimaryScreen** and **ReTestScreen** folders.

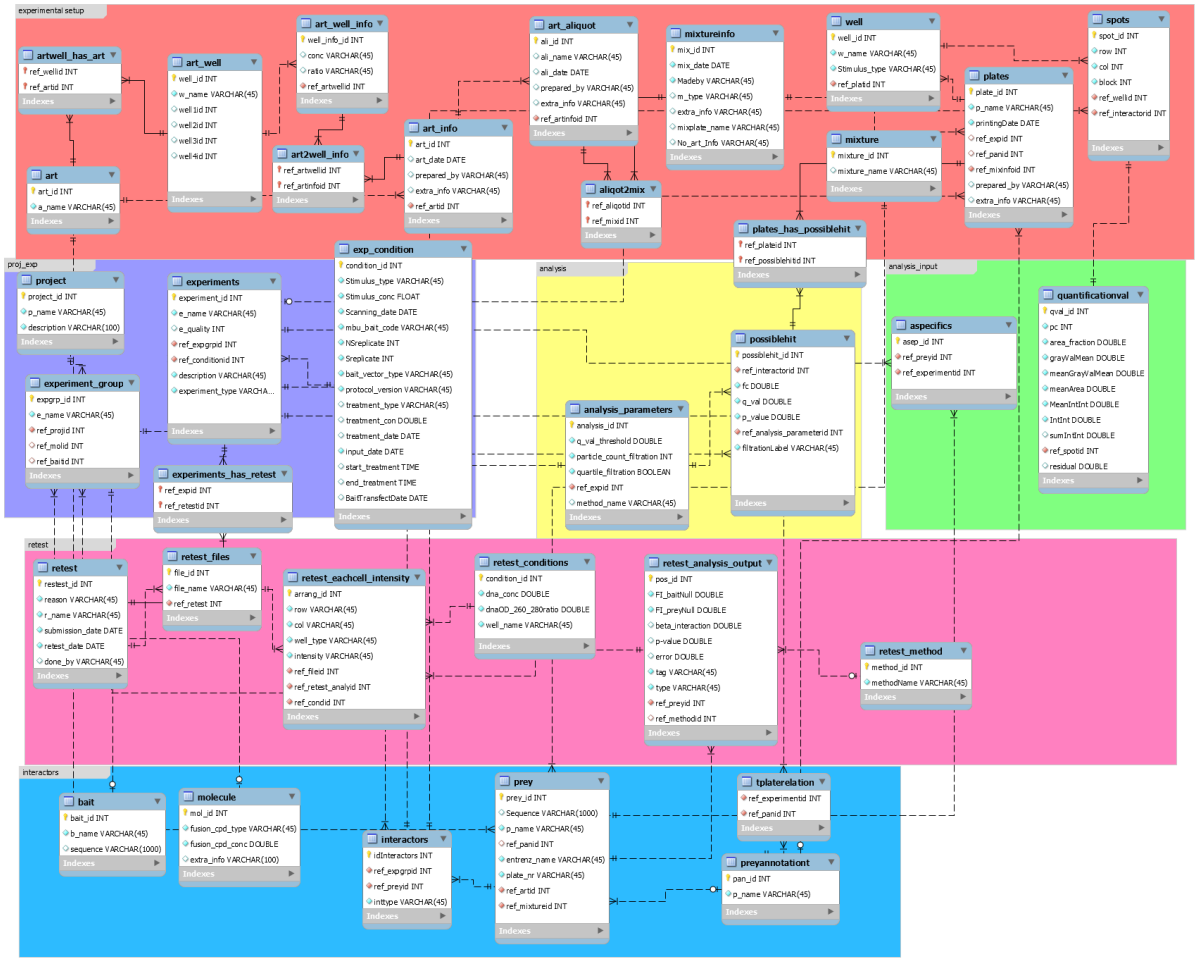

Figure 1 Database Schema: where orange color shows ART tables which is accessed by ART_Database_InputGui, yellow and green color represents table with analysis output and input values respectively, pink section is for Re-test values, blue for plate information and purple for project information

Input Pre-requisite

Primary analysis files

For analysis of primary screens as shown in *Figure 2*, you need three type of input files and one optional:

1. **XML File**: This is the scanner output file, containing 14 columns with 7 quantification parameters where each tab corresponds to different T-Plate.
2. **Linkage File**: This file contains 4 columns where each plate is linked to plate annotation files which define each data point in each plate with the corresponding prey protein information. You can prepare linkage file with : *Prepare Linkage File*
3. **Plate Annotation File**: These files contain annotation for each data point in each plate with corresponding prey information. Each plate corresponds to a different plate annotation file.
4. **A-specific files (optional)**: This file contains list of a-specific prey proteins which show positive interaction with all bait proteins because they bind to the receptor rather than to the bait protein itself.

Primary database submission files

During the data *Submission to MAPPI-DAT small window*, you need a concentration file for each ART 96 well plate which defines the concentration in each well.

Retest analysis files

For analysis of the retest of primary screen you need four type of files:

1. **Primary screen output file**: It is the output file from the primary screen with the list of positive prey interactors with two important columns: prey unique name (column no. 10) and EntrezID (column no. 13)
2. **Connection File**: It is the primary to retest connection file where the proteins from the primary output file are arranged in the retest file in same order in different plates. Steps to prepare connection file are defined here: *Prepare Connection File*
3. **Control Format File:** This file gives the arrangement of the controls in the intensity files with two columns. The rows in the first column corresponds to the controls in the intensity file with entrez names in the corresponding second column.
4. **Retest Files:** These files are in the format of 96 well plate with 12 columns and 8 rows. Each spot in the well defines the intensity of interactions. First 6 columns are without bait and next 6 with bait with each 3 non-stimulated and 3 stimulated wells.

Retest database submission file

During the data *Submission to MAPPI-DAT main tool for retest screen* you need two concentration file, one for control files and one for test files which defines the concentration of each well in the 96 well plate format.

Analysis

To perform analysis, start the MAPPI-DAT tool using .exe or from command line. More description can be found in: *Tool download and usage*

