HCS Analyzer tutorial v1.0.4.1

This tutorial illustrates HCS analyzer functionalities, showing how the user can handle the analysis of an entire High Content Screening exported database.

For this tutorial, we used an artificially generated siRNA screening database composed of 25473 wells distributed over 68 plates. The plate format is a regular 384 wells plate, with 24 columns and 16 rows.

Content

Institut Pasteur Korea

Import Screening

Even if HCS analyzer is able to import MTR files, it has been principally designed to load CSV file. In such file 2 columns are mandatory: one column containing the plate name and another one with the well absolute position of the well. Each row represents a well where any descriptors can be added in a **double** format. The user has the possibility to add 3 others specific data:

- A name (compounds or siRNA name by example) as a **string**

- An information as a **string**

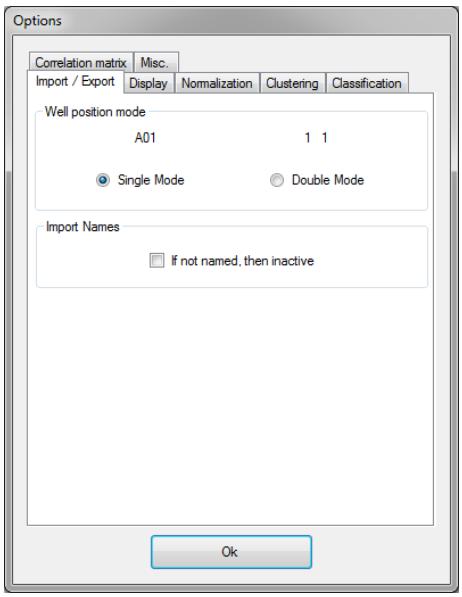
- A Locus ID as an **integer** that will allow the software to make the link to KEGG database.

Here is a typical example of such file generated by MS Excel and save in CSV format.

Note: The names as well as the position of the column do not have to follow any rule as the software loader is flexible enough to deal with this kind of variability.

Plate384	Well 384w	Cat	Catalog Numb	Gene Symbol	Description	Locus ID	Fluo1	Size	Volume	Ratio1	Fluo2	Texture	Rand1	Rand2	Rand3	Rand4
SRP000459	D24	Druggable	1	1	adrenocortical dysplasia homolog	65057	0.822859	0.982368	59.26912	144.2518	90.04089	104.5234	48.66298	35.7505	36.28192	14.83044
SRP000479	P20	Druggable	2	2	deleted in esophageal cancer 1	115123	0.72335	0.987039	183.9027	234.0327	176.5919	231.5573	0.90994	7.755161	22.86228	23.36147
SRP000502	G11	Druggable	3	3	hypothetical protein CG003	57574	0.909086	0.724062	137.0043	139.0109	74.81597	195.0901	34.82006	5.351135	40.45101	0.397935
SRP000469	P09	Druggable	4	4	hypothetical protein 15E1.2	54708	0.75825	0.943229	58.89768	196.4233	54.3536	118.3427	35.54298	47.71962	46.30442	9.891081
SRP000463	F05	Druggable	5	5	82-kD FMRP Interacting Protein	10299	0.913974	0.936443	144.9422	481.5817	40.2687	156.4878	33.01084	18.35786	16.16318	32.8679
SRP000474	K16	Druggable	6	6	hypothetical protein 384D8_6	64844	0.951024	0.940594	22.85816	499.7528	7.579783	113.2114	3.485309	5.603355	46.58561	42.21598
SRP000480	D10	Druggable	7	7	3' exoribonuclease	92979	0.983895	0.993165	23.72045	125.6528	19.88528	38.27095	17.06228	42.23741	13.09605	14.63739
SRP000466	E13	Druggable	8	8	gamma tubulin ring complex prot	23591	0.781394	0.883409	288.7793	1185.202	112.0497	310.5889	35.11503	24.39241	49.79913	9.485995
SRP000439	D16	Druggable	9	9	putative binding protein 7a5	1731	0.941405	0.86835	180.8949	209.4964	103.6978	140.4776	10.84225	7.070599	28.18577	36.48984
SRP000469	M24	Druggable	10	10	hypothetical protein FLJ13511	54464	0.912182	0.935283	146.6728	523.9838	132.9783	224.7909	6.941006	2.267799	18.93494	38.59139
SRP000439	D18	Druggable	11	11	8D6 antigen	55964	0.916994	0.882256	155.6142	187.1925	54.14138	91.08958	38.03543	4.182601	11.05194	11.11748
SRP000439	D20	Druggable	12	12	alpha-1-B glycoprotein	23157	0.917484	0.791711	154.0979	120.4616	122.1243	100.1224	13.09587	14.02423	6.747342	1.986987
SRP000465	I08	Druggable	13	13	ataxin 2-binding protein 1	23176	0.937335	0.848266	113.7	171.2941	25.2004	33.61339	3.128092	1.025906	6.887327	31.69431
SRP000484	I21	Druggable	14	14	ataxin 2-like	151011	0.861691	0.974512	61.80003	157.4689	96.43404	89.46256	47.26814	33.45952	5.308625	29.8281
SRP000438	A05	Druggable	15	15	alpha-2-macroglobulin	55752	0.347707	0.899089	67.99509	159.2583	90.66948	154.9841	25.47527	14.39126	1.822695	33.67102
SRP000483	M07	Druggable	16	16	alpha 1,4-galactosyltransferase	143501	0.727905	0.732156	173.724	630.3225	146.8105	148.0179	44.21375	14.64555	22.50402	47.52822

Note: HCS analyzer can deal with two different well positioning modes that can be selected in the options window.

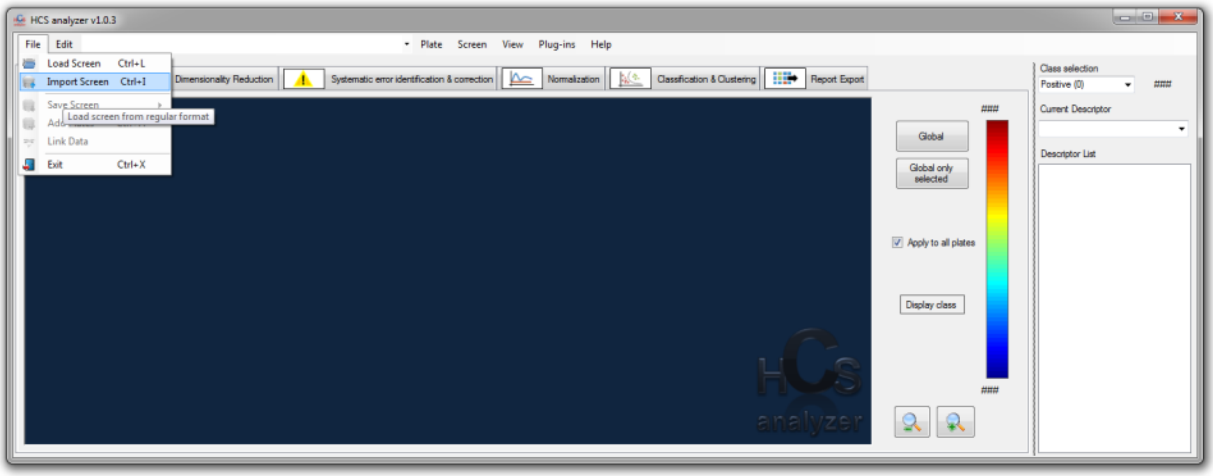


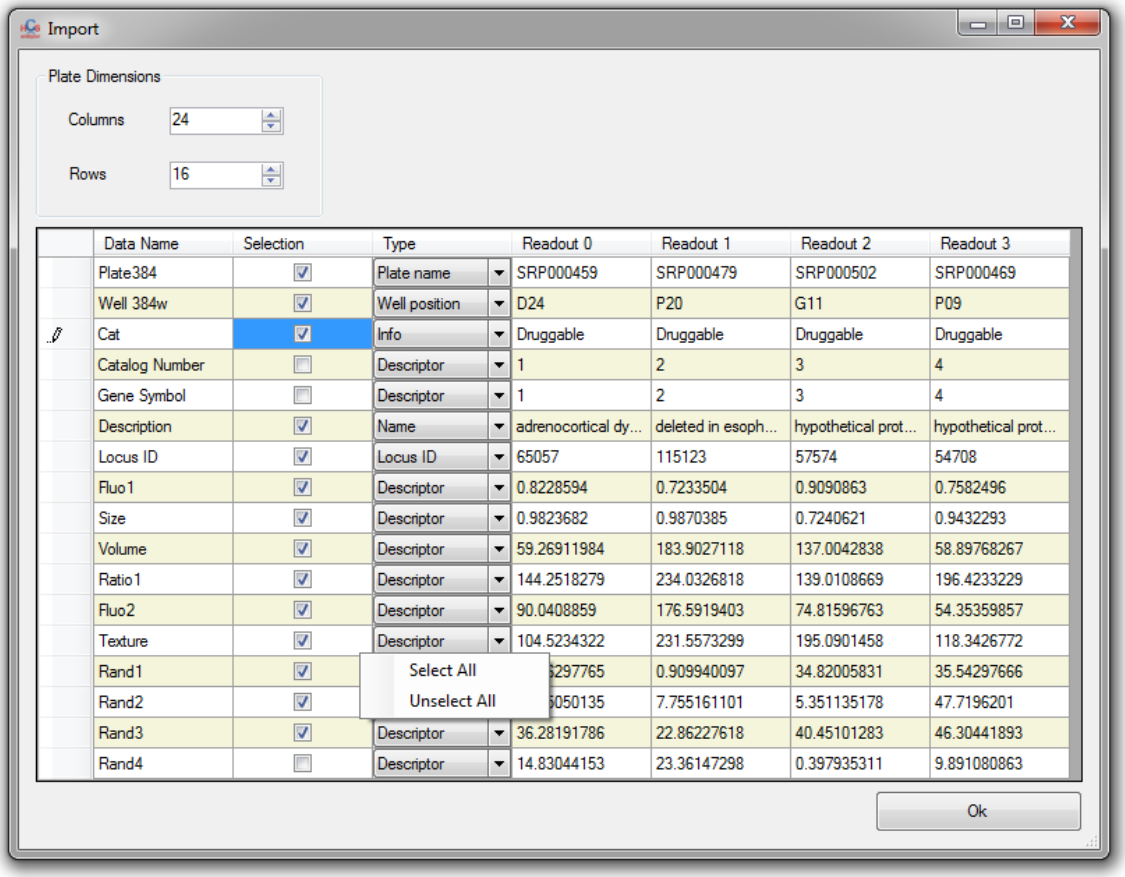
1- Single Mode: The row position is given by an alphabetical character. Row and column positions are merged together in a single column.

2- Double Mode: One column has to be created for the row position and one for the column.

Regarding the selected mode, the loading window options will be adapted.

First, click on “Import Screen” from the “File” Menu, and locate the file to be loaded.





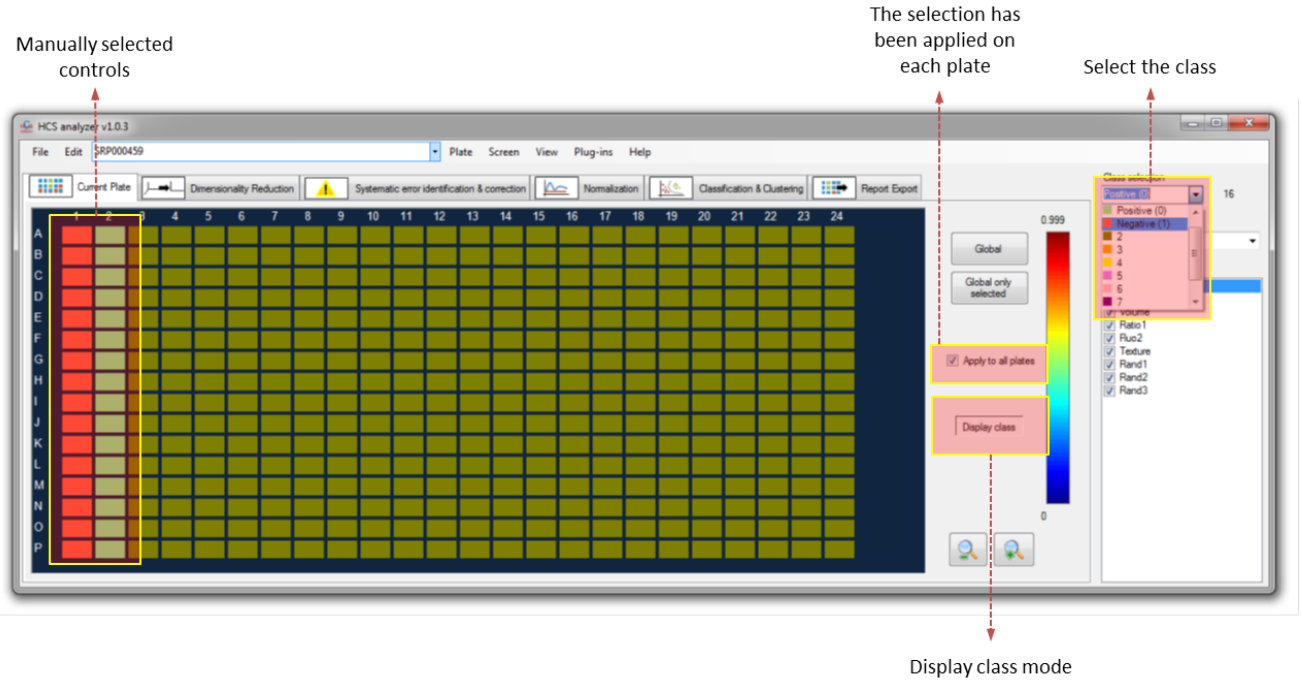
The Import window should popup such as below. The software reads 4 rows to help the user to define its choice.

Define the plate dimensions, select all the requested descriptors as well as the other useful information such as Locus ID, Info or Name (if needed) click “Ok”. A message box should appear giving you the information about the number of loaded well.

Note: In order to avoid any trouble during the process, the software does not accept any undefined well. It means that if any readout is missing, the concerned well will be automatically rejected.

Define Controls (if present)

Positive and/or negative controls have been performed during the screening process; the user can define them manually. First choose the desired class using the dedicated control, then using the mouse select the corresponding wells.

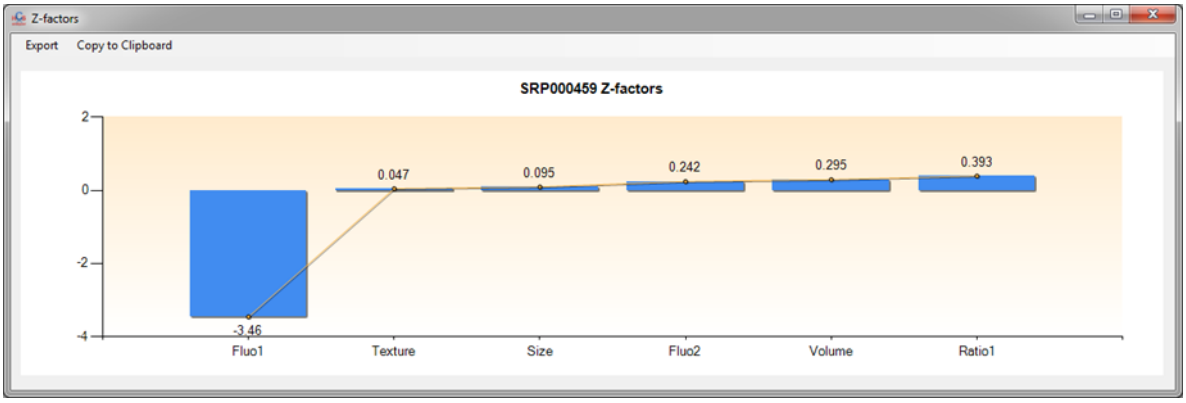


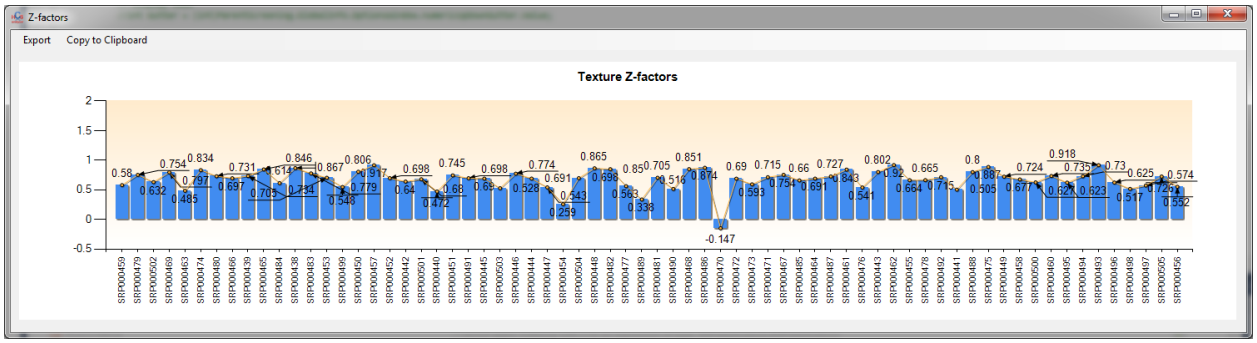
Note: It is easier to see the selection by switching to the "Display Class Mode".

Note: By default "Apply to all plates" is checked. Thus the current selection will be performed on every active plate.

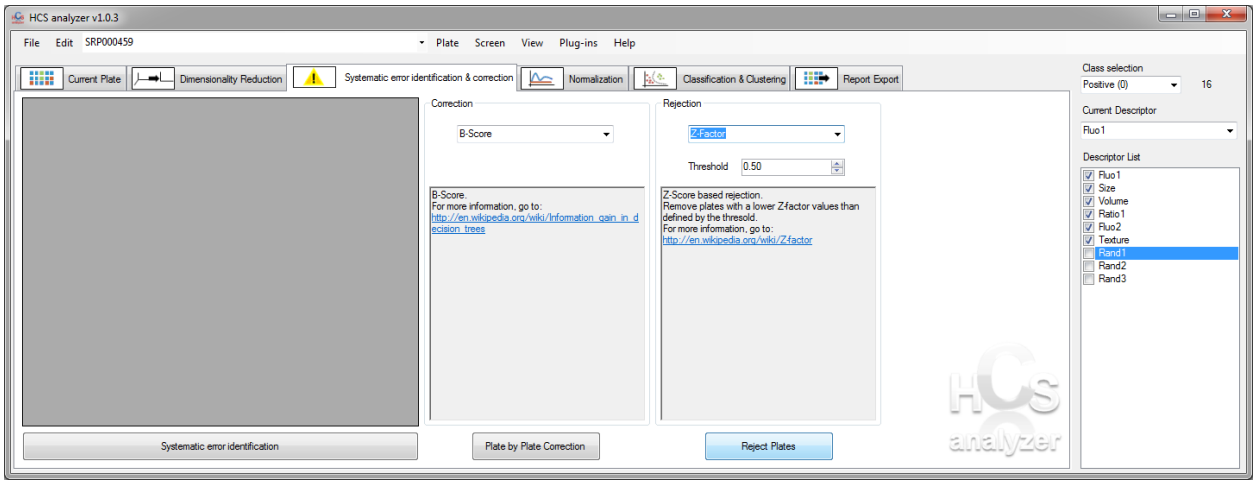
Quality control

Many different way control the quality have been implemented in HCS analyzer. The Z-factor is a commonly used one that the user can test either on the current plate by clicking on "Plate->Quality Controls->Z-Factor" or "Screen->Quality Controls->Z-Factor".