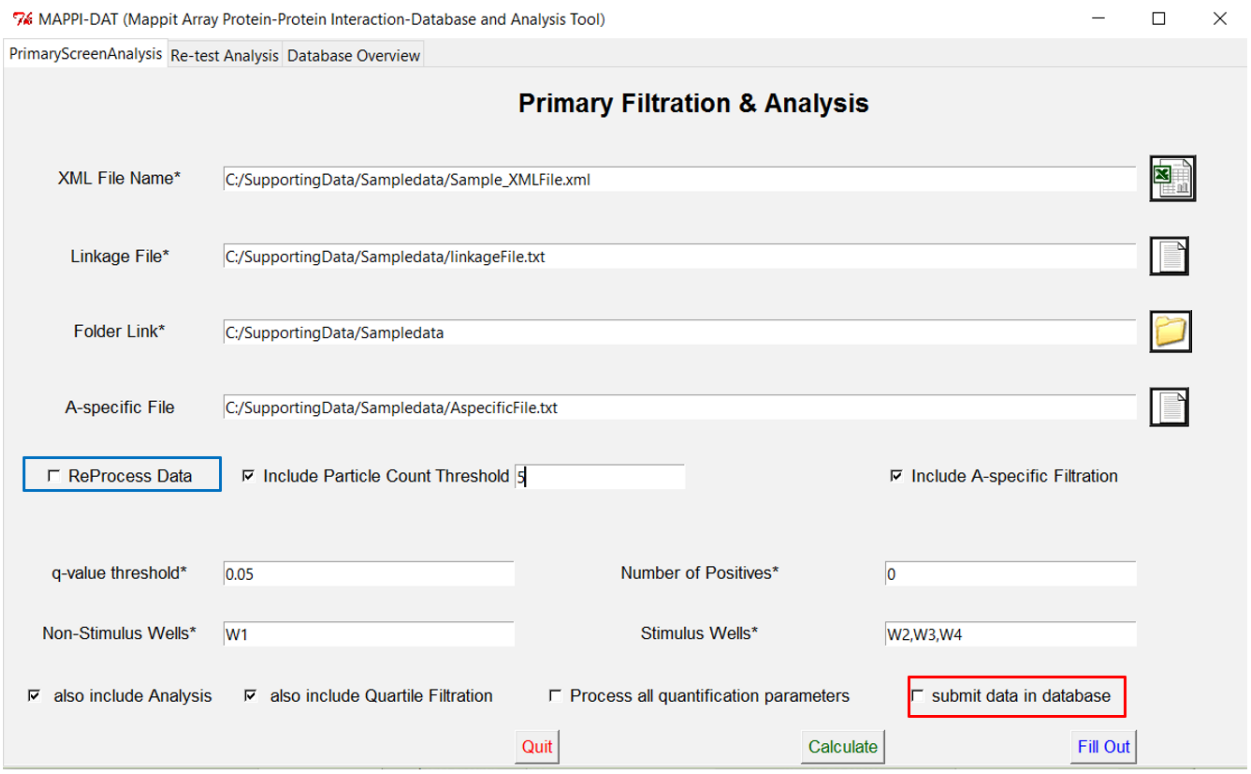

Figure 2 Graphical user Interface for MAPPIDAT tool

Primary Screen Analysis

For analysis you need three type of files (*Figure 2*). More info can be found here: *Primary analysis files*

1. Import xml file in tab "**XML File Name**"
2. Do not forget to change the path to TFiles in the **linkageFile.txt**, given in the sample folder.
3. Then click on "**Fill Out**" button which will use the path from the XML File and set default parameter.
4. Post filtration can be included by clicking on check box corresponding to type of post filtration. More explanation about each filtration can be found in section *Particle Count filtration* and *Quartile based filtration*.
5. If you have a list of a-specifics you can locate the path of the file and then click on the "**Include A-specific Filtration**" check box.
6. Highlight "**also Include Analysis**" check box.
7. Then finally click on "**Calculate**" button.

To perform submission before calculation check section *Submission to MAPPI-DAT main tool for primary screen* for more detail

The black window with the tool will show the progress of the method. Once analysis is done, you can find a folder name "**MAPPIDAT_OutPut**" in the path defined in the "**Folder Link**" in primary screen tab. More info can be found here: *Primary output data*

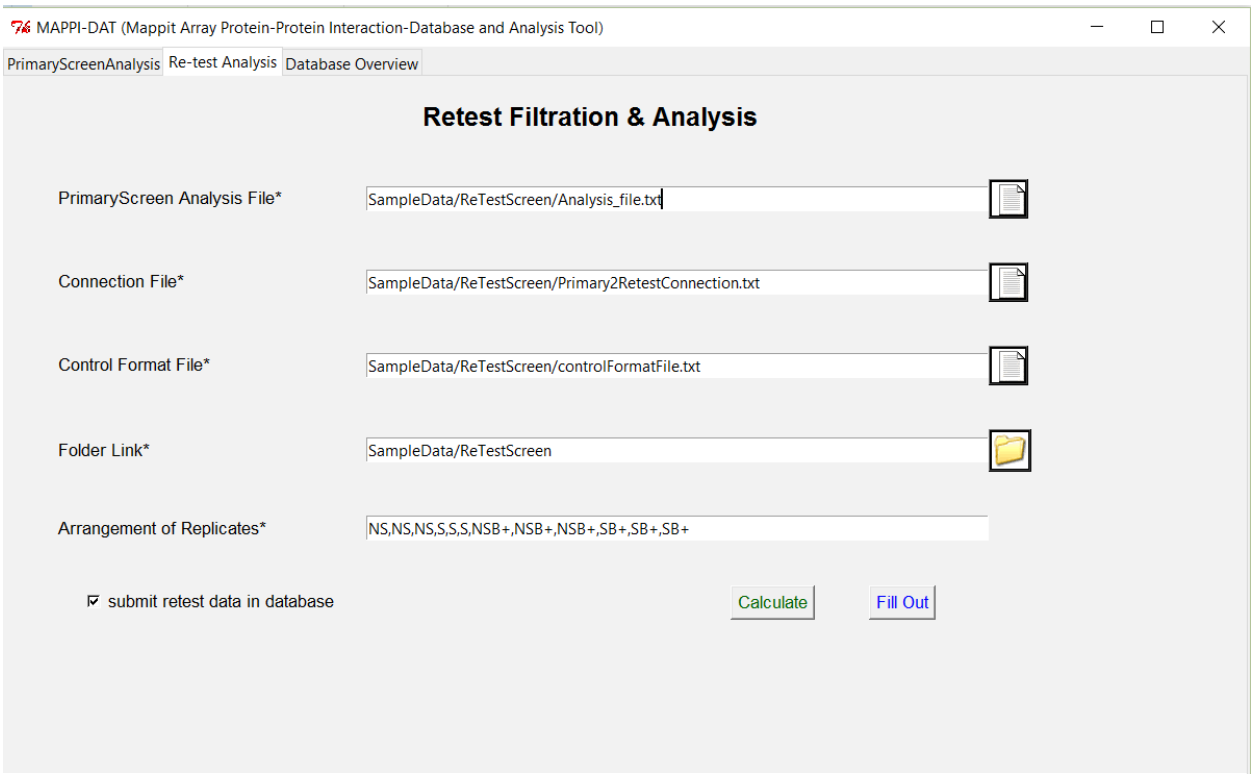

Figure 3 Re-test analysis window

Re-Analysis

Export selection allows you to download analyzed data as well as raw data in form of tab-delimited txt. You can also download all raw data for the selected experiment using **RawDataFiles** check button on export window, in form of tab-delimited txt file (Figure 16). You can re-analyze re-submit all this raw data using the same MAPPI-DAT analysis panel.

To re-analyze the downloaded data you need three type of files; RawDataPre_Processed file, linkage file, T-files, detailed information can be found under section *Raw export data*. All these files are included in the defined place during export.

- First select "**ReProcessData**" button on the MAPPI-DAT analysis panel, as shown in *Figure 2* (marked with blue colour).
- Import "**RawDataPre_Processed**" file in the xml panel, provided in the export folder.
- As the downloaded data is already pre-processed which include the annotations for prey therefore you can skip defining A-specific file.
- Rest all steps are similar as defined under section *Primary Screen Analysis*

Re-test Analysis

For the retest analysis you will need four type of files (*Figure 3*). More info can be found here: *Retest analysis files*

- Import primary output file, connection file and control format file in the desired panel.
- Once all the required fields are completed, click on "**Calculate**" button.

To perform submission before calculation check section *Submission to MAPPI-DAT main tool for retest screen* for more detail

The black window with the tool will show the progress of the method. Once analysis is done, you can find the output in the path defined in the "**Folder Link**" in re-test Analysis tab. More info can be found here: *Retest output data*

Database Submission

To submit data into the database you need to connect the local database to the tool. For more details: *Connection to database*

Once the database is connected to the MAPPI-DAT tool you can submit the data in the database during analysis.