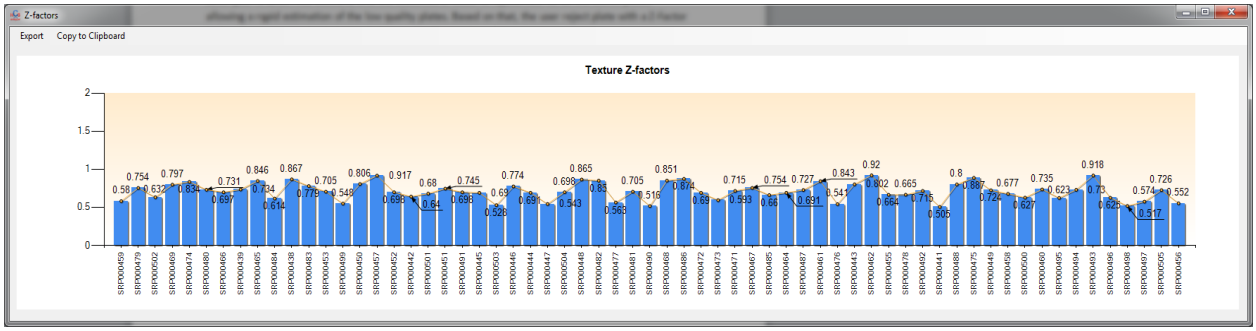




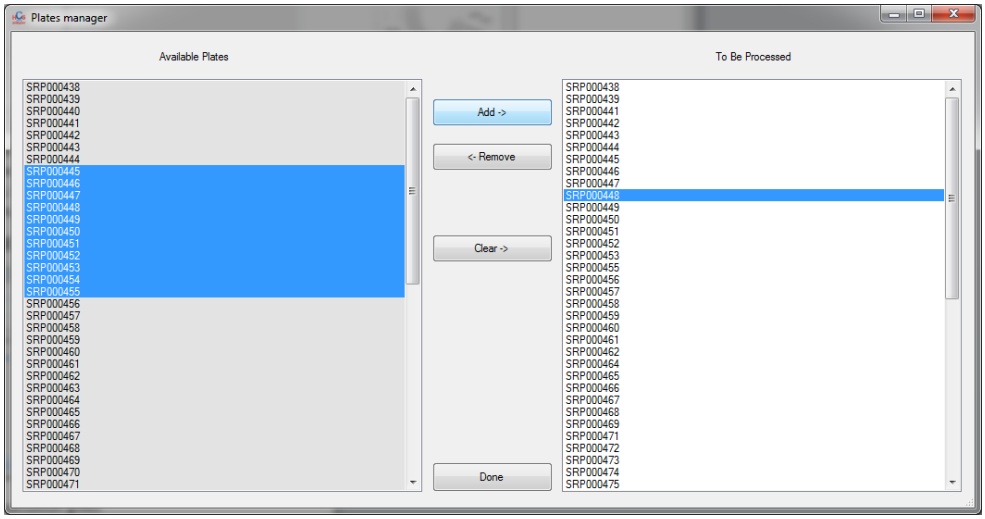
In the first case, the Z-factor is computed on the current plate for each active descriptor and the results are sorted from the highest to the lowest. If performed on the complete screen, only the active descriptor is taken into account, and the associated Z-factor is represented for all the active plates allowing a rapid estimation of the low quality plates. Based on that, the user can reject the plate with a Z-Factor higher than a certain threshold by clicking on the “Reject Pates” button.



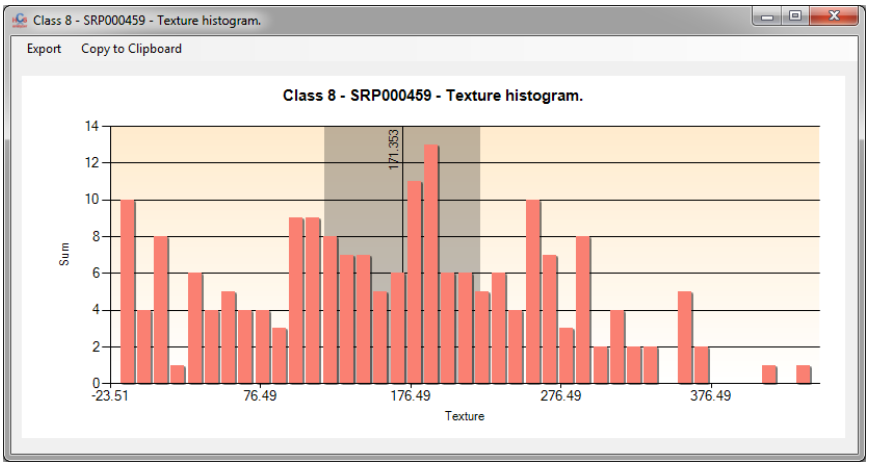
The active plate list has then been modified, and a new Z-factor visualization gives:



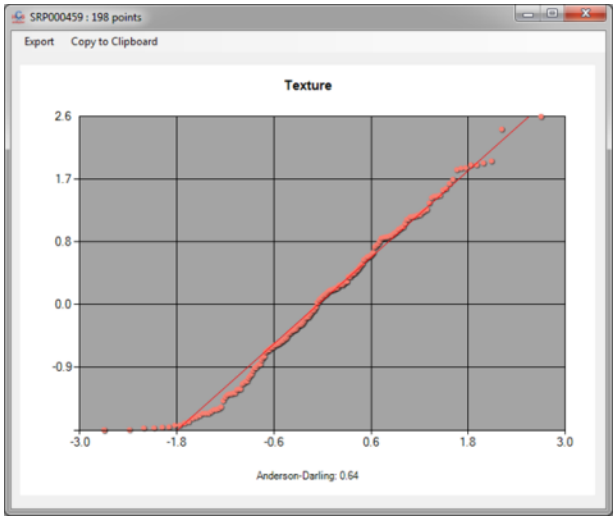
The user can always display and modified the plate list by clicking on “View -> Plates Manager”. The left column represent the complete set of loaded plates and the right column the set of active plates that will be processed.



In certain situation, assessing the normality of a descritpor readout on a control set of wells can also provide a precious information about the data quality. To check this, the user will select a descriptor, a class and then click either on “Plate -> Visualization -> Histogram”



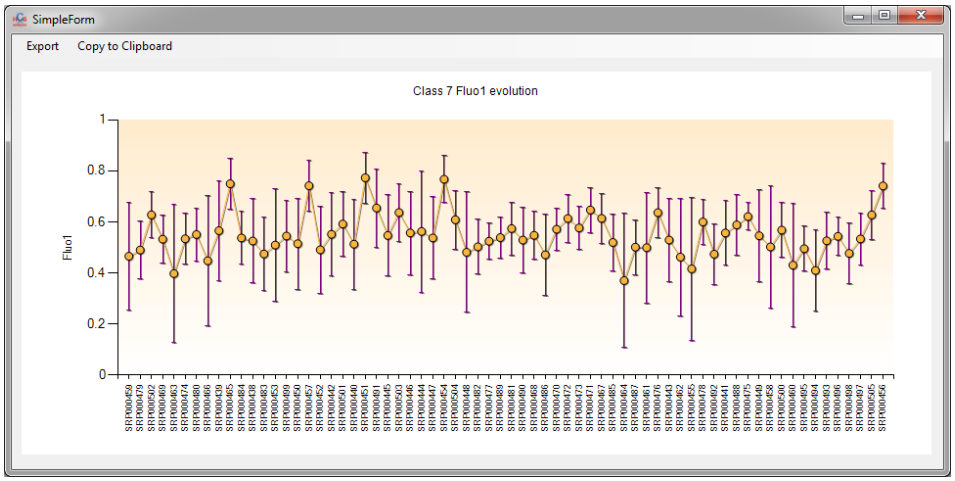
<u>Note:</u> The histogram display and computation is highly modifiable through the "Edit->Options" window.



or "Plate -> Quality Controls -> Normal Probability Plots".

<u>Note:</u> Those displays can also be generated for complete screen by going to "Screen" menu.

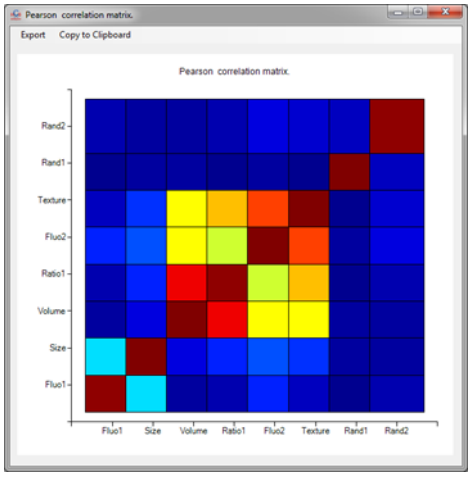
If the user is interested by assessing the quality of a particular set of well among the complete screen, he can click on "Screen->Quality Controls->Descriptor Evolution" and check the evolution of the average values and standard deviations of the selected class.



Descriptor Selection

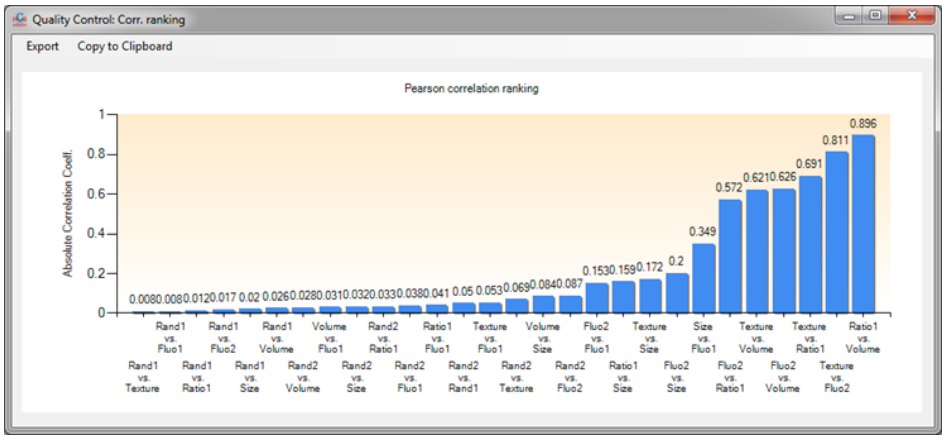
The presented application is clearly targeting screening multivariate data. In such context one important step in the data analysis is identify redundant source of information. Typically the identification of two correlated descriptors has to be identified either to perform better in the classification or clustering step, but can also be of great interest in understanding some biological mechanism of action.

If the user is interested by the quality of the descriptors in term of information, he can click on "Plate->Quality Controls->Correlation Matrix" (or its equivalent "Screen->Quality Controls->Correlation Matrix" to display the correlation matrix of the active descriptors). The process will display first the correlation matrix (by default the absolute value of the Pearson coefficient),



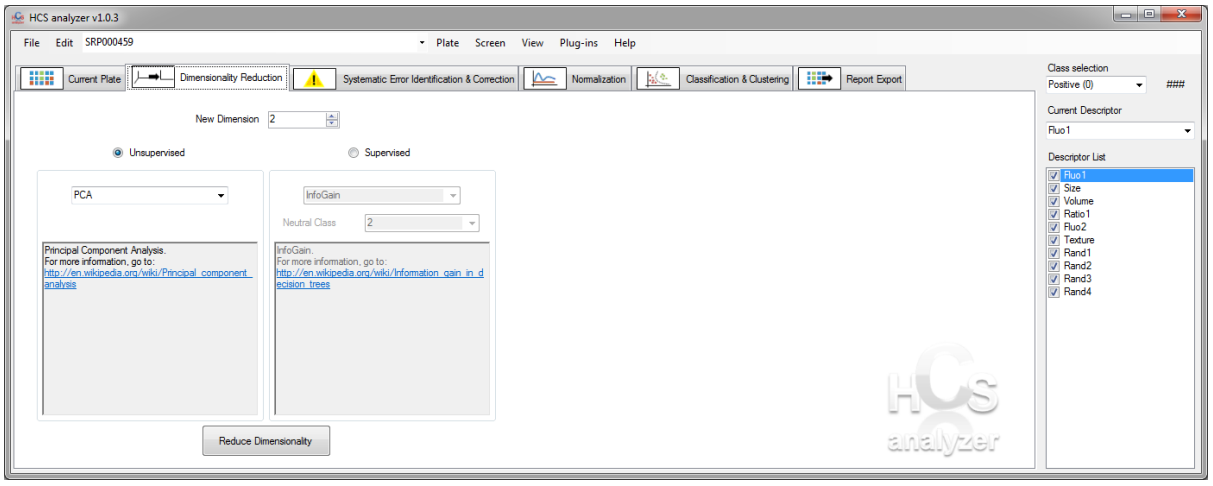
<u>Note:</u> For this process all the activated wells are taken into account (no specific class).

then the ranking of the coefficient to easily identify correlated coefficients.



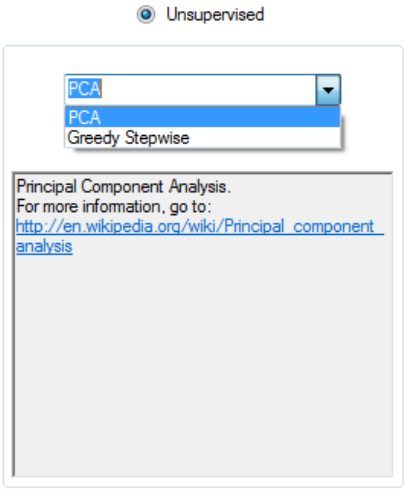
<u>Note:</u> The user can select the Spearman coefficient and remove the ranking display by changing the options in the "Edit->Options" menu.

HCS Analyzer provides different approaches for automated feature selection. Those methodologies are available by clicking on the “Dimensionality Reduction” tab.

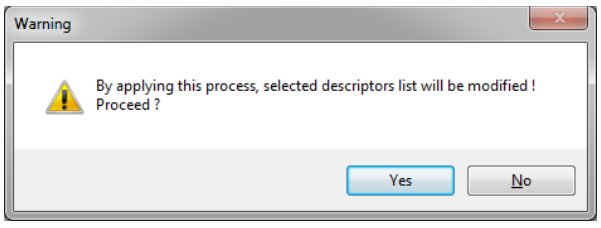


At this stage, one can choose between two different kinds of approaches: Unsupervised or Supervised. Compare to Supervised, Unsupervised dimensionality reduction methods do not required any predefined classes.

Unsupervised: In this version of the software, two unsupervised methods are available: PCA or Greedy Stepwise.



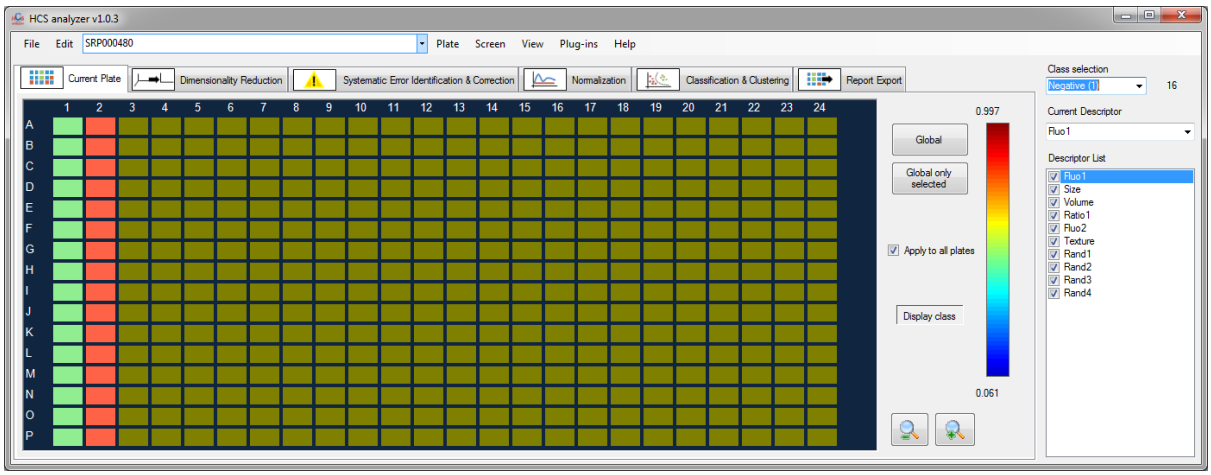
<u>Note:</u> whatever the feature selection algorithm is selected, a warning message box will appear prior to the process, reminding the user that the active descriptor list will be modified.



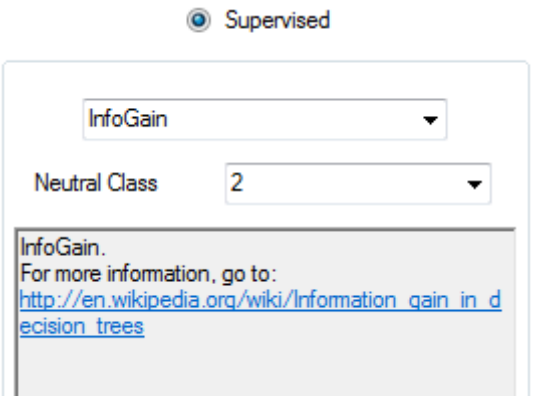
If the PCA algorithm is selected, all the active wells will be used to identified the most important descriptors (i.e. the features maximizing the deviance of the data), reducing the dimension to the selected value.

Unlike the PCA, the greedy stepwise does not reduce the number of feature to a user selected value. The final number of features is automatically chosen by the software.

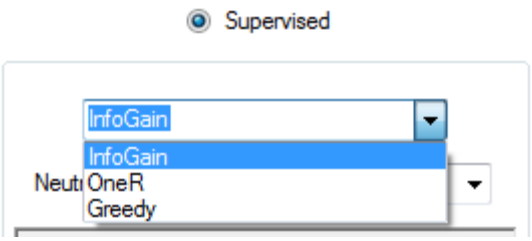
Supervised: As said above, supervised methods require a pre-defined set of classes. Let's select two controls as class 0 and 1 and keep the rest of the plate as class 2.



The algorithms will then choose the features that separate the more efficiently the classes, but will not take into account the well corresponding to the neutral class. In such way, the user can focus on identifying important features required for maximizing the difference between controls only.



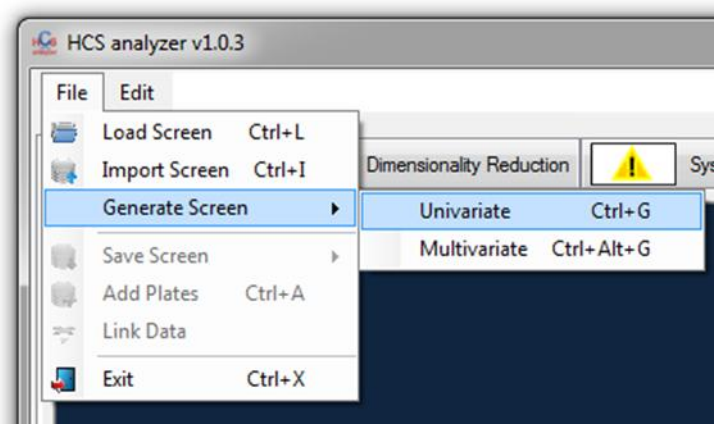
At this stage, the user has access to 3 different methods: InfoGain, OneR and Greedy.



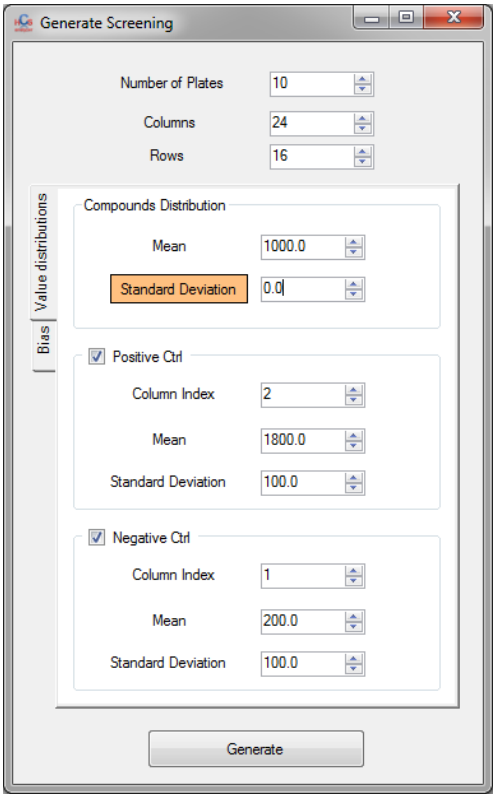
InfoGain will operate regularly by reducing the dimensions to the user selected value. OneR operates the same way; however, in this case, the user has a feedback from the algorithm through the text box, giving him the opportunity to check the weights associated to every descriptor.