But first you need to add information about the plate using ***Art_Database_InputGui.exe*** from folder **MAPPI-DAT-SmallGui**, steps to follow: *Submission to MAPPI-DAT small window*. This will allow to store information about each ART plate, ART plate aliquot, mixture and plate information. This information will be directly linked to the data while analysis is done.

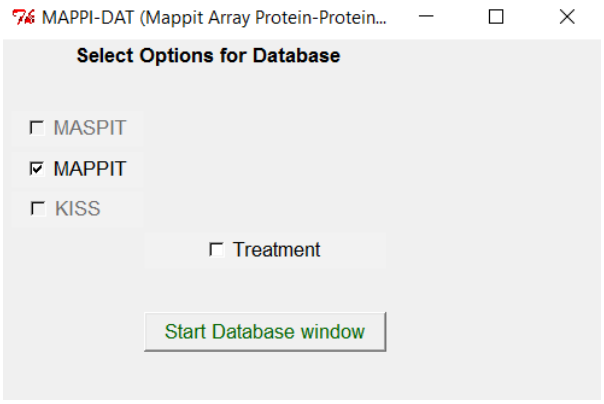

Figure 4 Pre-database selection window: for selection of type of experiment that need to be submit

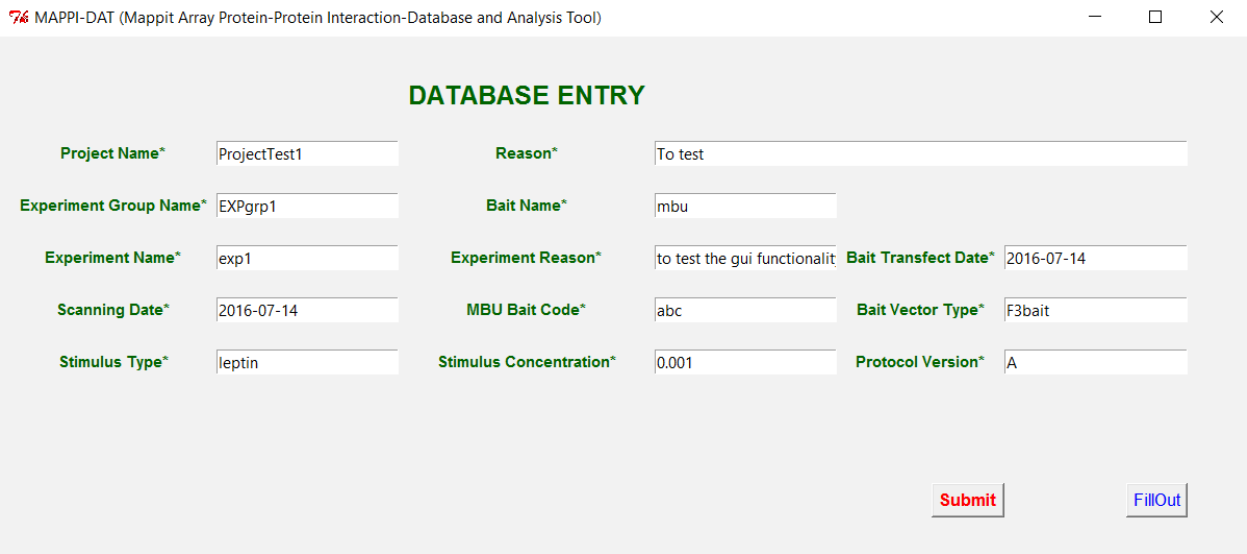

Figure 5 Database submission window for all analyzed and raw data from MAPPIT cell microarray system

Submission to MAPPI-DAT main tool for primary screen

- To input data in the database during analysis, click on "**submit to database**" check box as marked with red in *Figure 2*.
- From the pop up window (shown in *Figure 4*), select type of experiment you want to enter data for.
- After you made your selection, click on "**Start Database Window**" button.
- Fill information about the project in the next open window, as shown in *Figure 5*.
- You can try the "**Fill Out**" button which will allow you to automatically fill the fields default values and path. Change fields according to input data before submitting the data in the database.
- After completing all fields in the panel you can click on "**Submit**" button which will close the current window.
- Cross check all the information in the Analysis panel (*Figure 2*) and then click on "**Calculate**" button. This will perform analysis and also add the data in the database.

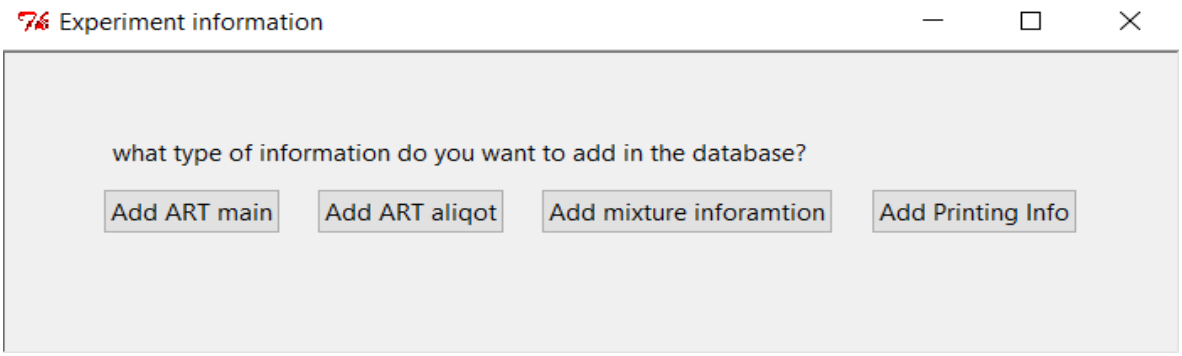

Figure 6 Art_Database_InputGui: selection window to select what type of information need to be added in the database

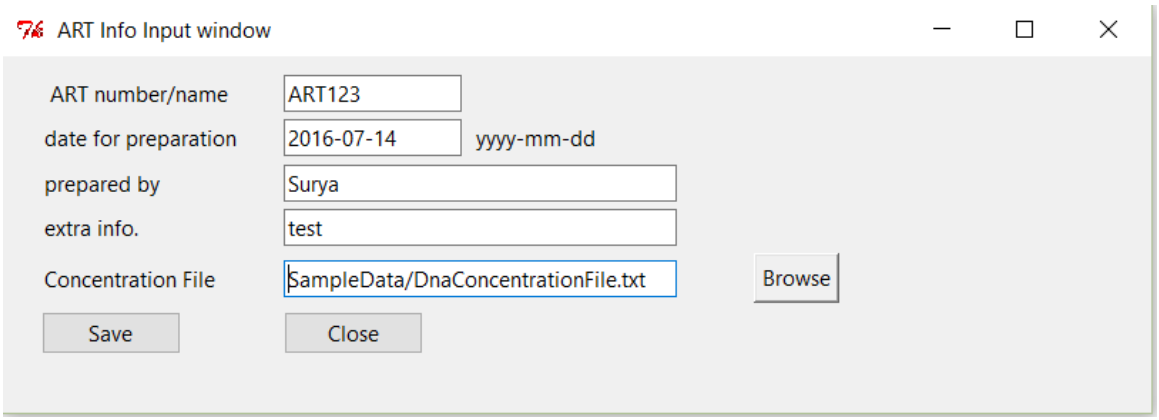

Figure 7 Art_Database_InputGui: ART info window to enter information about each ART plates