Phenotypic clustering

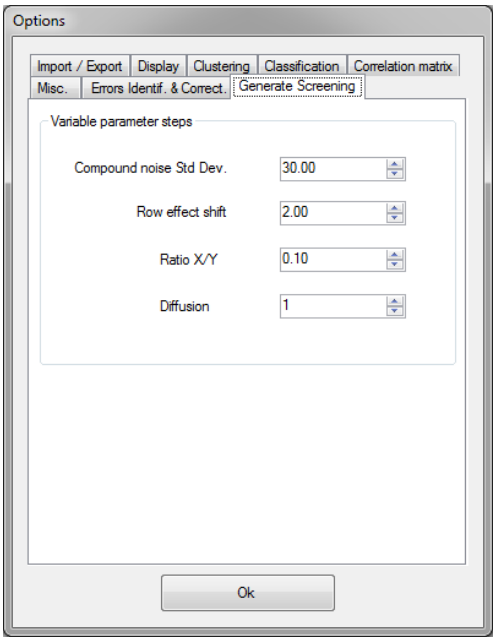
For the purpose of this section, let's start buy generating an univariate artificial screening by clicking on *file->Generate Screen->Univariate*



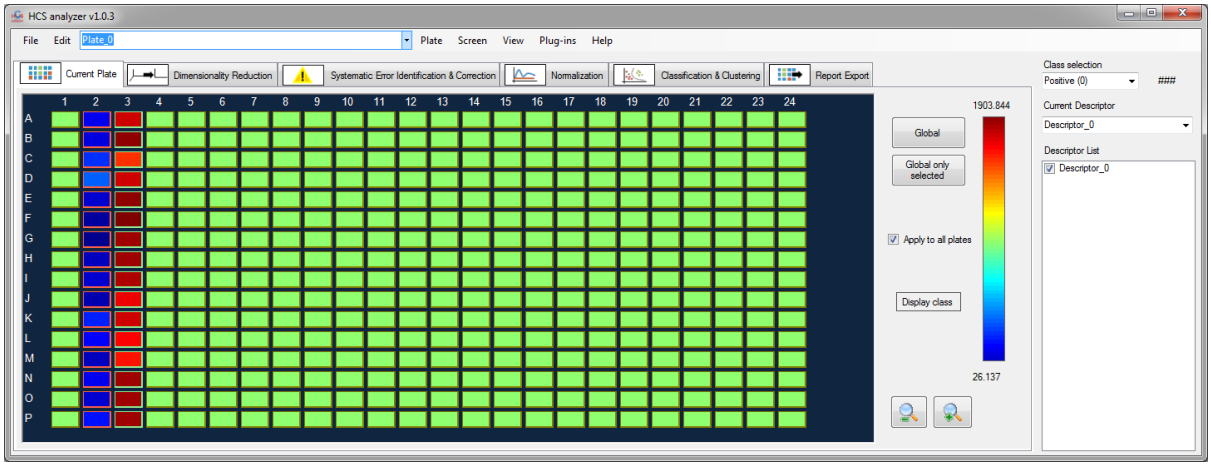
Change the standard deviation of the compound distribution to 0 and push the standard deviation popup button as it is shown on the picture below.



This means that the compound distribution (Gaussian) will have a variable standard deviation starting 0 and increasing by 30 (defined on the option window as show below) on every plate (here 10 plates are defined).

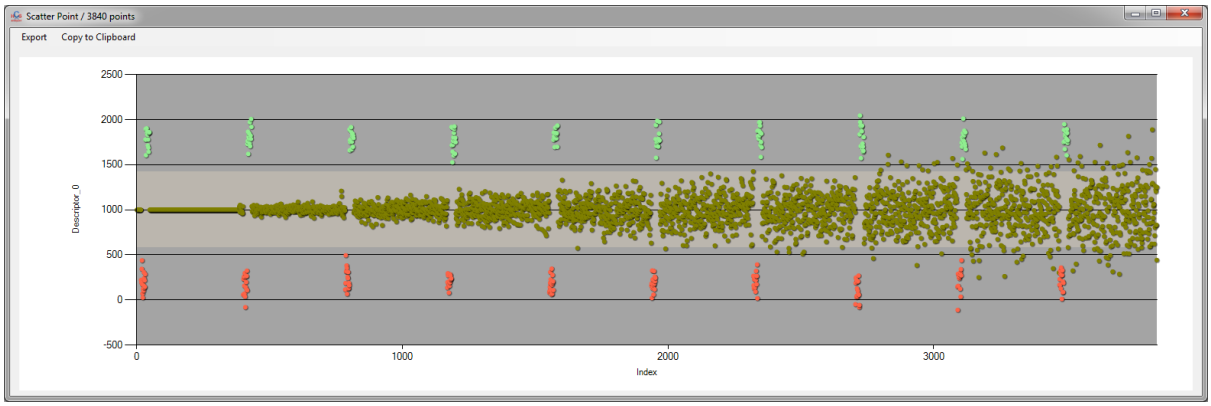


Click Ok to generate the 10 plates. Starting from:

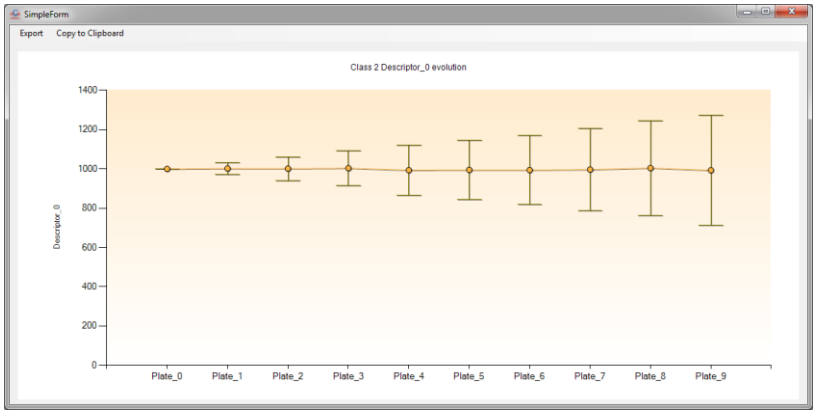


To

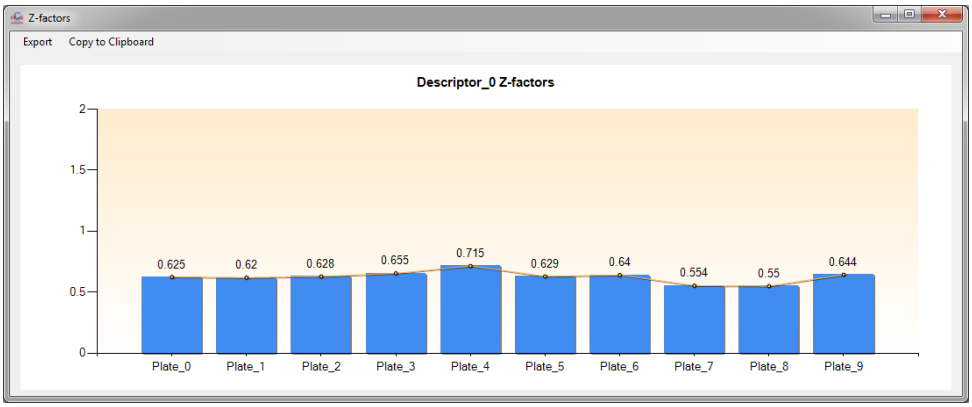
The evolution of the screening data quality can be monitor by clicking on *Screen->Visualization->Scatter Points*



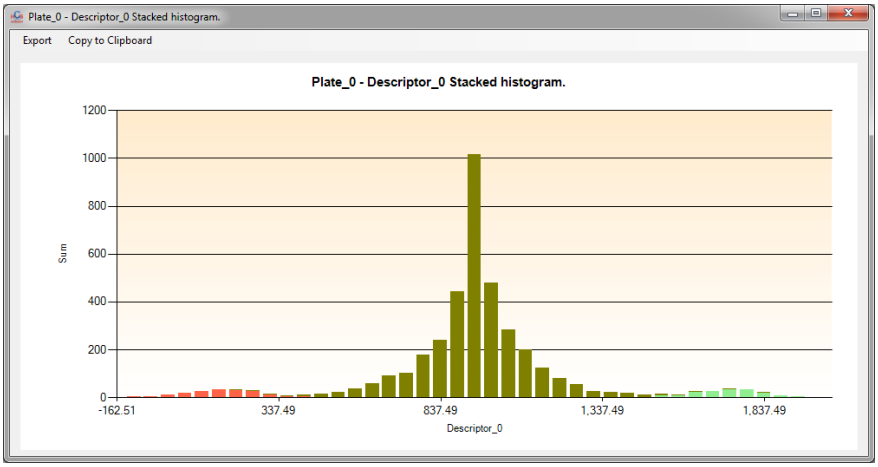
More precisely, using *Screen->Quality Controls->Descriptor Evolution* and by selecting the class 2 (here the compounds), we obtain:



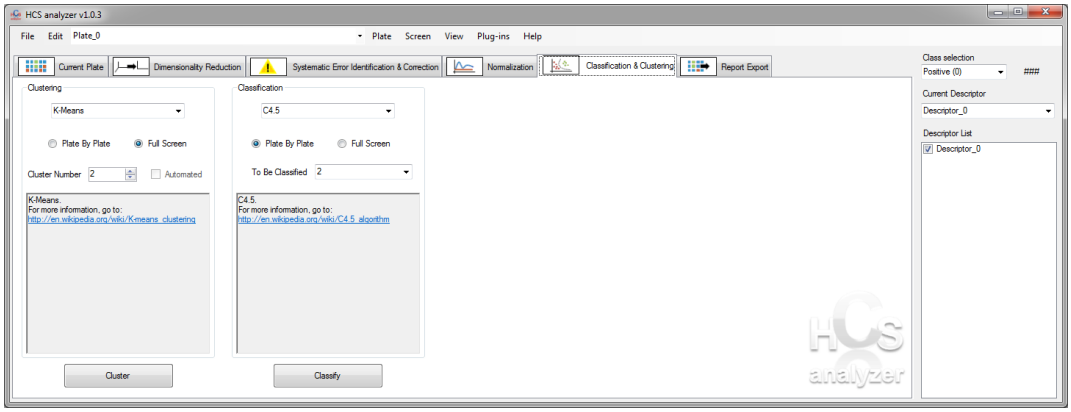
<u>Note:</u> As only the compounds distribution has been subjected to a variation over the plate index, a quality control such as the Z-factor should not vary over the screening. This can be assess by clicking on Screen->Quality Controls->Z-score



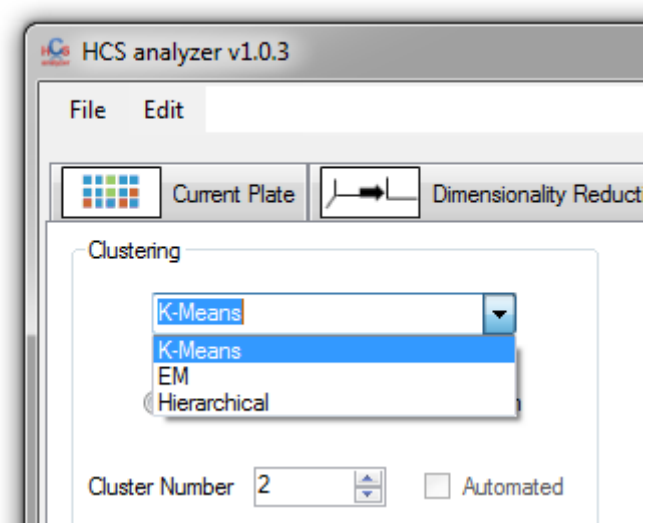
The stacked histogram (available by clicking on *Screen->Visualization->Stacked Histogram*) provide also a convenient way of displaying the data distribution:



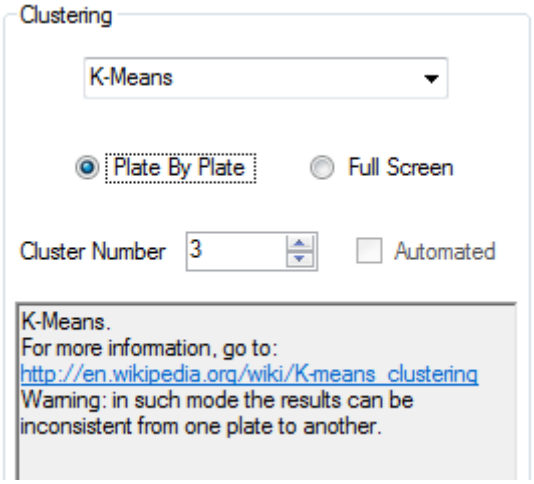
Let's jump directly to the clustering step by clicking on the *Classification & Clustering* tab.



First select the clustering algorithm:



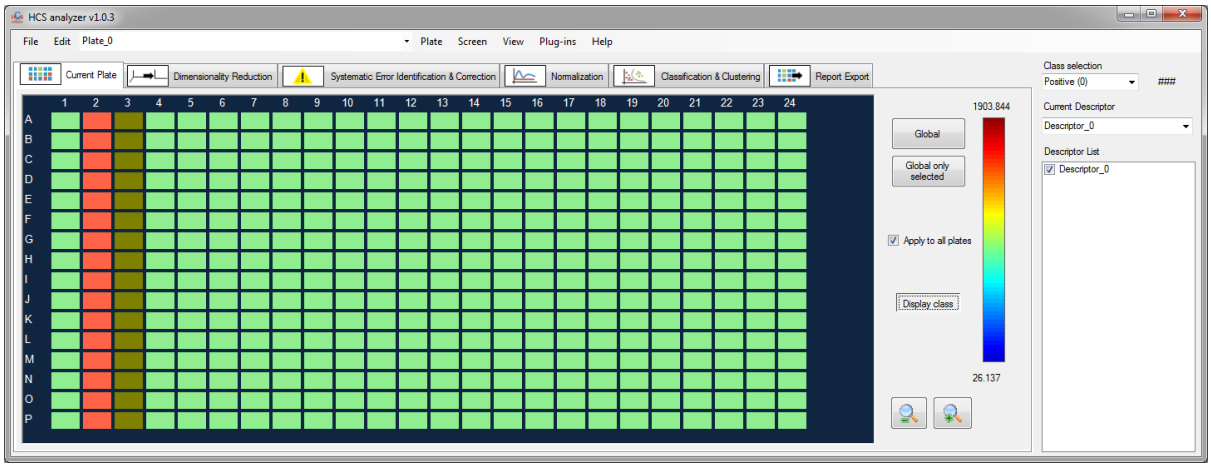
When using K-Means, the user has to select the number of expected class. Let's choose 3 (2 controls + 1 compound distributions).



First, we will start by checking the *Plate By Plate* radio button. In this case, the clustering will be operated on each plate independently.

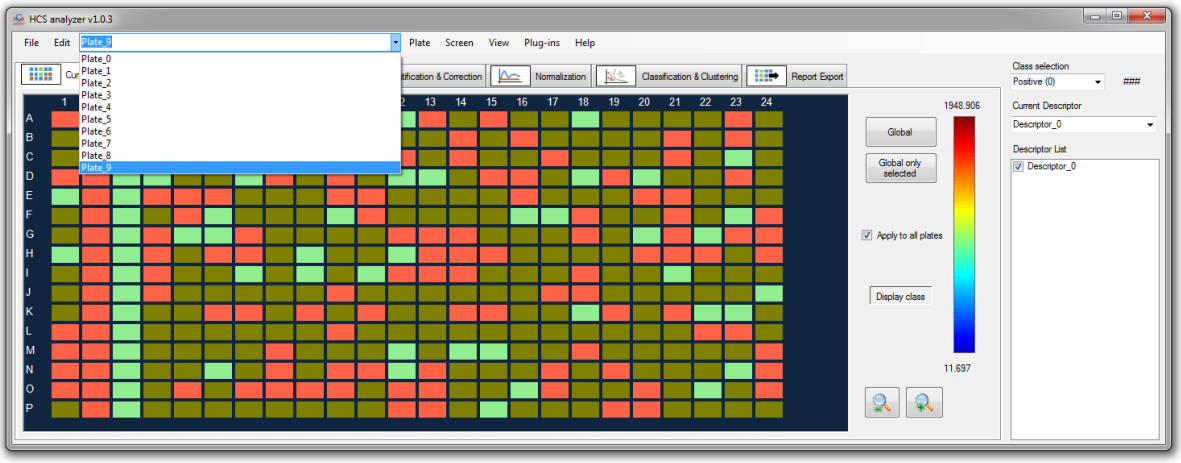
<u>Note:</u> In such case, no consistency is guarantee for the class association from one plate to another.

Click on Cluster to process the entire screening. Let's come back to the plate visualization tab, and press the popup button *Display Class* to monitor the result.

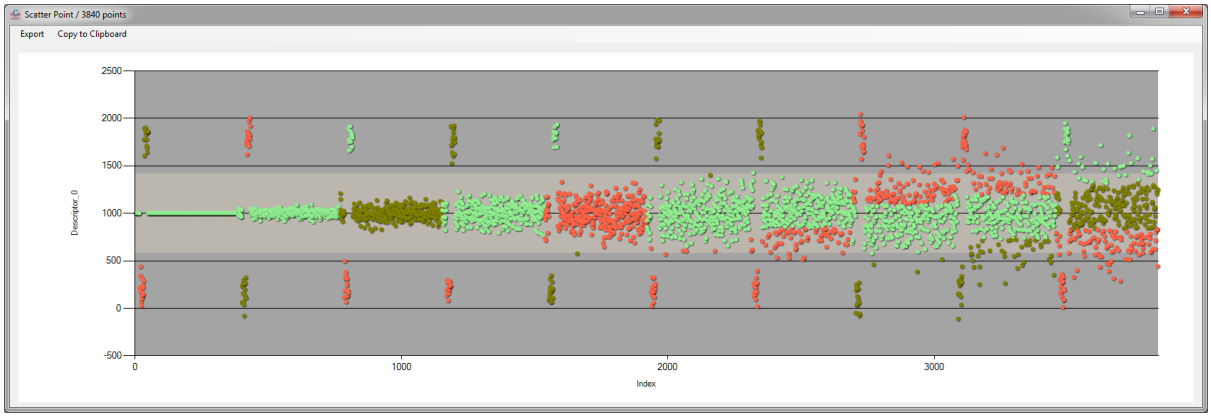


As said above, the class association is random (here the compound and the controls have been inverted). However, in term of clustering, the process was successful by differentiating the controls and compounds properly.