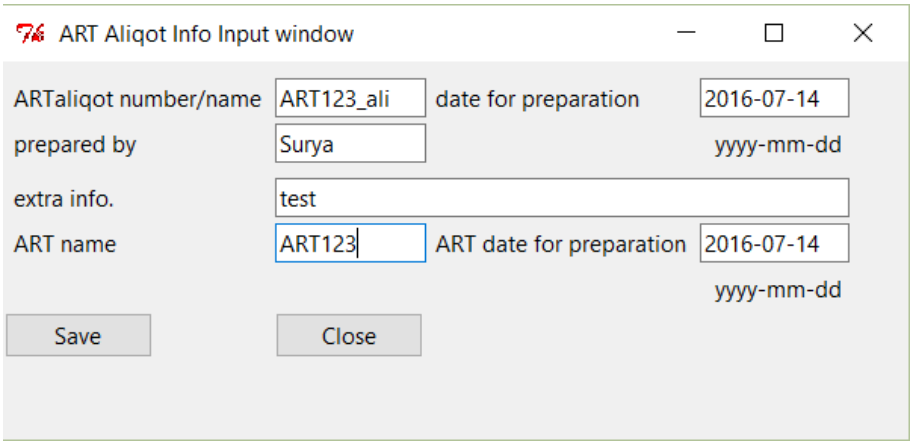

Figure 8 Art_Database_InputGui: aliquot info window to enter information about aliquots for respective ART plates

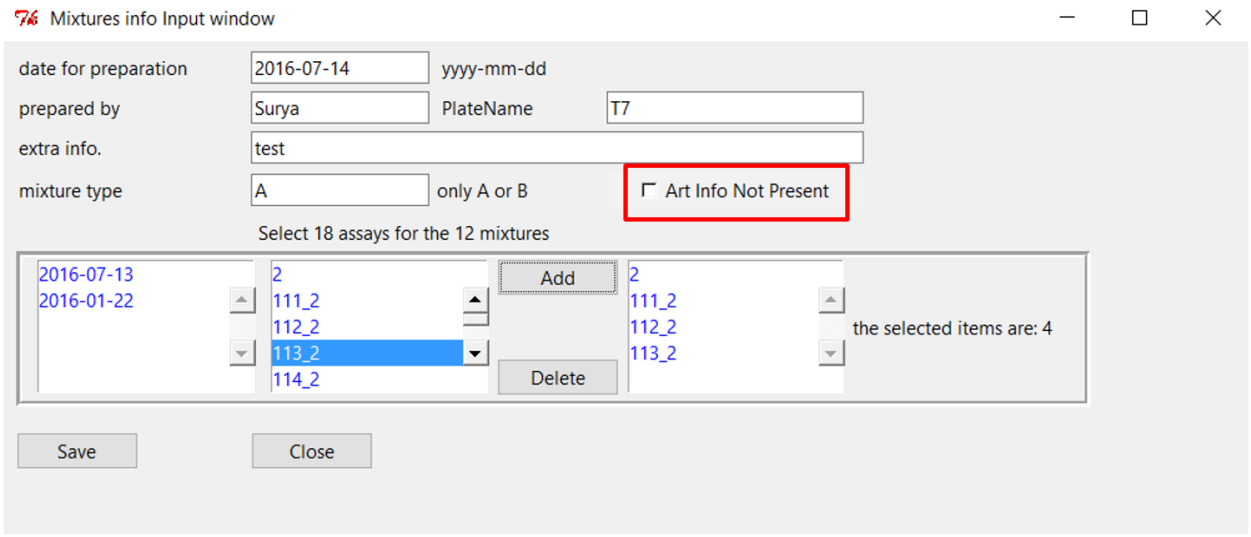

Figure 9 Art_Database_InputGui: Mixture info window to enter information about mixtures corresponding to each plate

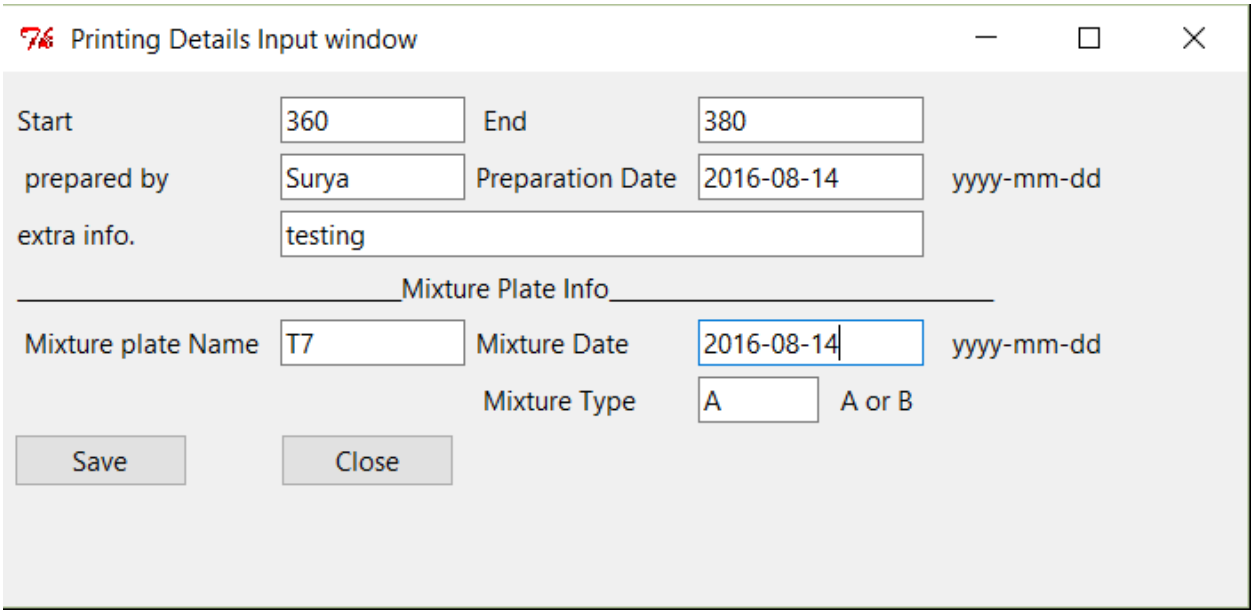

Figure 10 Art_Database_InputGui: Plate Printing info window to enter information about each microarray printing plate before MAPPIT analysis

Submission to MAPPI-DAT small window

To submit plate information, start the ***Art_Database_InputGui*** using .exe from **MAPPI-DAT-SmallGui** or from command line. For more description: *Tool download and usage*. The schema for ART main to printing plate can be found in *Figure 18*. Check for required file before submission: *Primary database submission files*

1. Running ***Art_Database_InputGui*** will open a window which allow you to select type of information you want to enter in the database as shown in *Figure 6*.
2. If you do have Art and Art aliquot information about the plate than you can skip step 3, 4, and 5, and directly can go to step 6.
3. Clicking on "**Add ART Main**" button will open a window as shown in *Figure 7*, which will allow you to enter ART information in the database. You will also need to import concentration file for respective ART. Once done, click on "**Save**" button to save ART information in the database and then press "**Close**" button.
4. Then you can proceed with aliquot information by clicking "**Add ART aliquot**" button from the main window. In new pop up window as shown in *Figure 8*, along with aliquot information you will also need to specify ART name and date, for which you are entering the aliquot information. Once you done click on "**Save**" button and then close the window with "**Close**" button.
5. **With ART-aliquot information:**

   To enter mixture information for each plate, you need 18 ART-aliquot entries.

   - To start, click on "**Add mixture Information**" button in the main window, a new window will pop up where you need to specify the details of mixture as shown in *Figure 9*.
   - click on the dates in the lower left panel which will show all art-aliquot entries entered in that date.