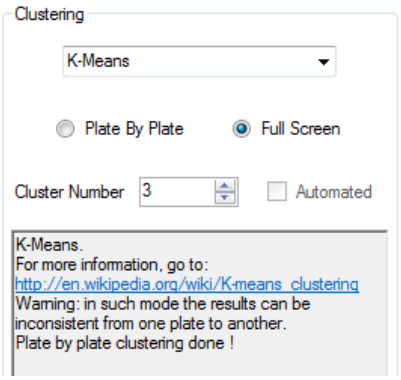
As the noise is increasing the clustering becomes fuzzy as show on the next figure:



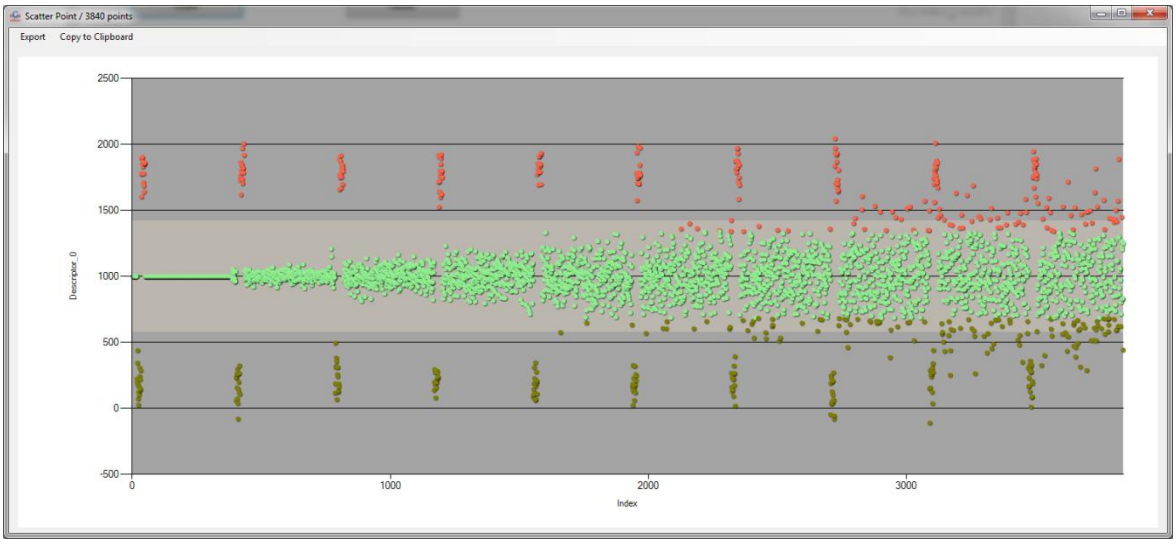
In such case, the *Screen->Visualization->Scatter Points* give us a good overview of the phenomena.



Let's apply the same clustering method but this time performed in one time over the entire screen.

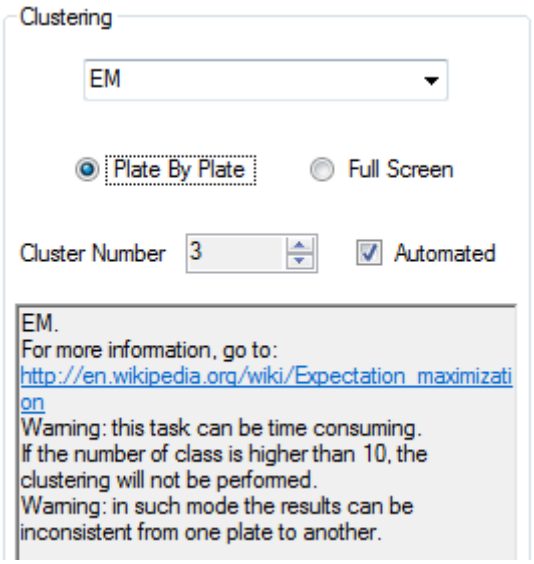


The resulting scatter points graph gives:



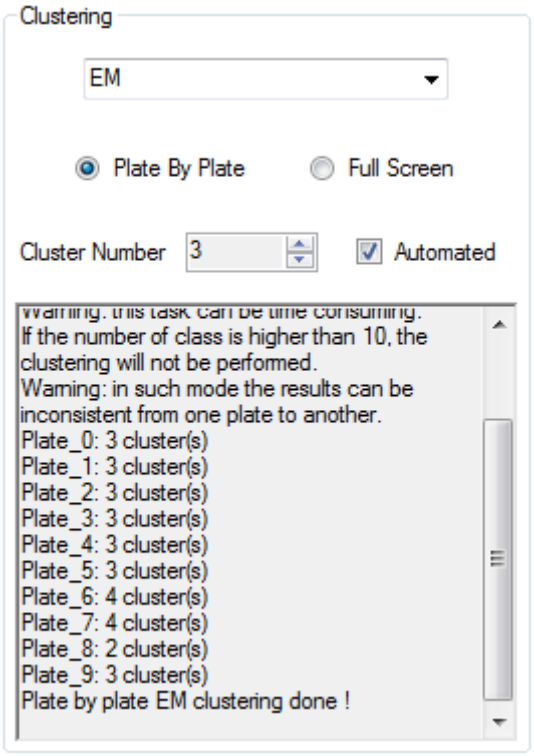
As usual the class association is still random, but now as the screening is process globally, there is a consistence from one plate to another.

By choosing another clustering approach such as the Expectation–maximization algorithm (EM), it is possible to automatically identify the optimal number of cluster. Let's perform it his way

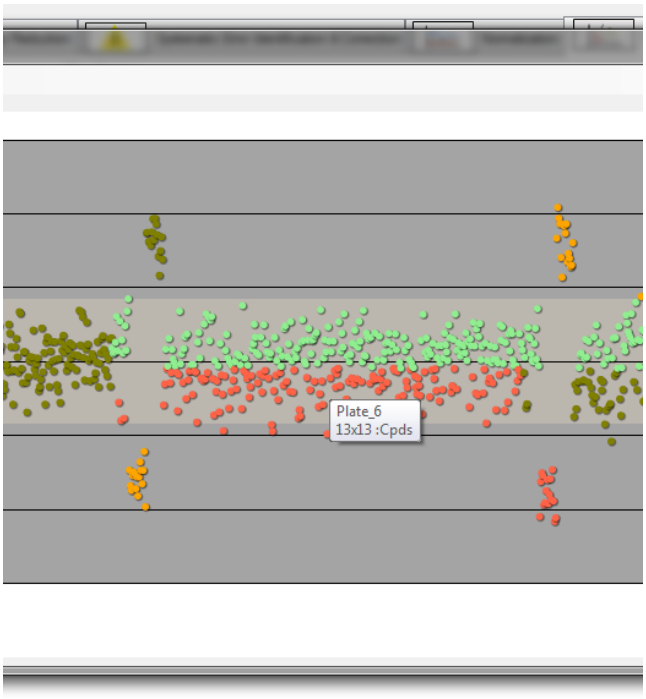


<u>Note:</u> This operation can be time consuming.

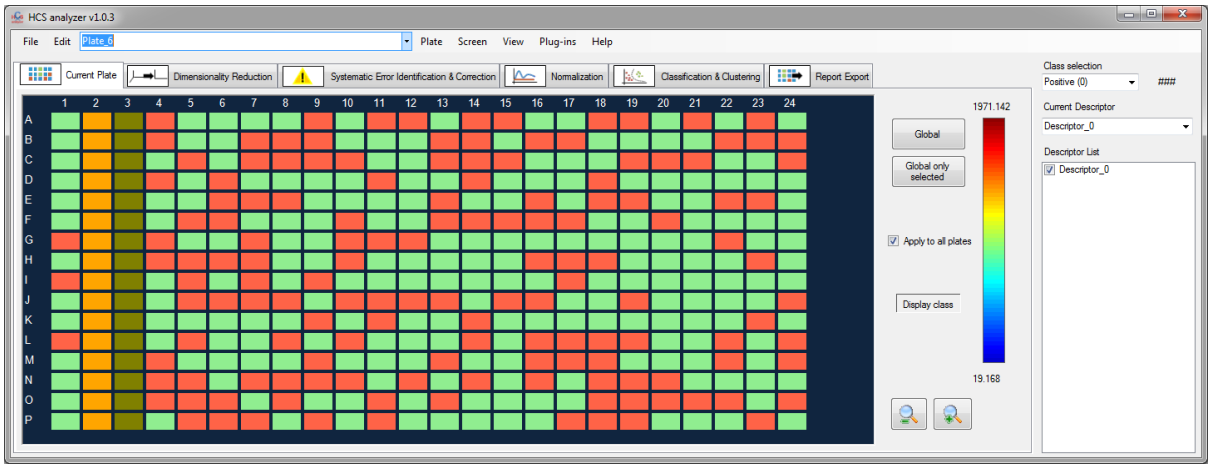
The number of identified clusters can be check on the dedicated console.



It is interesting to see that on plate_6, four clusters have been detected.

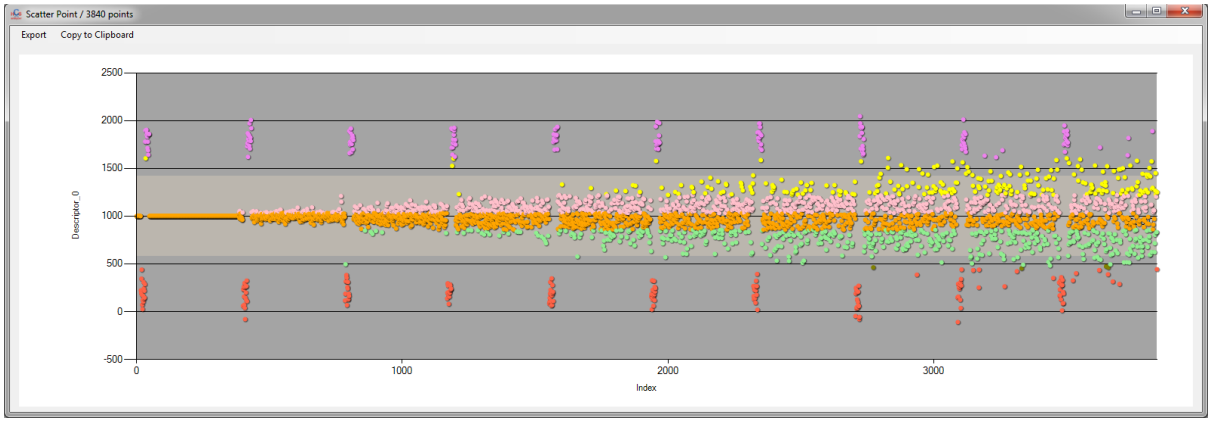


In this case, the controls have distinguished as well as two phenotypes within the compounds class.

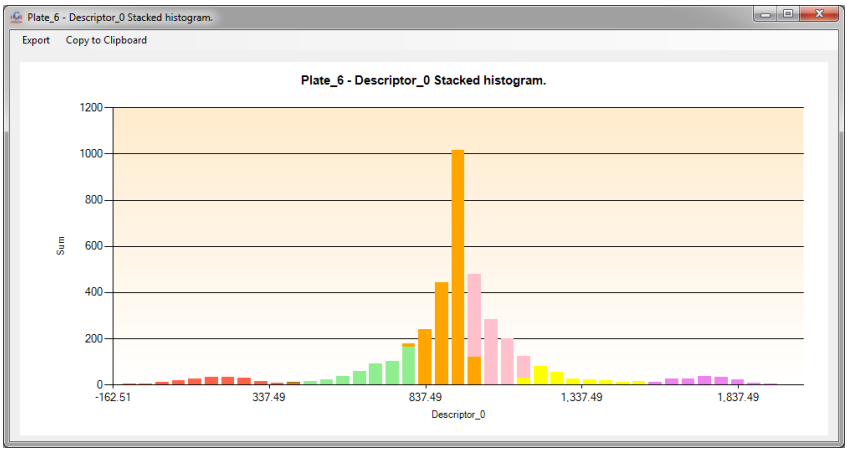


If the same operation is performed in the entire screening in once, 7 clusters are obtained.

<u>Note:</u> if the cluster number identified by the EM is higher than 10, then the process is cancelled.

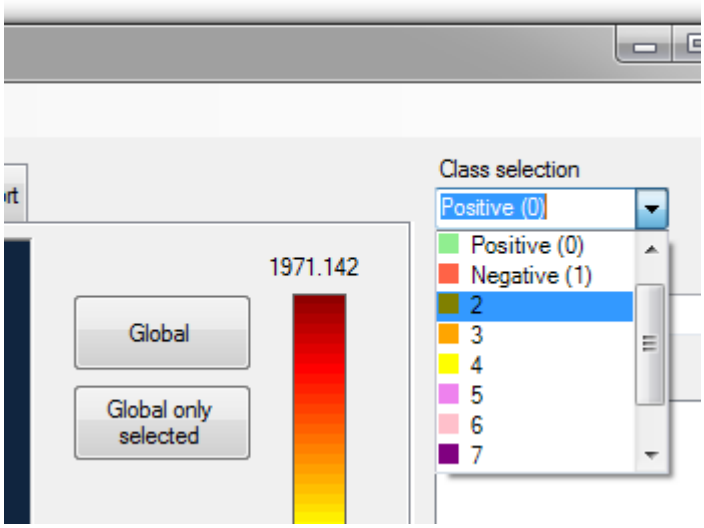


For such an approach, the stacked histogram is the most relevant way for understanding the clustering.

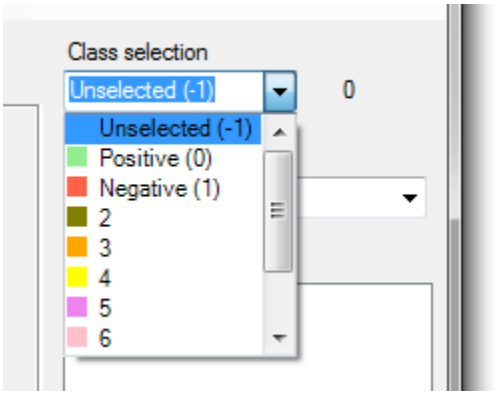


The controls have been properly identified; however, the compound distribution has been split in five distinct Gaussian distributions.

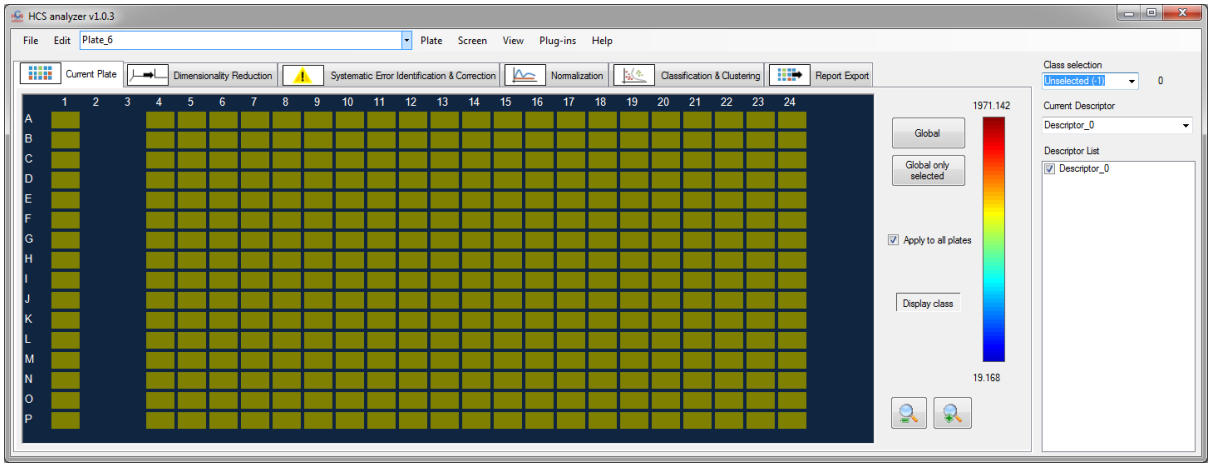
Now we can convert the entire screen wells into class2 (click on *Global*).



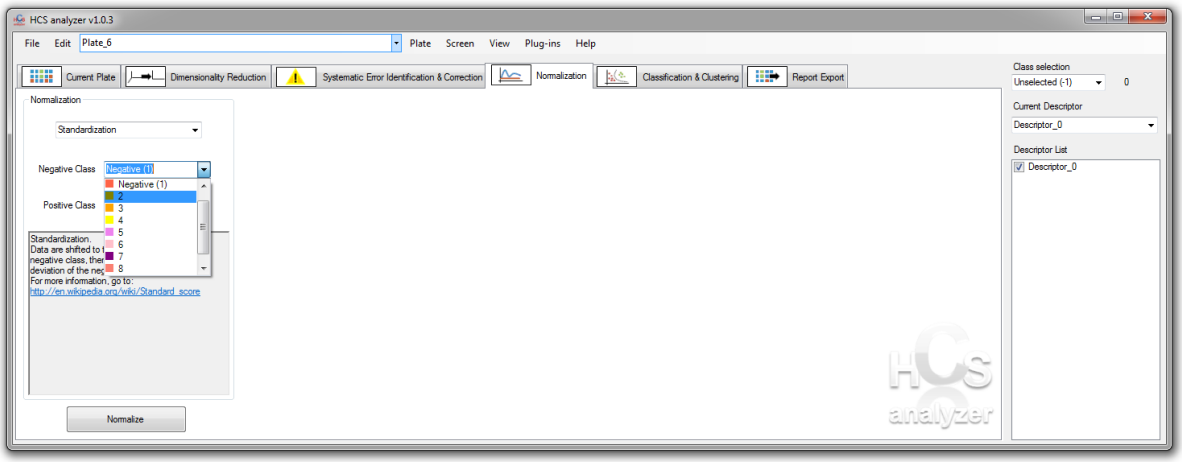
Then un-select the two controls, in order to remove them from further analysis.



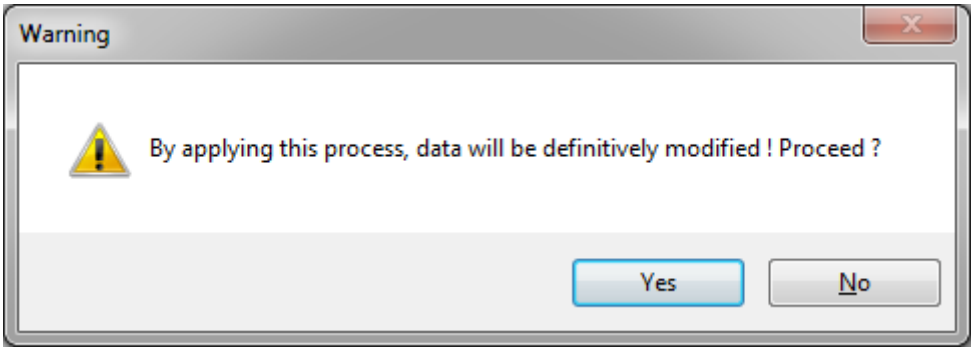
The plate design should look like this:



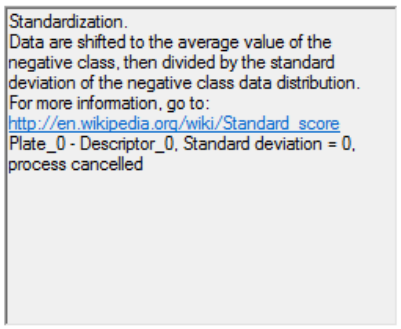
The idea now is to perform a standardization of the data. For that, click on the *Normalization tab,* select *Standardization* and apply it to the class2 previously defined.



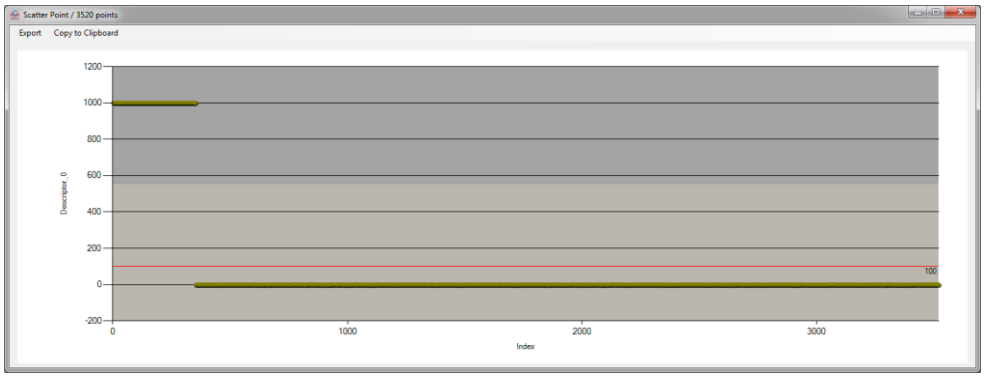
This operation is one of the few process that will modify the data.