- You need to select at least 18 ART-aliquot entries to proceed by clicking add button in the window.
- Once done with the selection of aliquots, then press "**Save**" button to save the entry.
- Click on "**Close**" button to close the window

6. **Without ART-aliquot information:**
   If you do not have ART-aliquot information, then follow these steps:
   - To start, click on "**Add mixture Information**" button in the main window, a new window will pop up where you need to specify the details of mixture as shown in *Figure 9*.
   - Then click on "**Art Info not present**" check box (marked with red in *Figure 9*), which will allow to save the mixture information without aliquot information.
   - Save the information by clicking on "**Save**" button and then press "**Close**" to close the window.
7. To add the printing information about plate, click on "**Add Printing Information**" which will display a window where you can add the plate information as shown in *Figure 10*. You will need the mixture name and date to enter the printing information. Once done, click "**Save**" to save and continue with "**Close**" button to close the window.

After this you can add the data in database with main MAPPI-DAT Gui with analysis with *Submission to MAPPI-DAT main tool for primary screen*

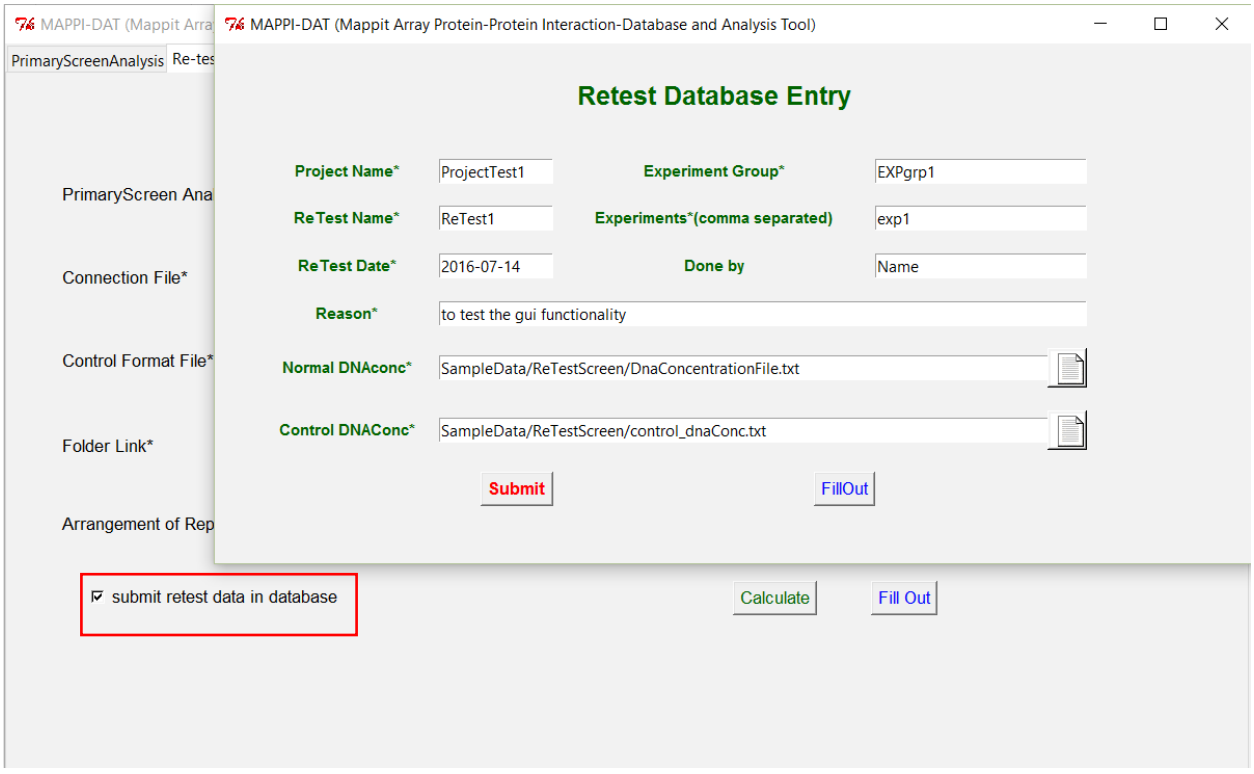

Figure 11 Re-test database submission window

Submission to MAPPI-DAT main tool for retest screen

You need some files to submit data to the database. More info can be found here: *Retest database submission file*

- To input data in the database during retest analysis, click on "**submit to database**" check box as marked with red in *Figure 11*. This will start a new window as shown in *Figure 11*.
- Enter the details about the project and experiment for which you want to add data for retest analysis.
- You can try "**Fill Out**" button which will allow you to automatically fill the fields default values and path. Change fields according to input data before submitting the data in the database
- After completing all fields in the panel you can click on "**Submit**" button which will close the current window.
- Cross check all the information in the Analysis panel and then click on "**Calculate**" button. This will perform analysis and also add the data in the database.

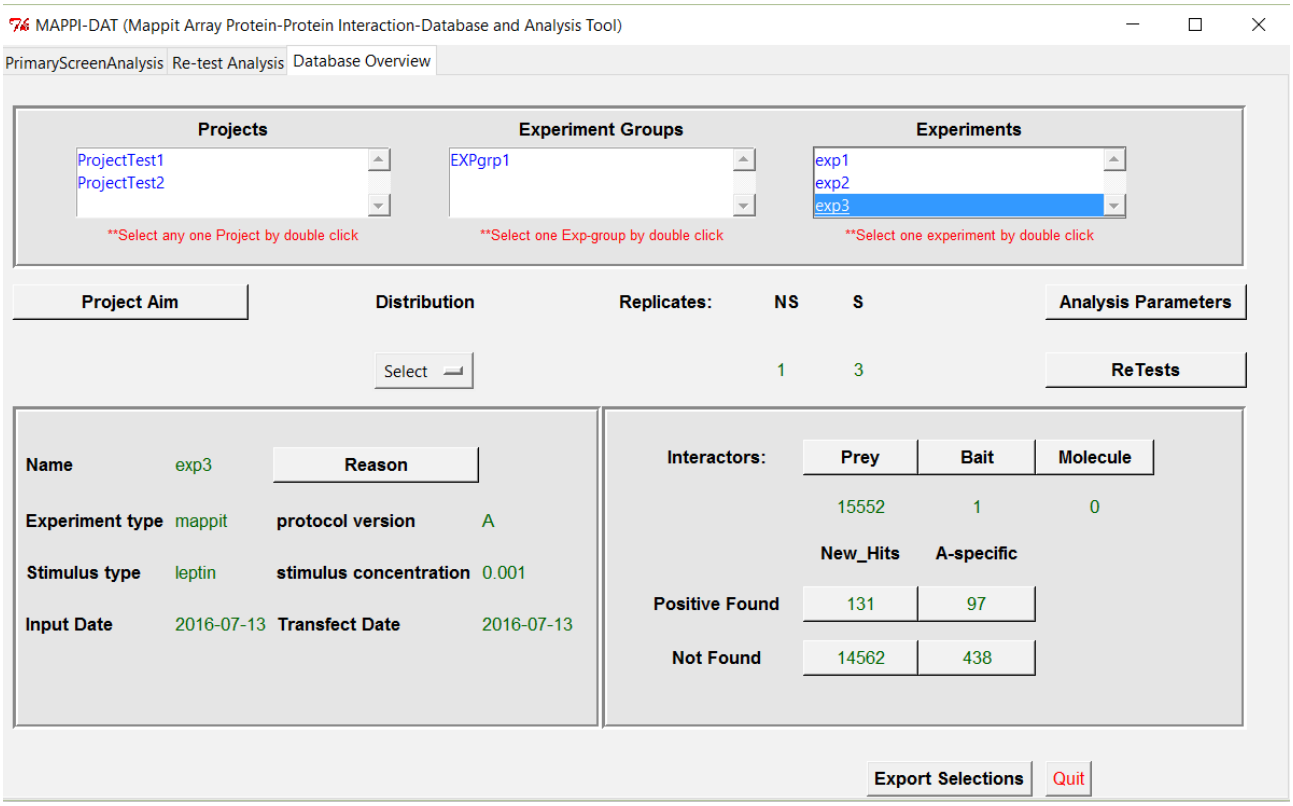

Figure 12 Database overview window allows user to check all the existing data in the database.