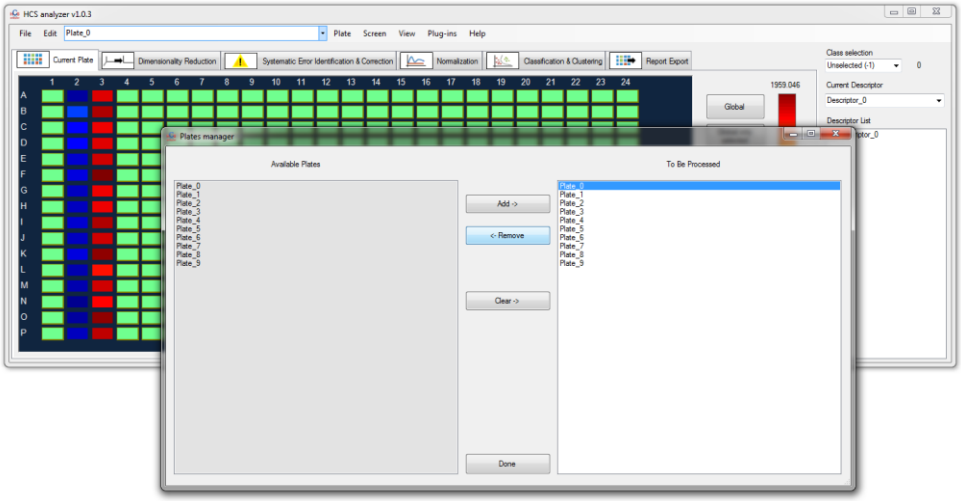
Due to the data structure, the first plate has not been processed (null standard deviation).



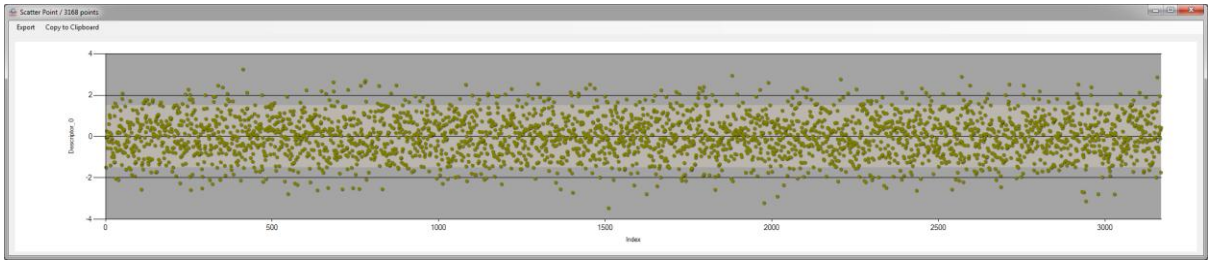
The full screening scatter points are then difficult to read.



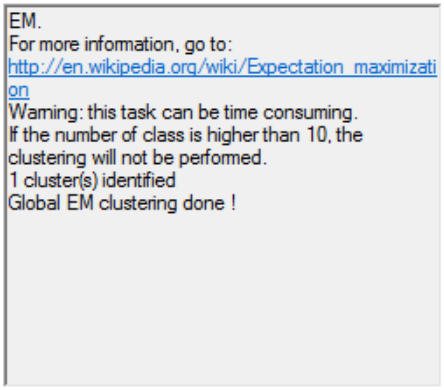
This can be fixed by removing the plate0 from the list of active plate (*View-> Plates Manager*).



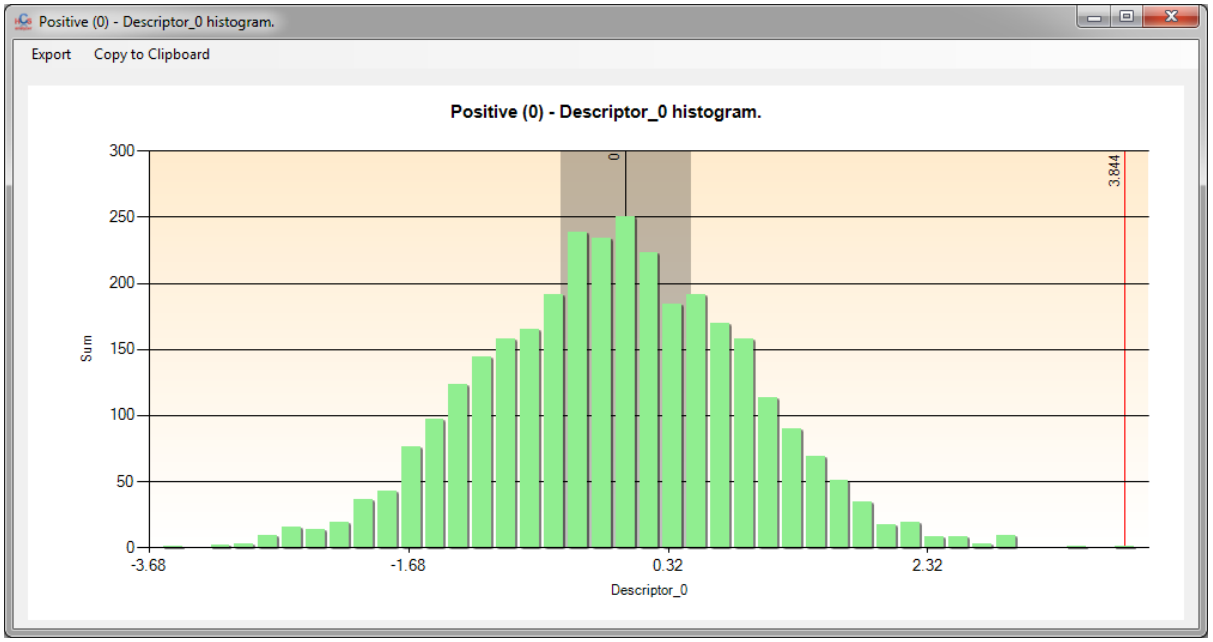
The scatter points become then:



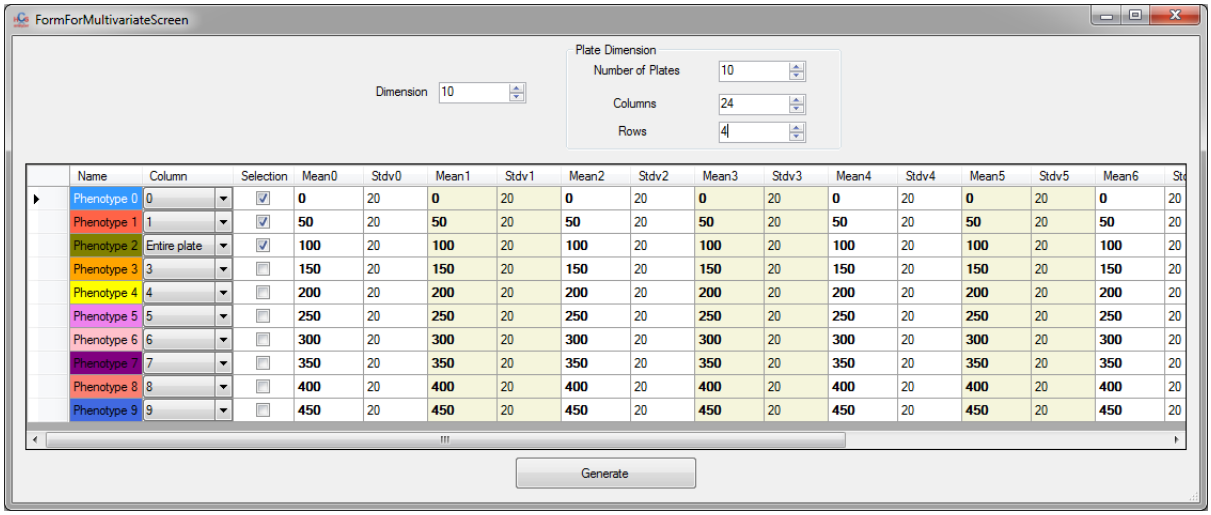
Finally, we can apply an EM clustering with an automated number of cluster identification.



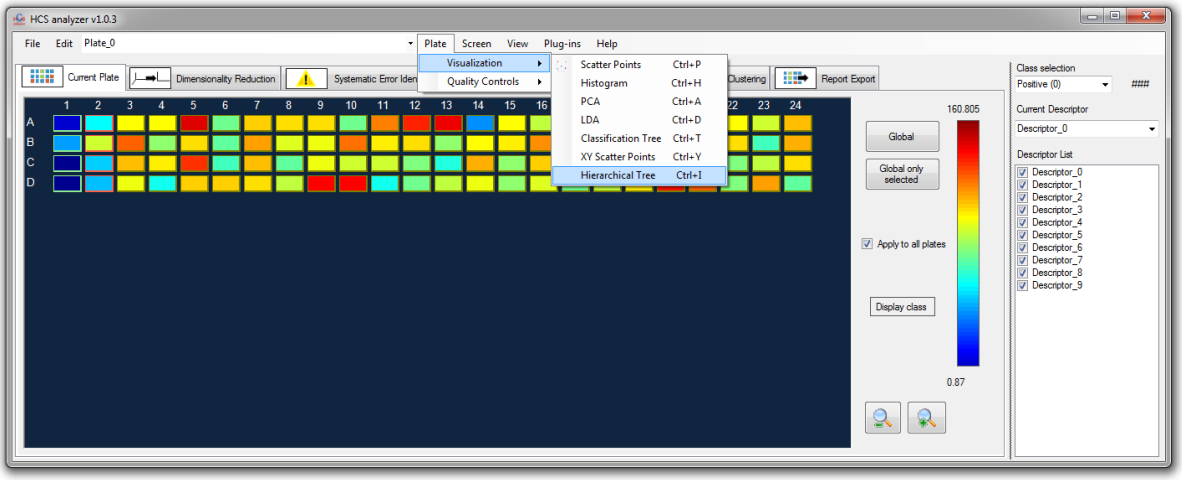
As expected, only one cluster has been identified, as it is display on the histogram visualization.



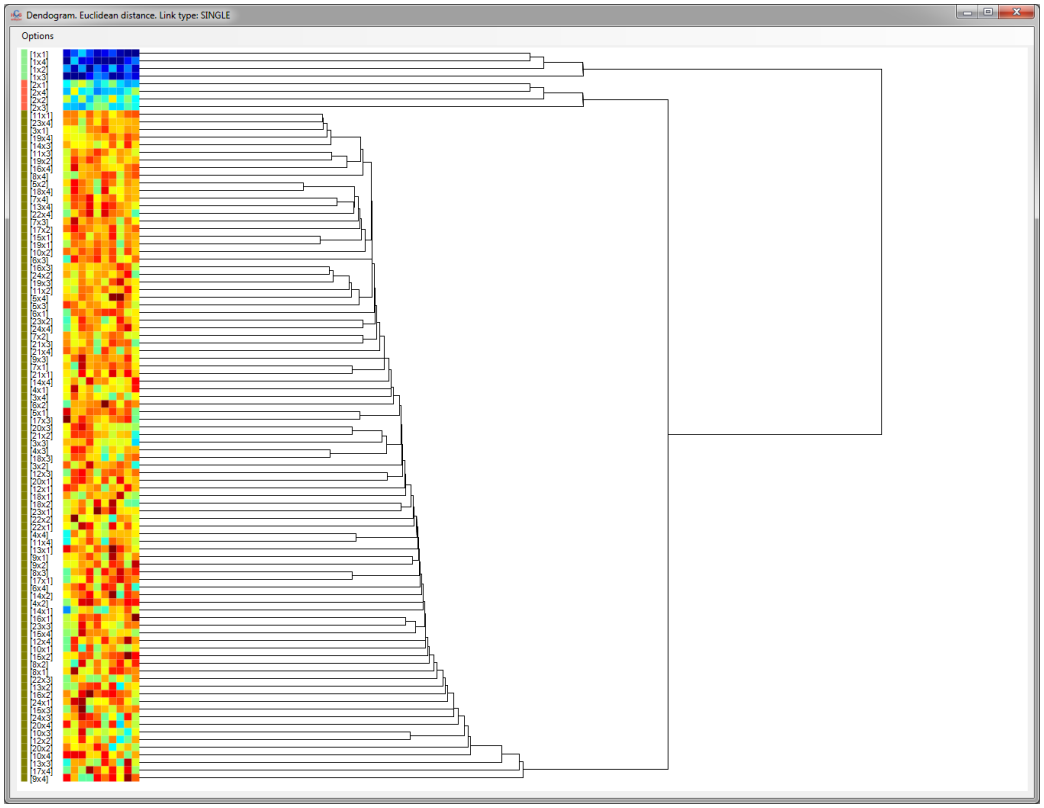
HCS analyzer provides also a dendogram visualization of the data. The visualization is common in biology and provides a convenient way to check the signature distributions over a limited number of experiments. Let's start by generating a multivariate screening composed of a single plate with a 10 dimensions signature as follow:



Go to the menu *Plate->Visualization->Hierarchical Tree*



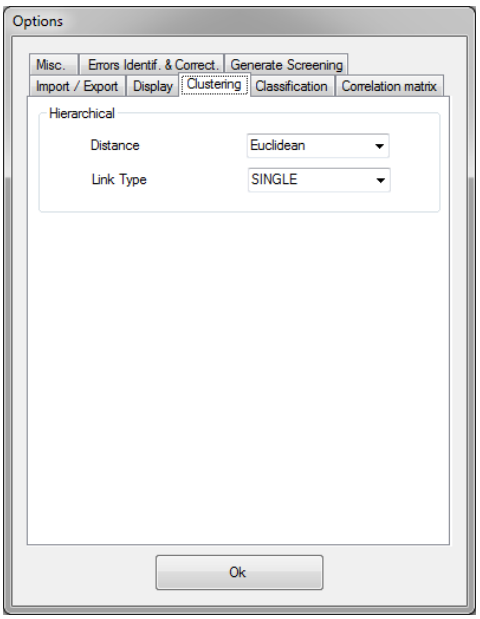
a window as follow should appear



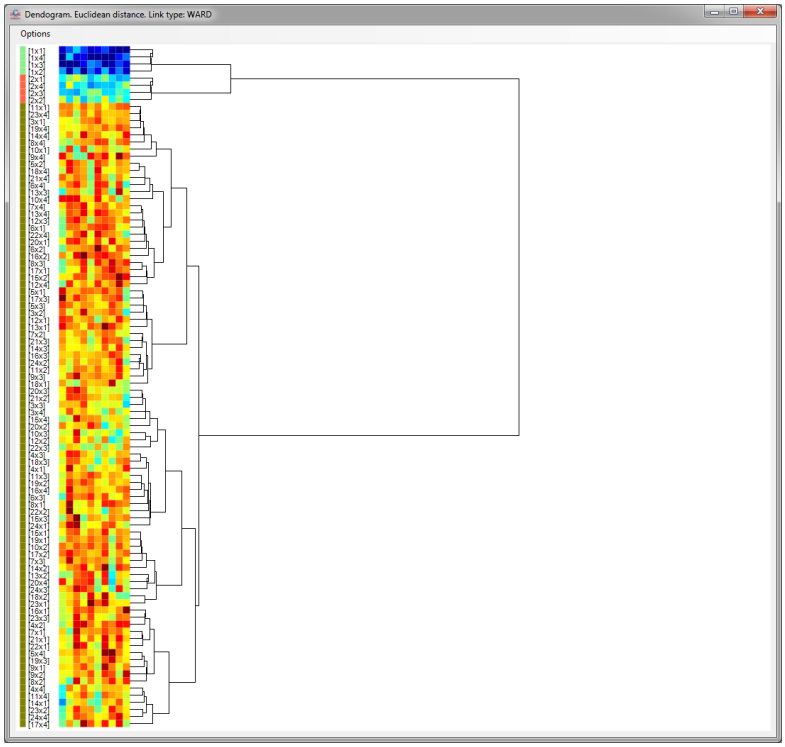
On the very left side of the dendogram, the user can see the color related to the well classes, followed by the well position. Then, next comes the signature associated to the well.

Note: each descriptor is separately normalized between -1 to +1.

And finally the tree by itself. One can access the tree options by displaying the *option window* and selecting the *Clustering->Hierarchical* box.



Note: those options will be also applied for the hierarchical clustering process

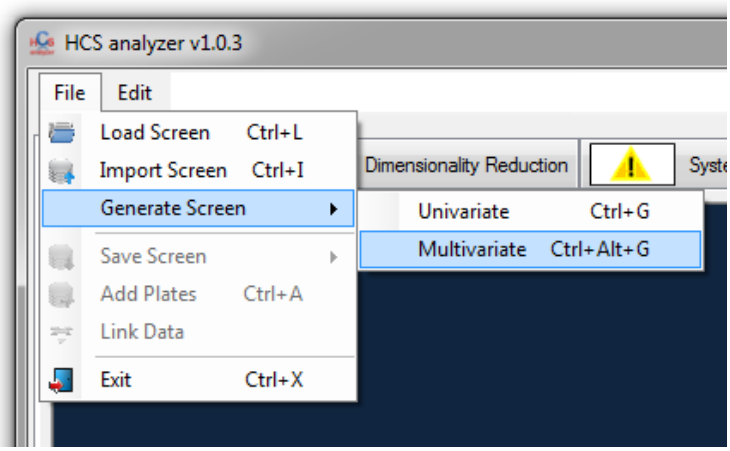


The user has to remember that such hierarchical clustering computation is highly related to the number of wells involved. Thus, the computational time can increase a lot for high number of data points, and the related dendogram display becomes quickly unclear.

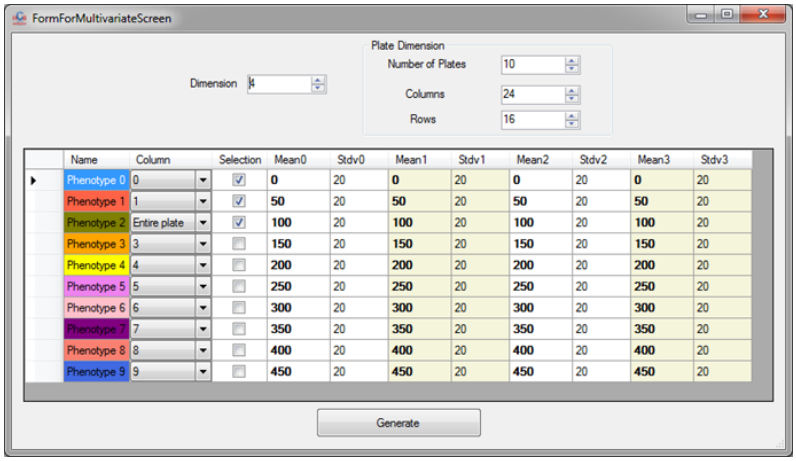
Phenotypic classification

For a clear understanding of the phenotypic classification, we will start this section by generating an artificial multivariate screening.

Click on *file->Generate Screen->Multivariate*.

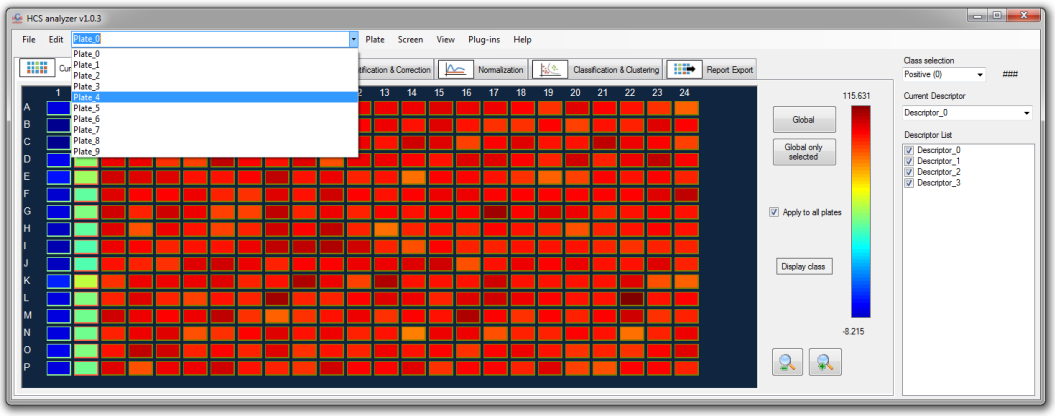


Increase the dimension of the data to 4, and keep the original set-up for the other parameters.

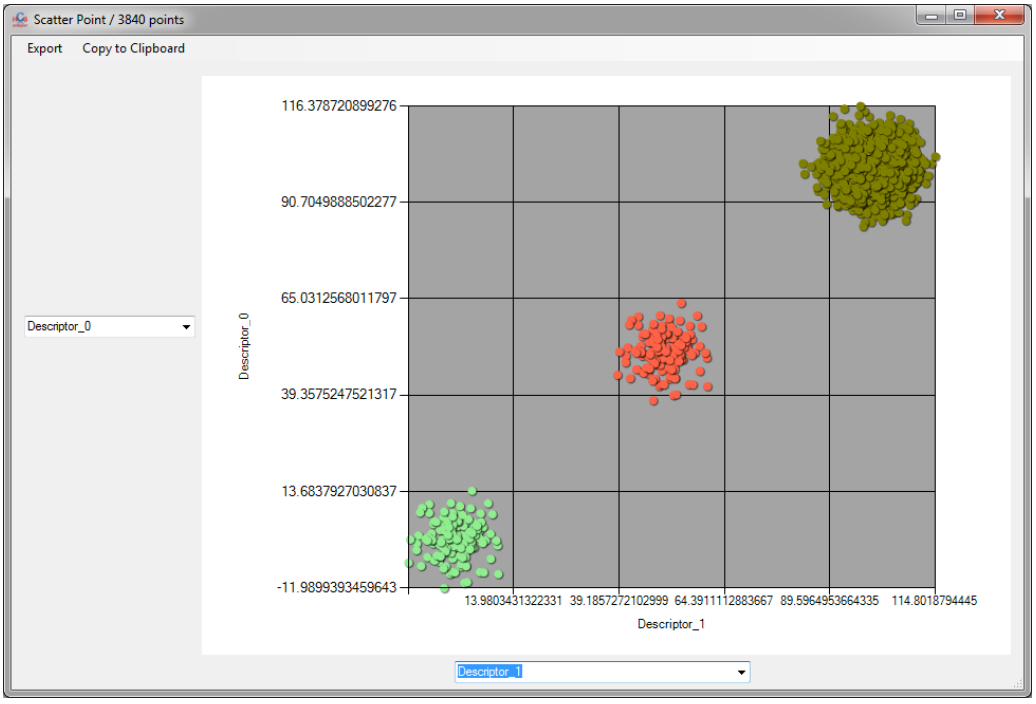


In such a mode 10 plates containing 3 different phenotypes (represented by multivariate Gaussian distribution, in a 4 dimension space, with means and standard deviations defined in the control). Phenotype 0 and 1 will be located respectively at column 0 and 1.

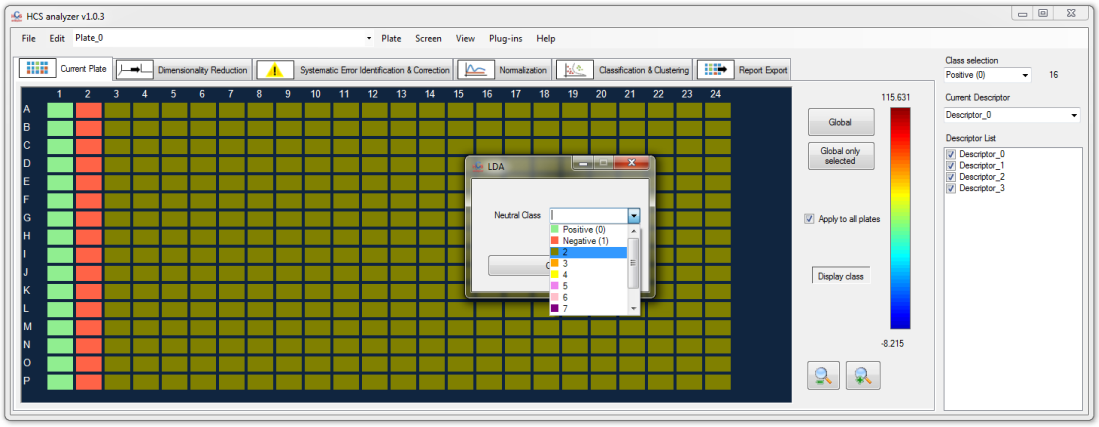
Click Generate.



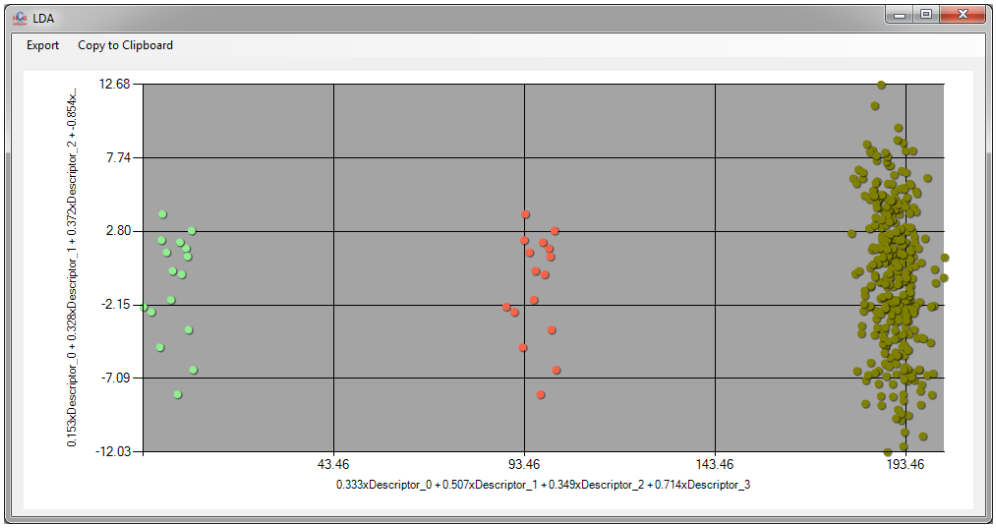
You can look at the points distributions along the different axis by clicking on *Screen (or Plate) -> Visualization -> XY scatter points*, and by choosing different descriptors along the axis.



In such example, the phenotypes are clearly defined and well separated. We can also visualize the linear discriminant analysis (LDA) on a plate. To do that, click on Plate->Visualization->LDA

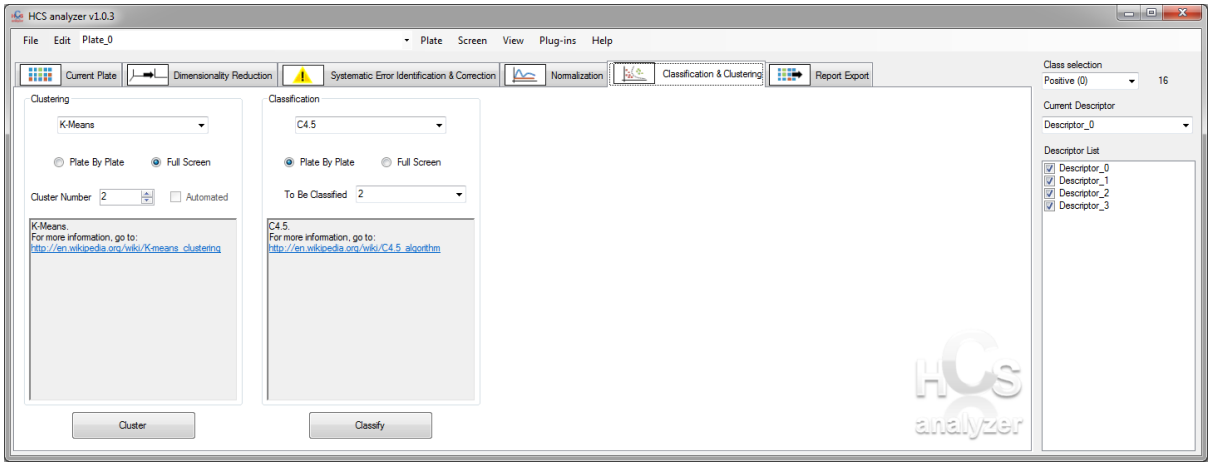


The user has to select a class that will not be taken into account for LDA computation but by projected in the resulting space. Let's choose Phenotype 2 (in dark green).

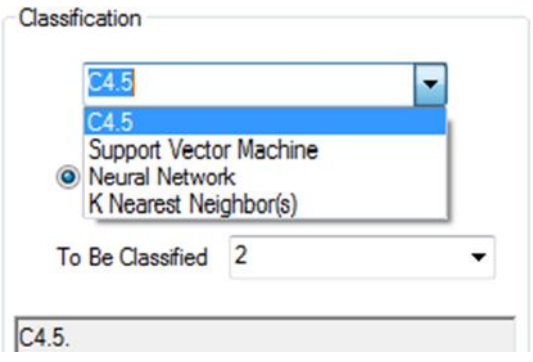


The horizontal axis of this space represents the linear component that maximize here the separation between the phenotypes 0 and 1 (light green and red).

If we consider those phenotypes as controls, we can then perform a supervised classification by displaying the corresponding tab.



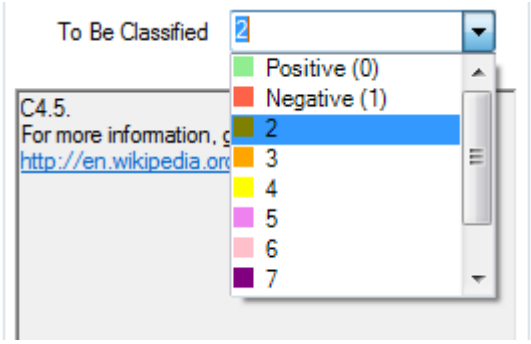
The classification algorithm can be selected here.



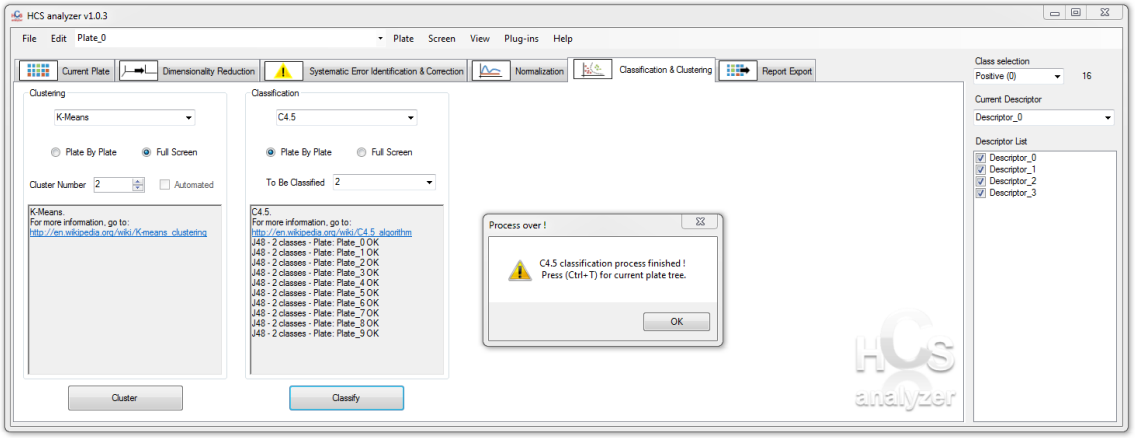
if the *Plate By Plate* radio button is checked, then each plate will be considered independently. It means that, the learning step will be performed on each plate and the classification will follow. If the *Full Screen* radio button is checked then, the learning step will be performed on all the screening data, then applied to all the wells.

Note: In a full screen process, the user will have to perform normalization step to make the data consistent.

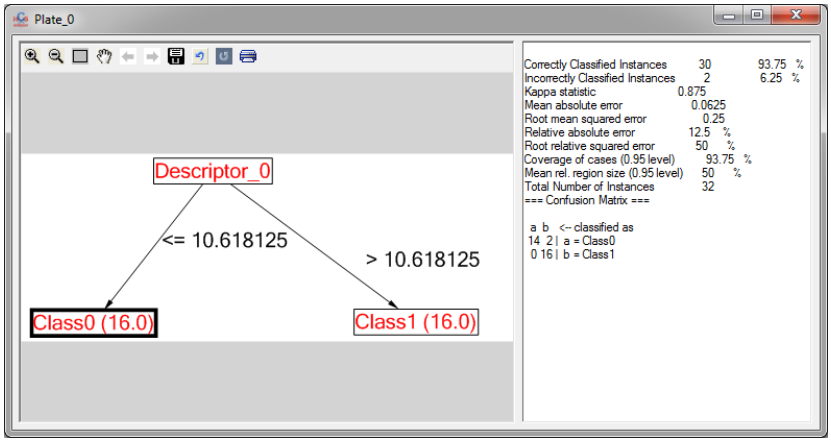
The user has to specified a *To Be Classified* class. This class will not be taken into account for the learning step, but will be (as all the other wells) classified. Typically, this represent the screened compounds.



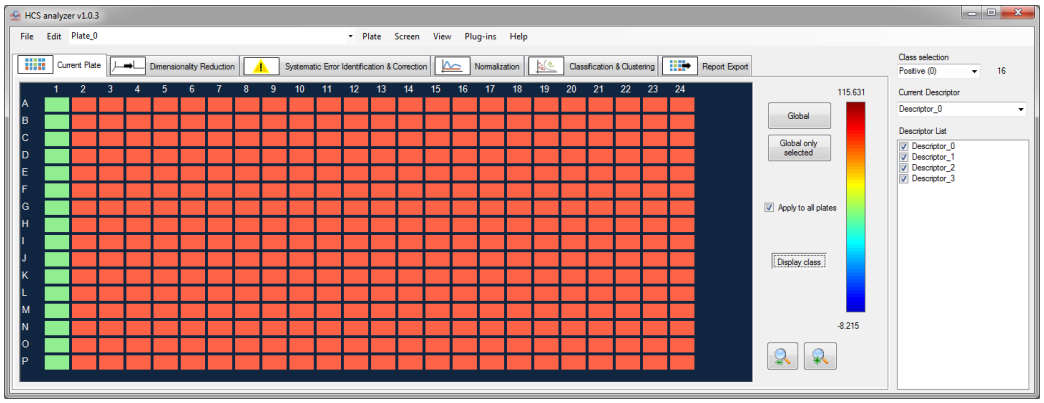
Let's choose Class 2, and then press *Classify* button.



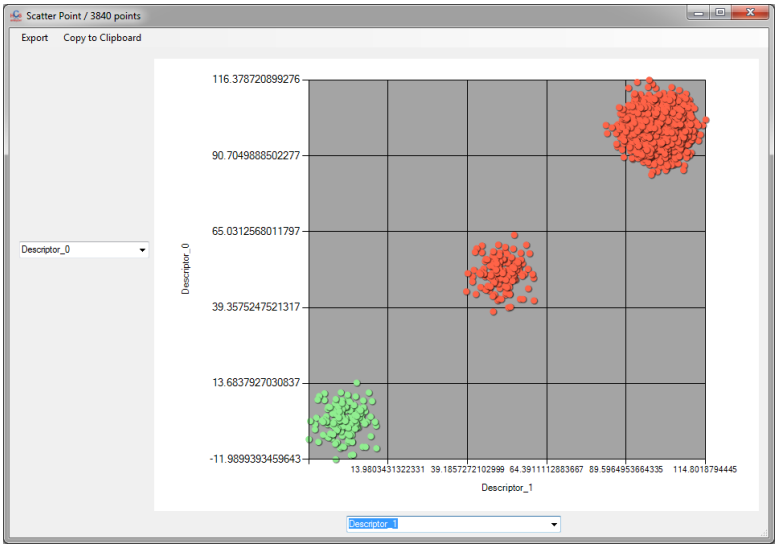
When done, information about the classification process is displayed, and in the case of C4.5 classification, the user has the opportunity to access each resulting tree. The trees are associated to each plate, and then the user has to selected the plate first and after either using the shortkey *Ctrl+T* or click on *Plate->Visualization->Classification Tree.*



By going back to the plate visualization, the user can assess the quality of the results.



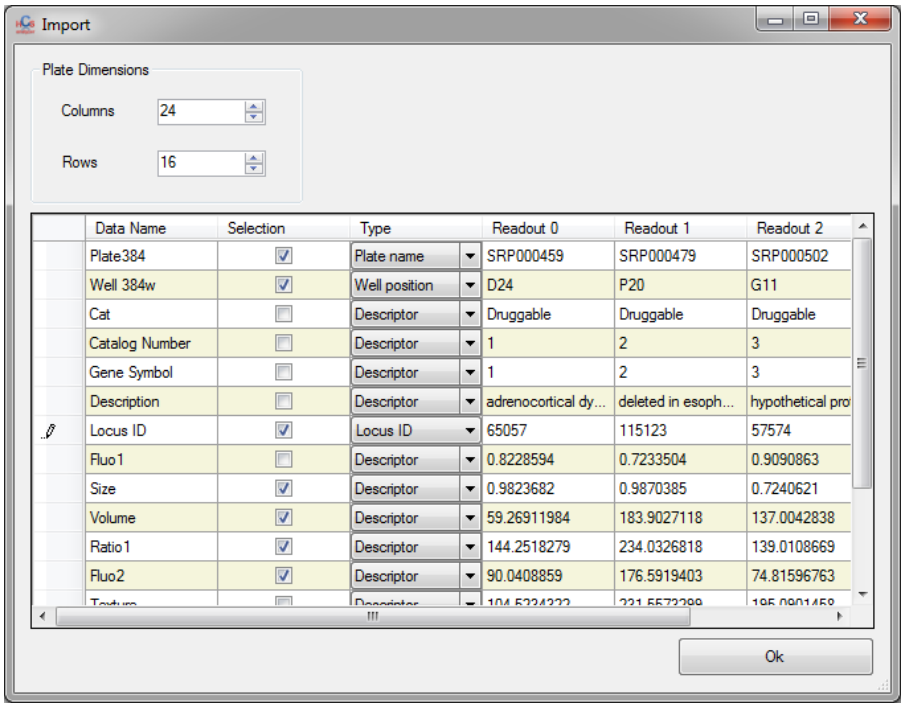
In this specific case, the classification is obvious, and the compounds have been classified in the same category as the phenotype 1.