Database Overview

With the database overview tab as shown in *Figure 12*, you can look at all the data that is present in the database. First you need to double click on the project which will show you all the experiment group that it contains. Double click on a particular experimental group will show you all existing experiments. When you select the experiment, it will show all meta information about that experiment.

"**Analysis Parameters**" button will show the parameter setting that is used while performing analysis.

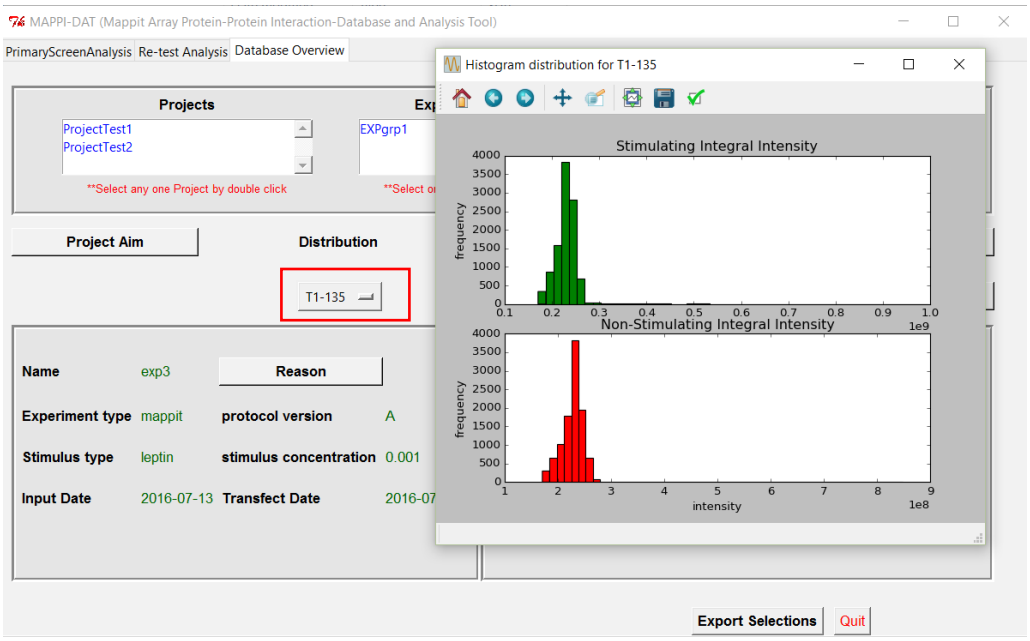

Figure 13 selection of plates from the drop box in the Distribution panel allow user to see histogram for stimulated and non-stimulated intensities.

Under the "**Distribution**" panel in the drop box you can select the plate name and it will show you the histogram plot for non-stimulated and stimulated intensities, as shown in *Figure 13*.

Lower left column will show all meta information regarding the experiment. However, on the lower right column you can find more information about the prey and bait.

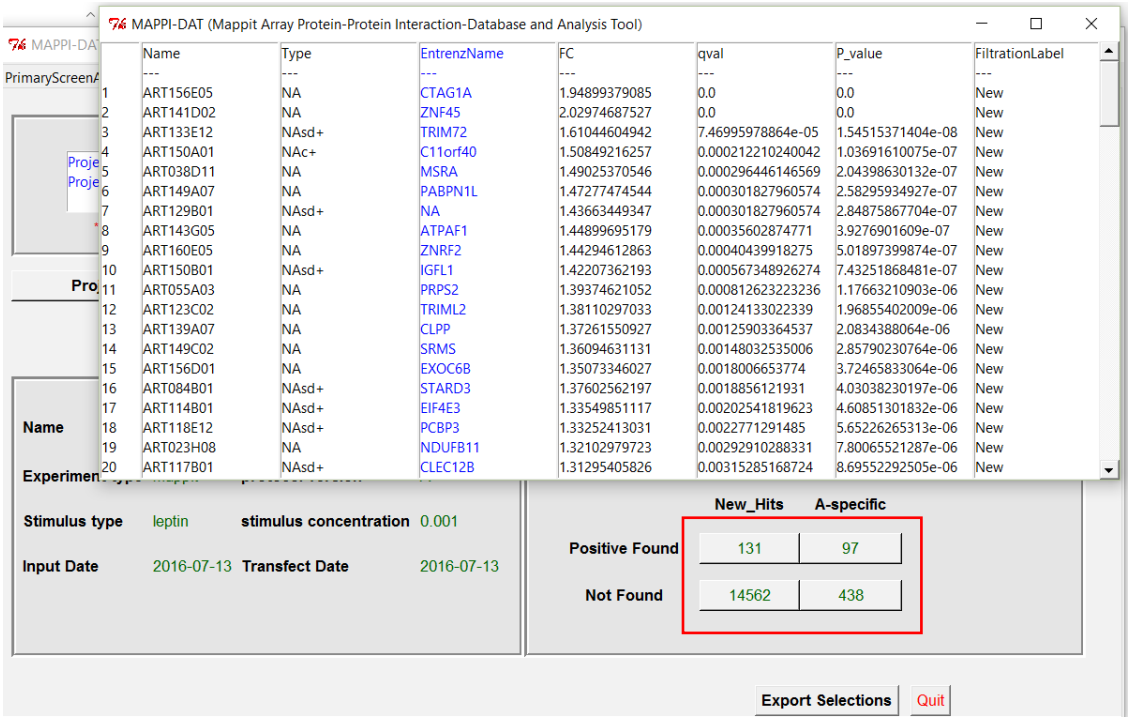

Figure 14 clicking on button with number on lower right corner allows user to look at the list of preys with their calculated values.

In the lower left column, you can also click on the number of positive found for the new hits and A-specific which will pop up a new window with the list of positives with their q-value and p-values, as shown in *Figure 14*.