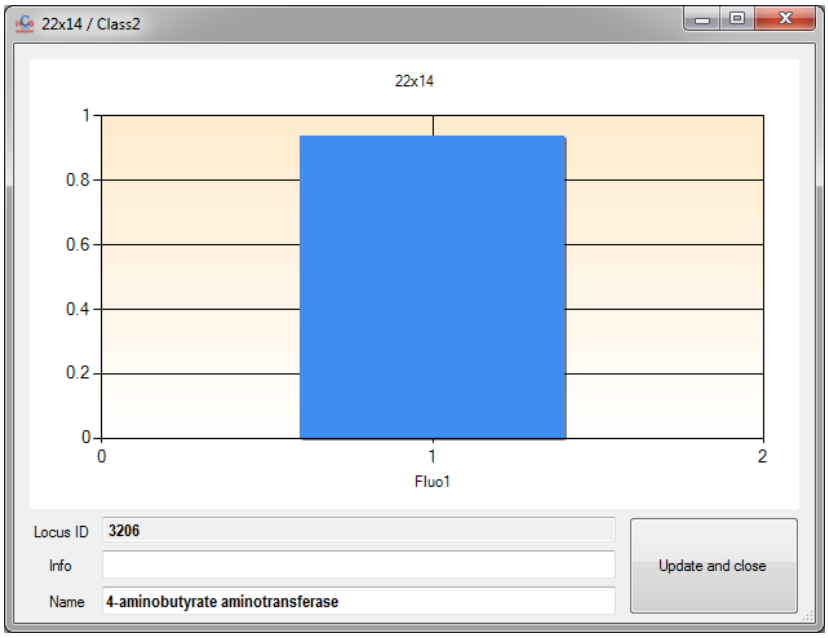
siRNA screening

In the context of siRNA screening, and if the user has locus ID associated to each well, we offer the possibility to connect the application to gather information about the genes involved as well as the pathways.

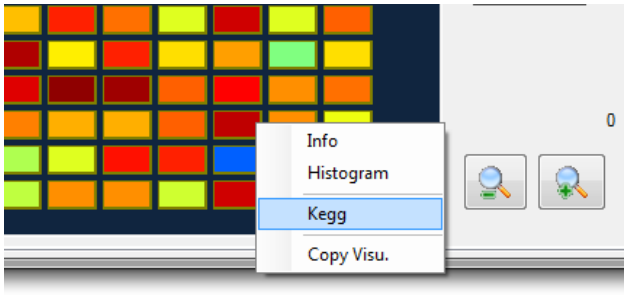
While loadint the screening the user has to specify the Locus ID such as follow:



In such context, the locus ID will be associated to each well. By right clicking and selecting *Info* from the contextual menu, the user can check the locus ID of a specific well.

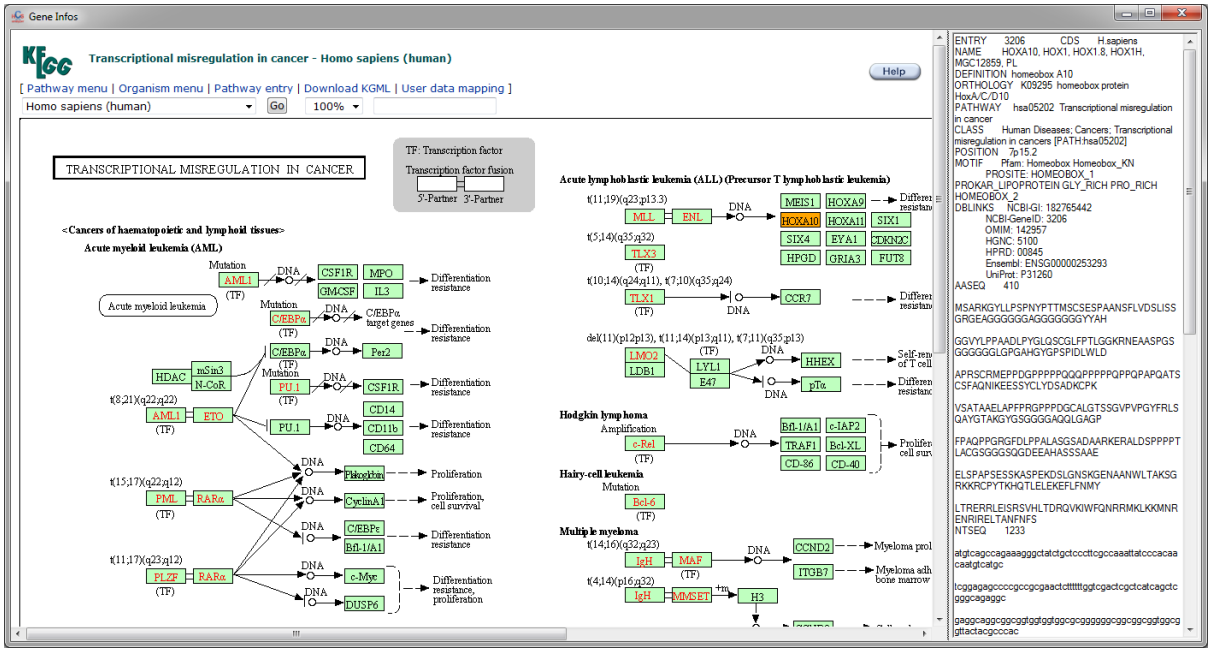


The user can also select Kegg to initiate a connection to the Kegg server.



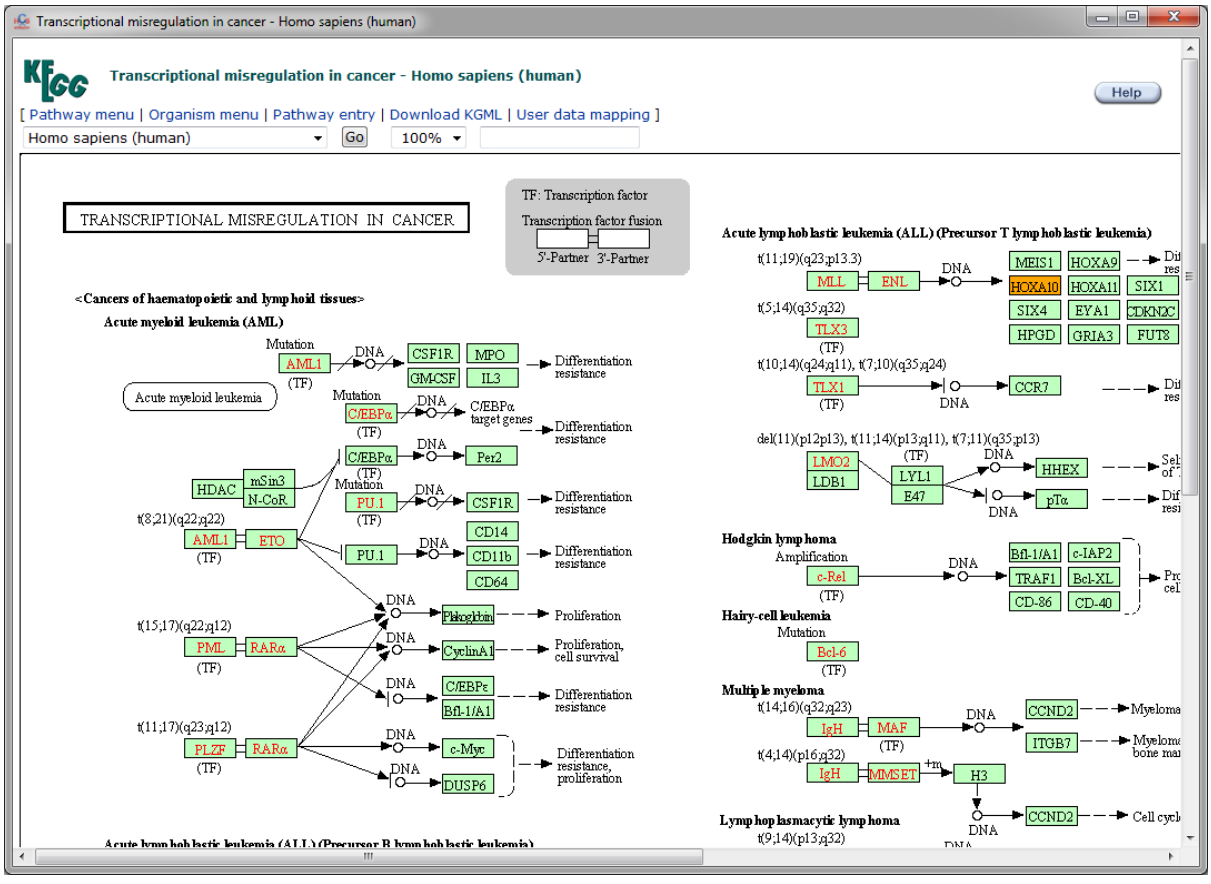
<u>Note:</u> for the option an internet connection is required.

If information about the gene are available, a window such as follow will be displayed.



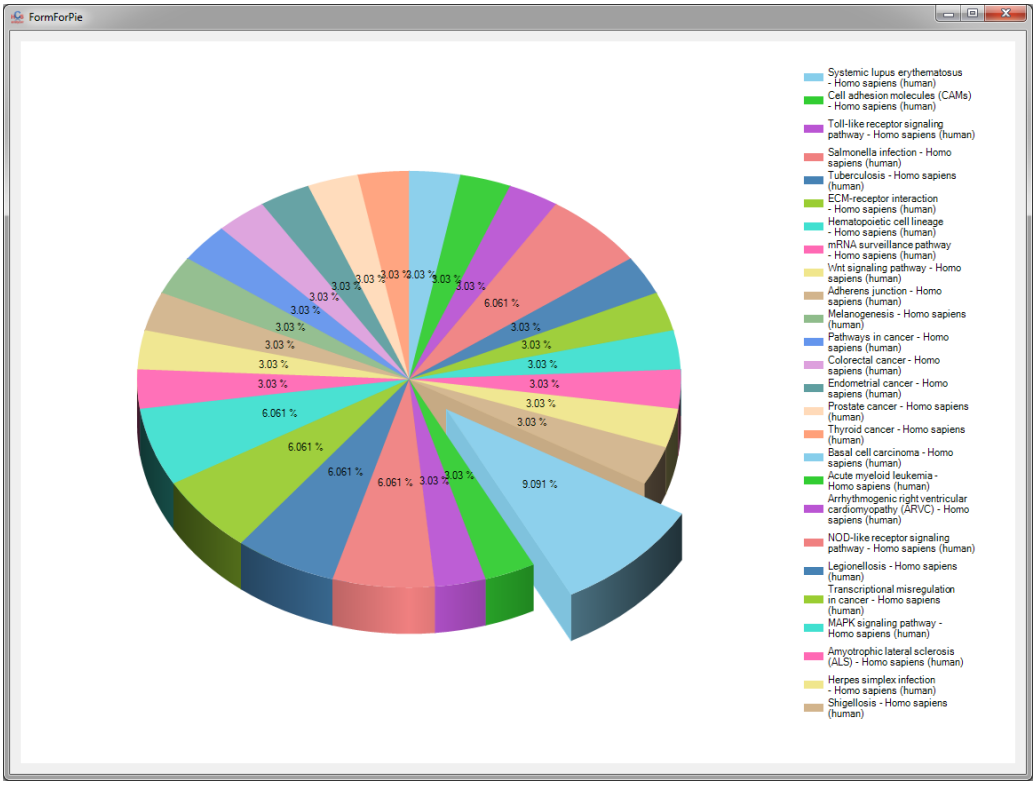
Where the gene of interest is highlighted in orange. In parallel, another window displays the pathway(s) related to this gene.

By double clicking on the pathway of interest, another window pops up, providing information about it.



In order to extract automatically information about the pathways involved in a certain phenotype, the user can select the following menu *Screen->Gene Analysis->Pathways analysis*.

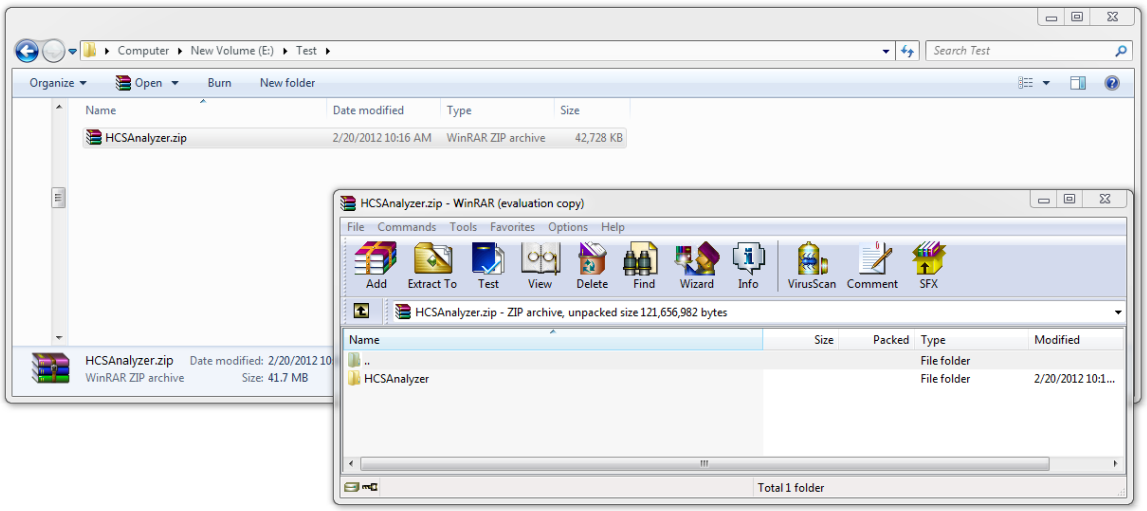
The application will then parse all the genes/well of the current selected class and gather the pathways related. At the end of the process, a pie chart displays the ratio of pathways involved and extracts the most recurrent.



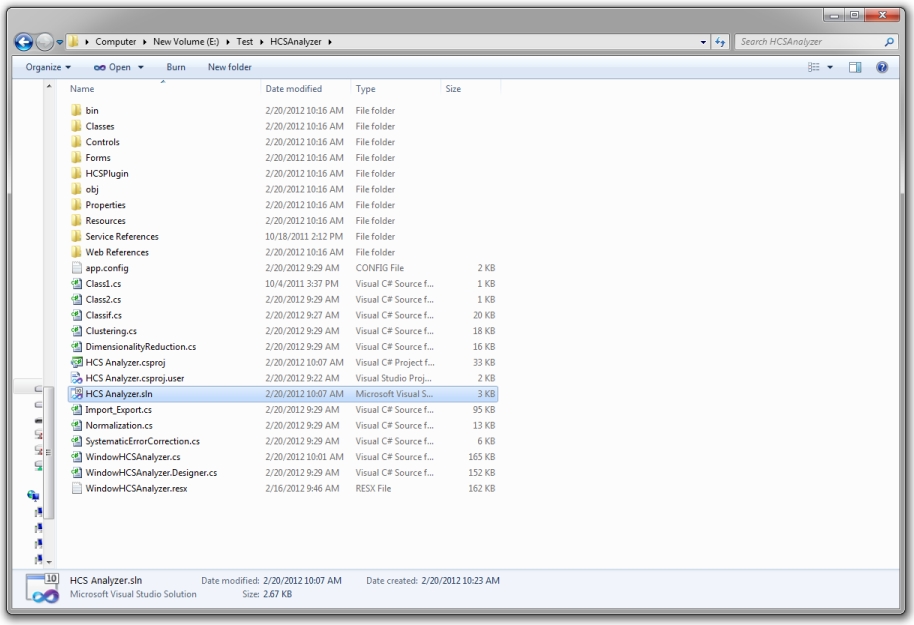
<u>Notes:</u> Due to the internet connections, this process is highly time consuming. This operation can be performed during the report process.

Development

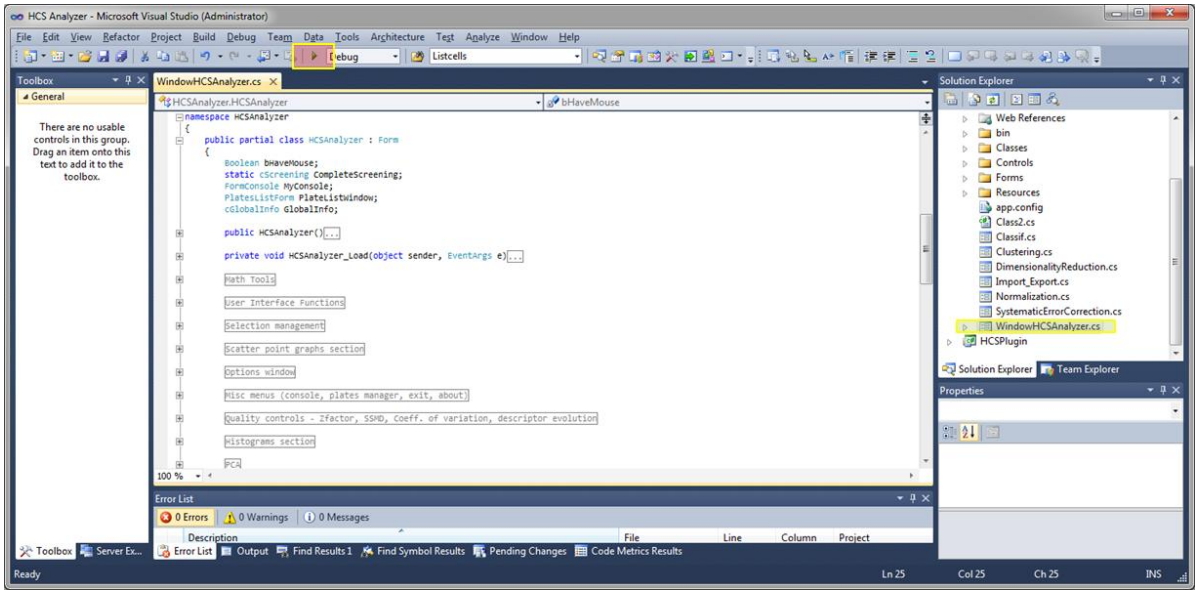
Start by extracting the complete solution from the “HCSAnalyzer.zip” file.



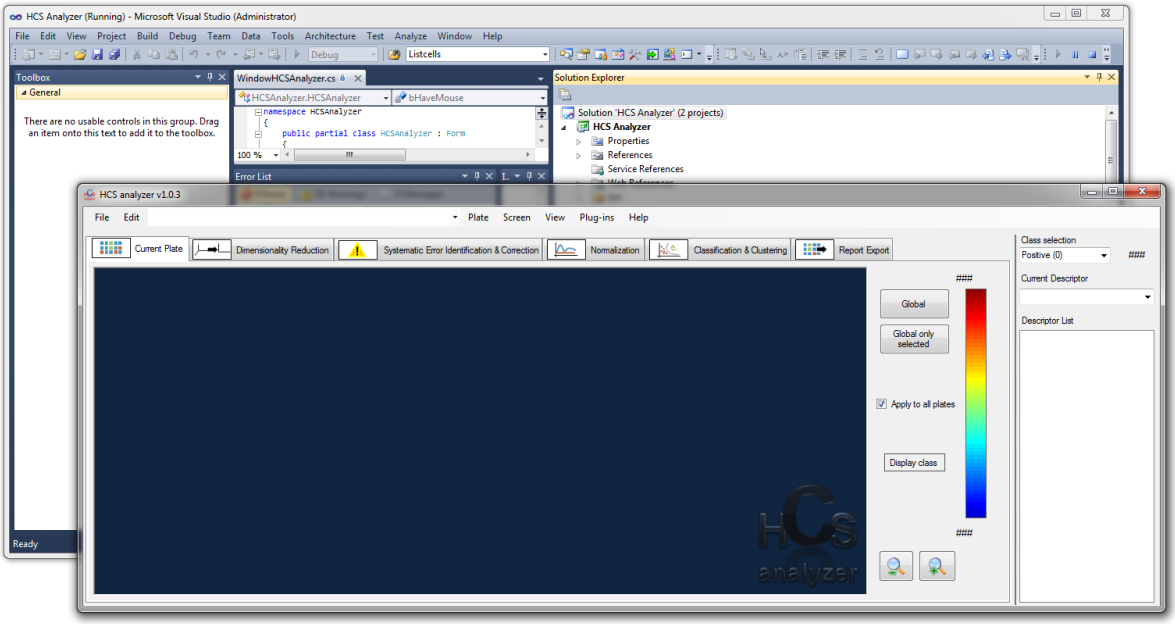
Double click on the solution file to launch the development environment tool.



One solution (“HCS Analyzer”) containing two projects (“HCS Analyzer” and “HCSPlugin”) has to be loaded.