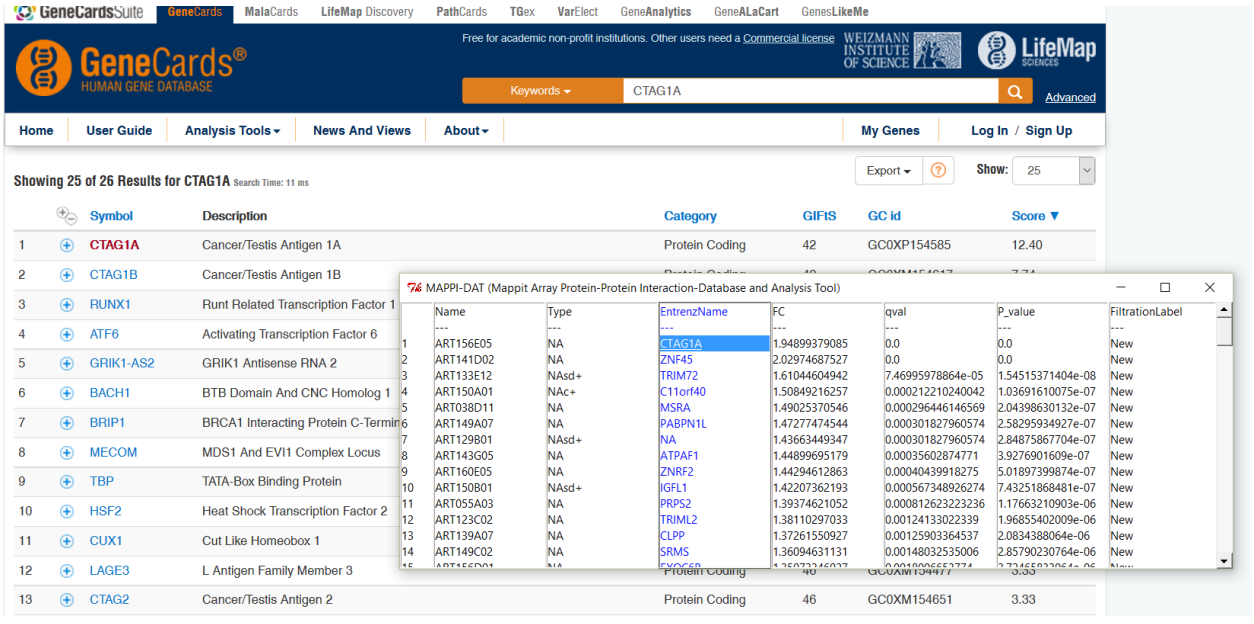

Figure 15 selection of one the prey from the list will redirect user to the GeneCard entry for the selected prey

If you double click on any *EntrezName* it will redirect you to the web page “Genecards.org” (Safran et al., 2010) for more details, as shown in *Figure 15*.

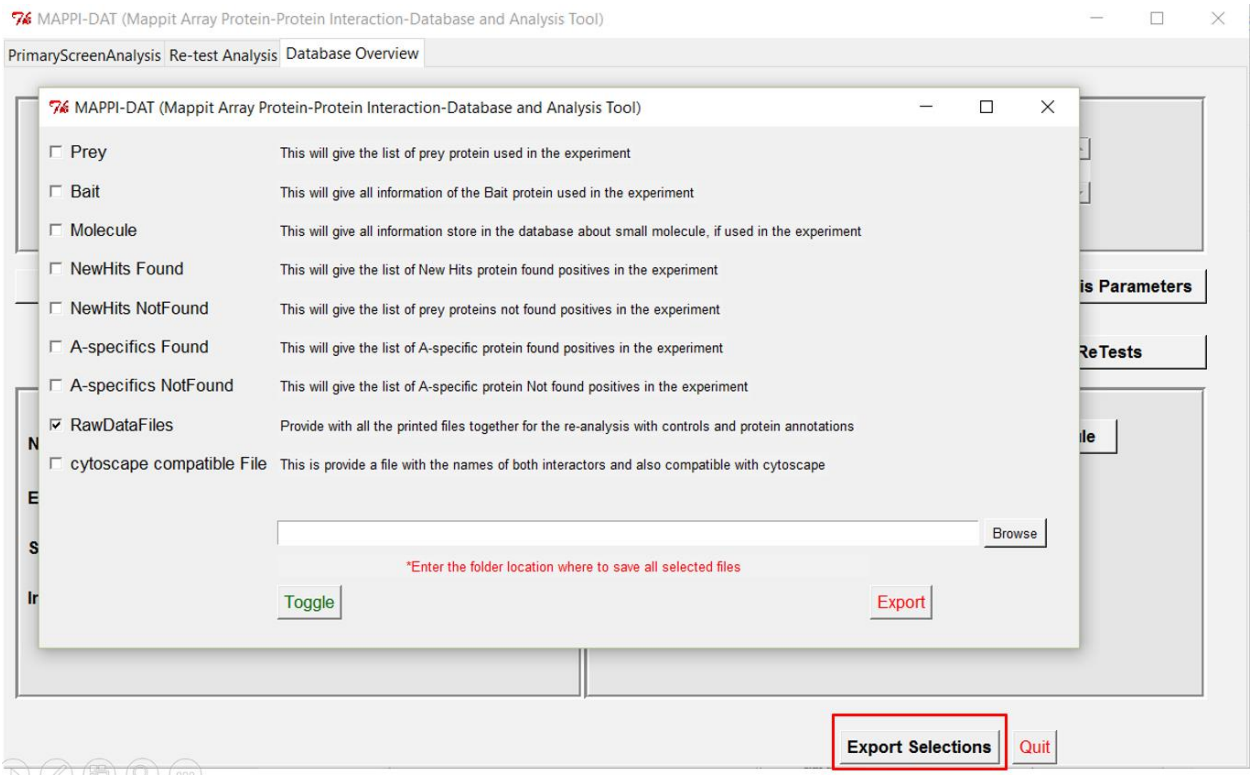

Figure 16 "Export" button in the lower left panel allow user to export all the analyzed data from database in the desired location

"**Export**" button on the lower right corner will allows you to download all data for that particular experiment in the defined folder, as shown in *Figure 16*.

Export selection allows you to download analyzed data as well as raw data in form of tab-delimited txt. For example, you can download cytoscape compatible file which will not only allow you to visualize the resultant protein interaction, but will also allow you to add the existing different database information on the network using different cytoscape's applications.

You can also download all raw data for the selected experiment using **RawDataFiles** in form of tab-delimited txt file (*Raw export data*). More information can be found under section *Re-Analysis*

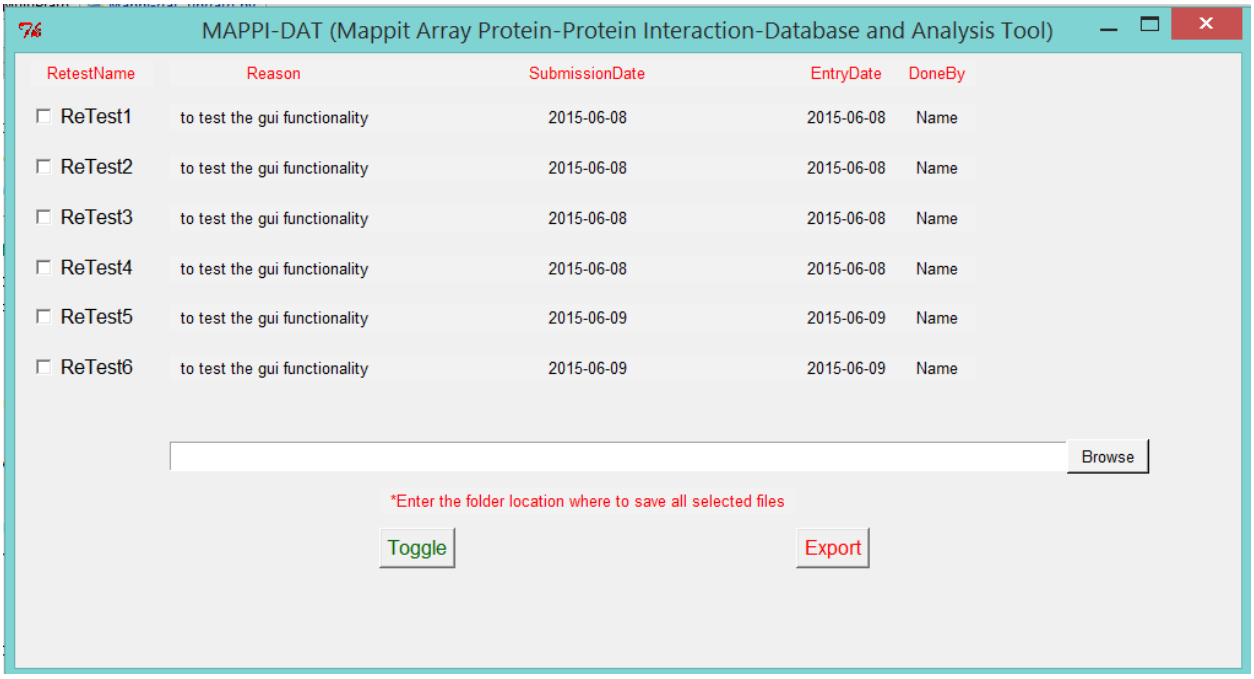

Figure 17 selection of "ReTests" button in the middle left, allows to look at all the performed retests for the selected experiment. This data can also be exported using "Export" button

"**ReTests**" button allows you to look at all existing retest analysis data in the database for selected experiment. It also allows you to save the information on the desired location, as shown in *Figure 17*.

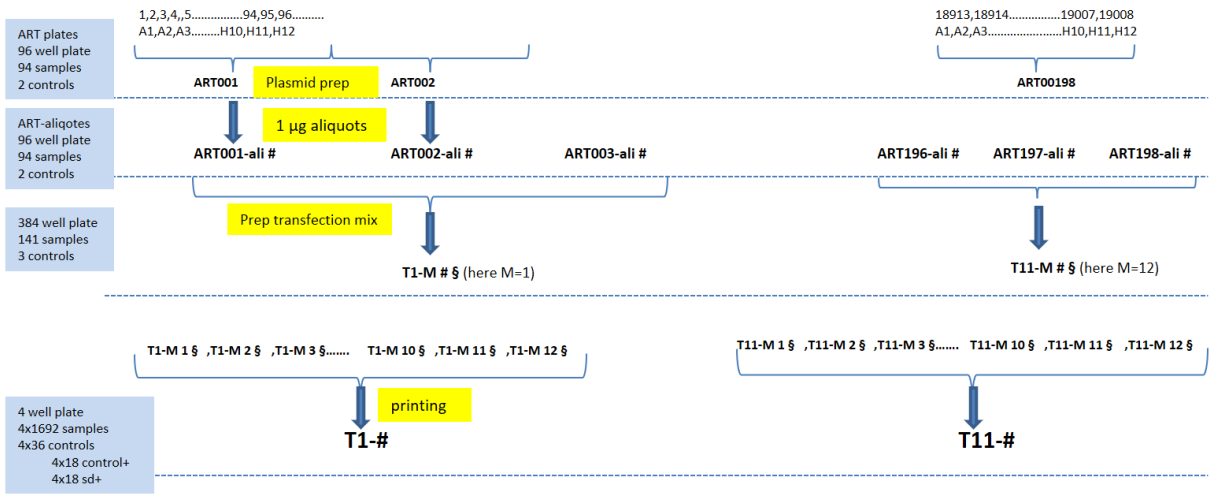

Figure 18 flow chart to show the arrangement of ART to printing plates

OutPut Data

Primary output data

After the primary analysis, you can find the all output data in the folder "**MAPPIDAT_OutPut**" in the path defined in the "**Folder Link**" panel in the analysis window. MAPPIDAT_Output folder contains 2 sub-folders:

1. **Processing:** This folder contains processed files before normalization: text file for each plate, all merged files with and without controls, and two box plots for control+ (stimulation independent control) and SD+ (stimulation dependent control).
   Control+ box plot is with respect to log2 intensities and SD+ box plot is with respect to fold change of stimulated over non-stimulated intensities.
2. **Analysis:** This folder contains all files after normalization including the after normalization file for integral intensity, particle count and residuals, and also analyzed output file.
   "**AllPlatesWithoutControlNormalized_SelectedAnalyzed.txt**" file is final output file which contains the list of preys that have passed the q-value threshold set by user.
   - First eight columns in the file defines the location of each point in the well and plate where it belongs to.
   - Next four columns define the annotation of prey followed by four columns which defines the analyzed value.
   - The next columns define: intensity, residual values, and, particle count.
   - The last column defines the type of interaction according to the post filtration for each replicates.

Retest output data

After retest analysis, you can find the output file "**ReTestOutputFile.txt**" in the path defined in the "**Folder Link**" in re-test Analysis tab.

First three columns show information about the prey, followed by intensity of the prey. Next seven columns define the fold induction, minimum of fold induction, standard deviation and coefficient variance of fold induction.

Column "**tag**" defines, a tag for each prey on the basis of the fold induction values calculated using intensities.

Raw export data

"**RawDataFiles"** button in the export panel, as shown in *Figure 16* will export all raw data for the defined experiment in an folder "**MAPPI_DAT_RawData"** which include three type of files:

- **RawDataPre-Processed File**: this is pre-processed file, which include all the plates information for the selected experiment. It is tab-delimited txt file. It also includes the prey annotation.
- **linkageFile**: it is similar to the linkage file defined in the section *Primary analysis files*
- **Tfiles** (folder): this folder contains all the t-files needed to re-analyze and re-submit this raw data in the database.

File Preparation

Prepare Linkage File

You need to prepare a linkage file before proceeding for primary analysis. Linkage file contain four columns.

- First column defines the bait name, if you have more than one bait in the project.
- Second column defines the type of mixture file.
- In third column each row corresponds to each plate in the XML file.
- In the fourth column you need to define the path of each protein annotation file.

Once you are done with the linkage file you can continue with the analysis.

Prepare Connection File

You need to prepare a connection file before proceeding to the retest analysis. The connection file contains four columns.

- First column defines the retest file names
- Second defines if it contains control sample or not
- Third and fourth defines which preys are tested in that plate with the number which come from main *Primary screen output file*: It is the output file from the primary screen with the list of positive prey interactors with two important columns: prey unique name (column no. 10) and EntrezID (column no. 13).

Definition

A-specific prey

A-specific preys are those preys which every time bind to the receptor rather than to the bait protein itself which in turn leads to high fluorescence intensity, showing it as positive interaction. However, the protein is not interacting with the bait.

Particle Count filtration

The microarray scanner also measures the interaction in terms of Particle Count, where each illuminated cell is counted. Particle Count filtration uses this measurement to eliminate those cases where either the median of particle count in stimulated replicates is lower than median of particle count in non-stimulated replicates or particle count in more than half of the stimulated replicate is less than the particle count define by the user in the analysis panel, under "**Include Particle Count Threshold**" (*Figure 2*)

Quartile based filtration

The quartile based filtration is a classical approach, mainly used to remove outliers in non-normal distribution. Here it is used to account for the cases where the intensity of the non-stimulated data point is higher than the background intensity but lower than the stimulated data point. The threshold is the

total of third quartile and 1.5 of inter-quartile range (IQRs). The threshold is only applied on non-stimulated replicates to label protein pairs where the intensity of the non-stimulated replicates is higher than the threshold in more than half of the non-stimulated replicates.

References:

Lievens, S., Vanderroost, N., Heyden, J. Van Der, Gesellchen, V., Vidal, M., Tavernier, J., & Heyden, V. Der. (2009). Array MAPPIT : High-Throughput Interactome Analysis in Mammalian Cells, 877–886. http://doi.org/10.1021/pr8005167

Safran, M., Dalah, I., Alexander, J., Rosen, N., Iny Stein, T., Shmoish, M., … Lancet, D. (2010). GeneCards Version 3: the human gene integrator. *Database : The Journal of Biological Databases and Curation*, *2010*, baq020. http://doi.org/10.1093/database/baq020