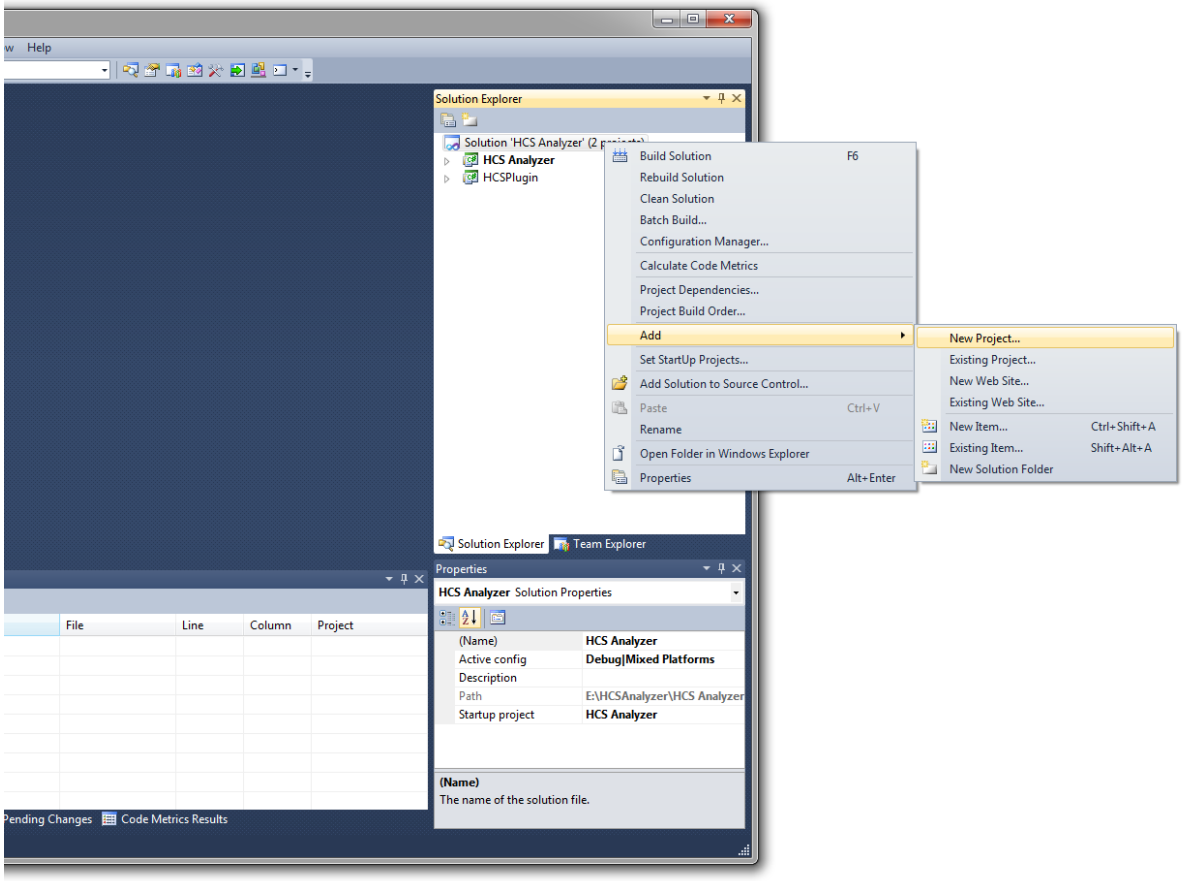


Press debug to launch the software.

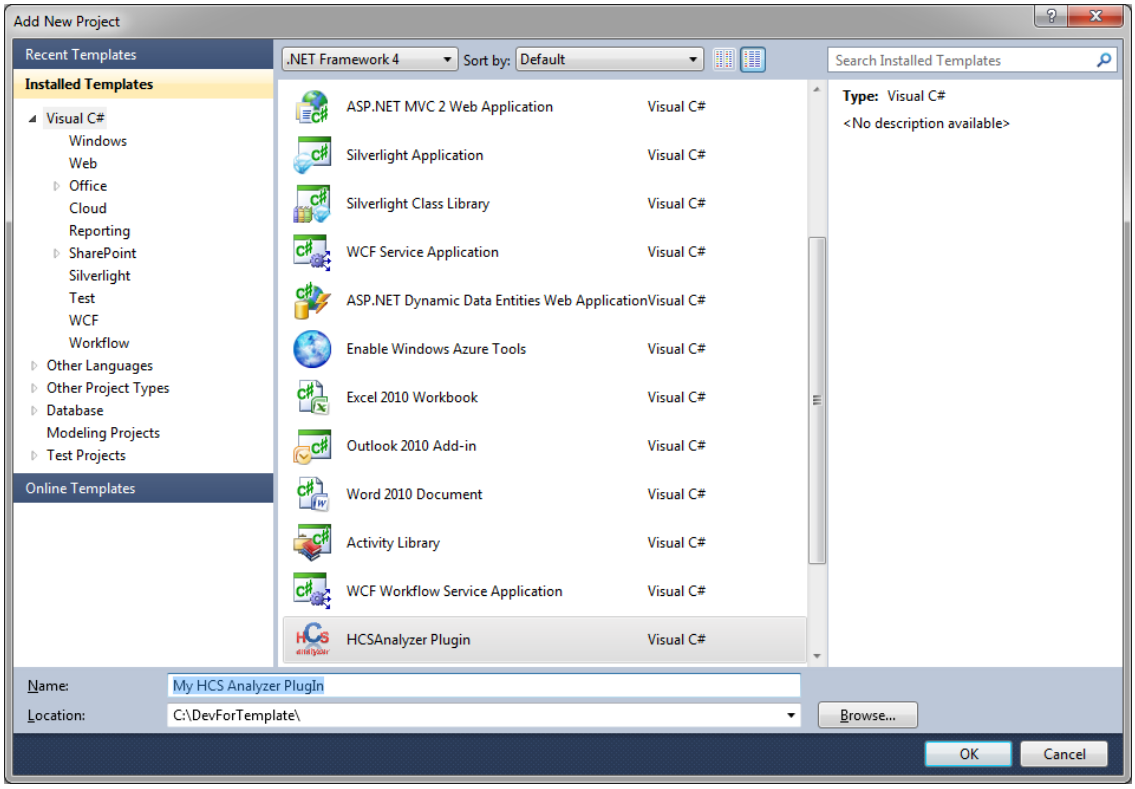


Plugin Development

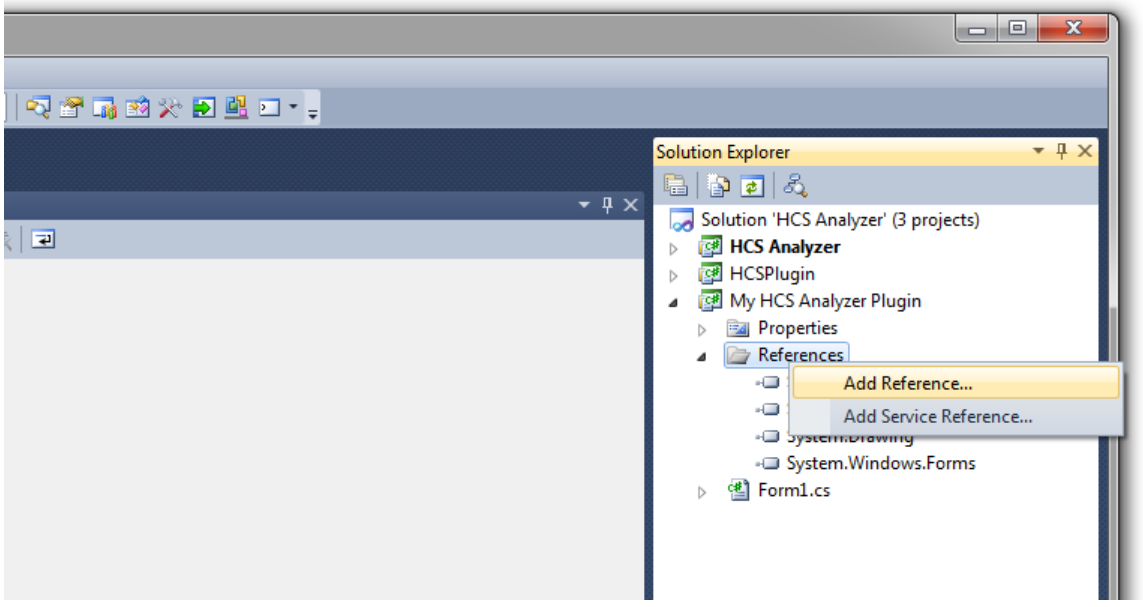
Open the HCSAnalyzer solution as describe above. Add a new project to the solution.



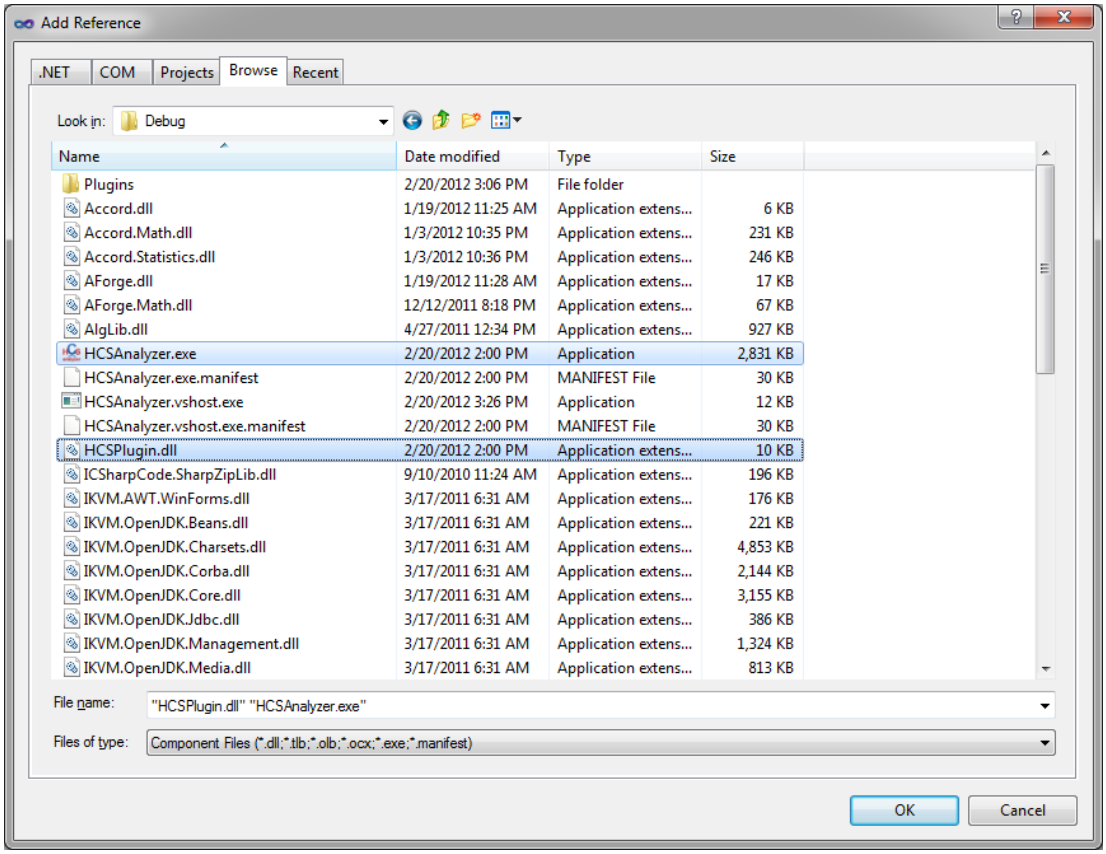
Choose the “HCSAnalyzer Plugin” template and specify the name:



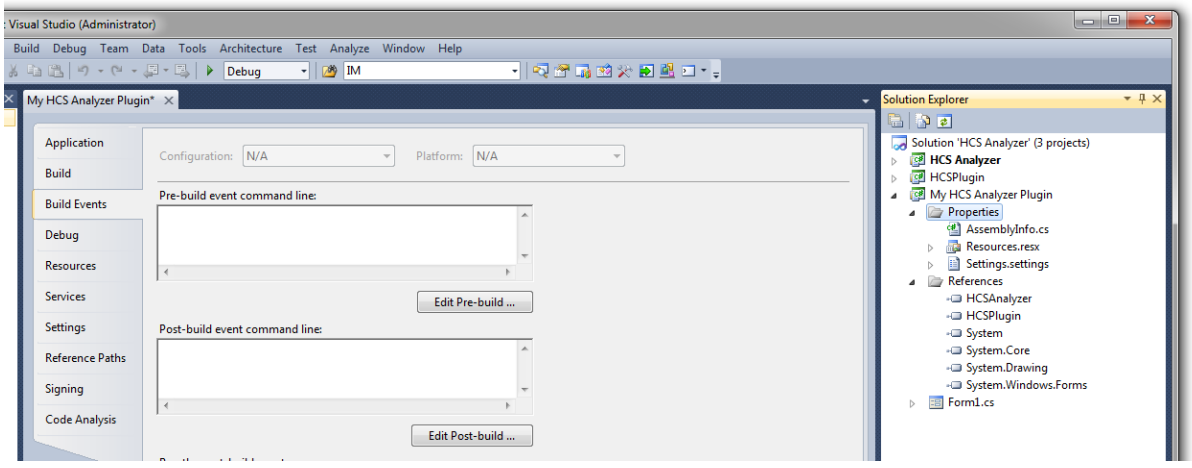
Note: To have access to this template, the “HCSAnalyzer Plugin.zip” has to be present in your Visual Studio template directory (e.g.:
“C:\Users\Myself\Documents\Visual Studio 2010\Templates\ProjectTemplates”).
At this stage you need to add two references to your plugin project :



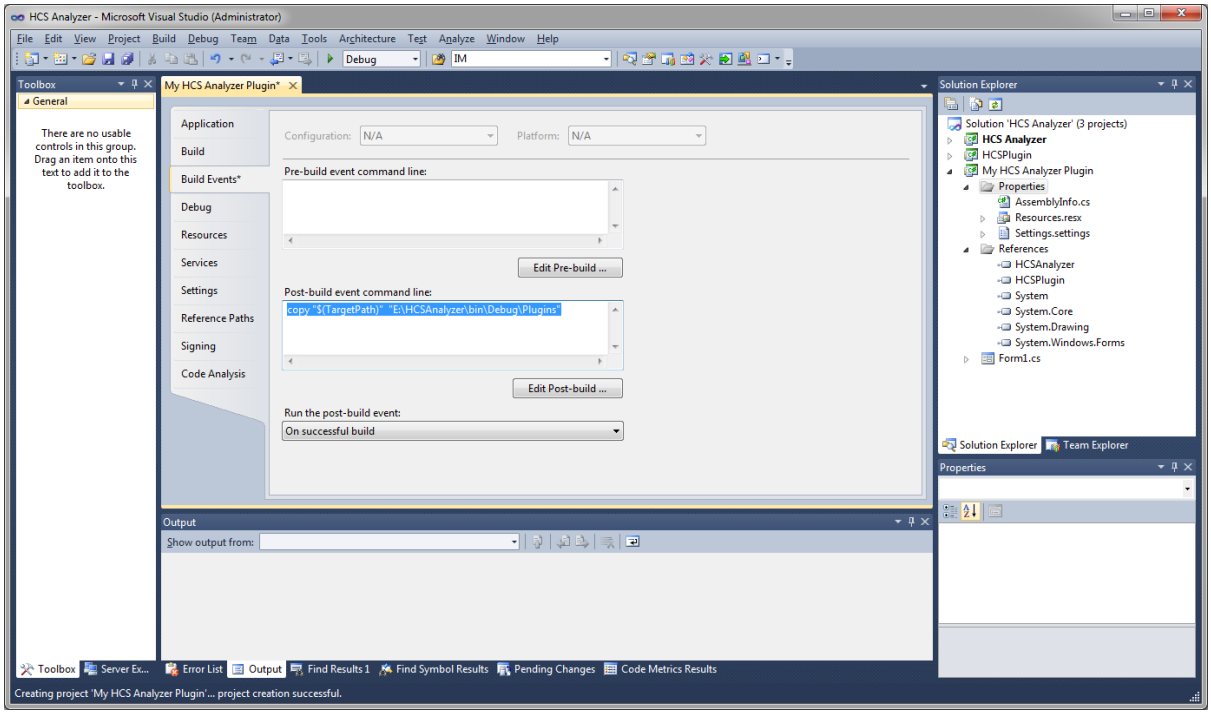
Click on the “Browse” tab, and go to your HCSAnalyzer development directory (probably something like “C:\HCSAnalyzer\bin\Debug”) and select “HCSAnalyzer.exe” and “HCSPlugin.dll”



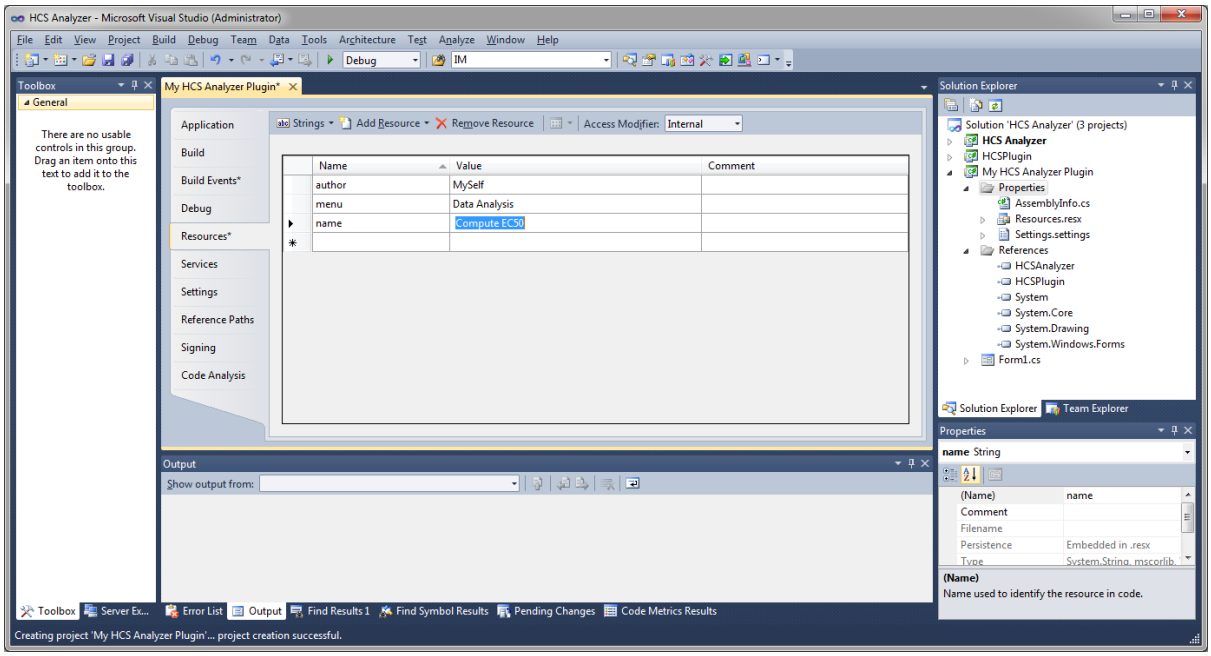
Once compiled, your plugin needs to be located in your HCSAnalyzer plugins directory. To do that automatically, double click on plugin Properties, and go the “Build Events” tab:



and specify your plugins directory in the “Post-build event command line:” (here “copy "$(TargetPath)" "E:\HCSAnalyzer\bin\Debug\Plugins" ”)



Finally give a name and a menu location to your plugin by clicking on the “Resources” tab and by filling out the three corresponding cells:



You can then develop your plugin and launch the HCSAnalyzer application. Your plugin will be available through the menu “Plug-ins”