

Supplementary information

GDA, a web-based tool for Genomics and Drugs integrated Analysis

Jimmy Caroli, Giovanni Sorrentino, Mattia Forcato, Giannino Del Sal, Silvio Bicciato

Step-by-step guide to a complete analysis in GDA

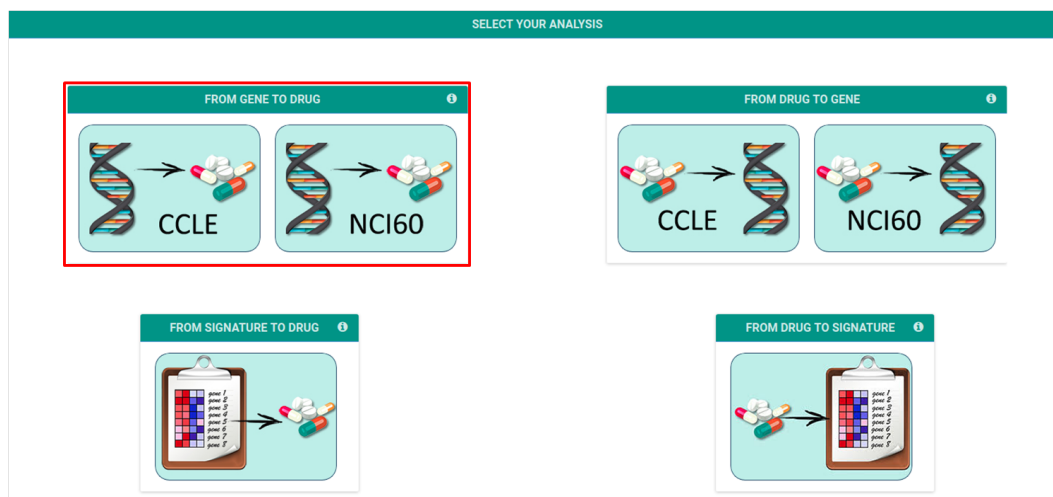
Introduction.....	1
The <i>from gene to drug</i> analysis module.....	1
The <i>from drug to gene</i> analysis module.....	9
The <i>from signature to drug</i> analysis module.....	11
The <i>from drug to signature</i> analysis module.....	14
<i>No Results</i> messages and negative cases with no results.....	17

Introduction

This tutorial illustrates how to perform the integrated analysis of genomic profiles and drug response data in cancer cell lines from the Genomics and Drugs integrated Analysis (GDA) portal. All queries used as examples in this tutorial have already been run in GDA to allow a faster retrieval of results.

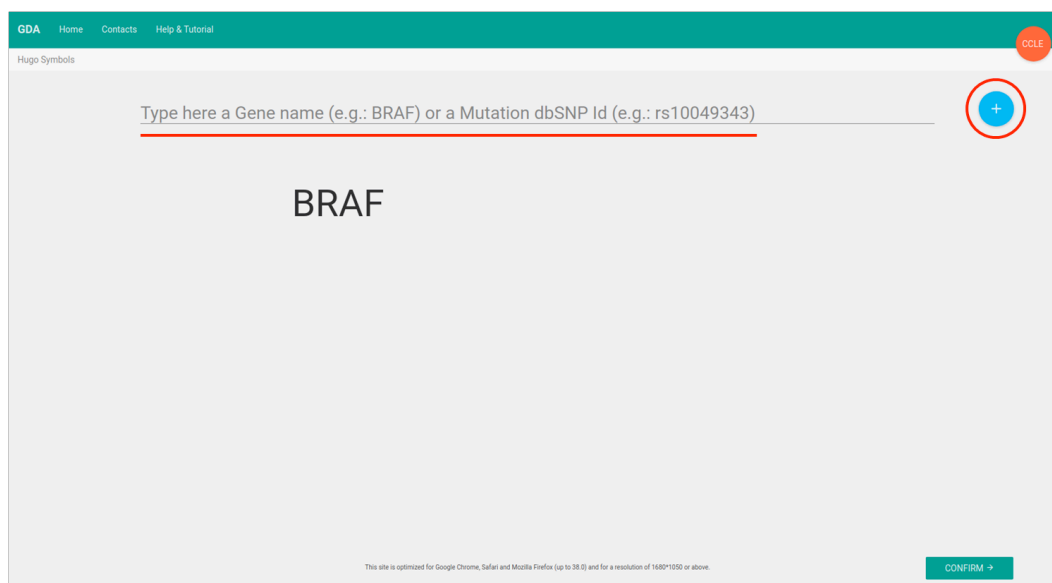
The *from gene to drug* analysis module

The *from gene to drug* query allows identifying drugs active in cancer cell lines bearing specific gene mutations. This query can be run selecting either **CCLE** or **NIC60** from the *from gene to drug* button of GDA home page (red box). The **CCLE** and **NIC60** buttons refer to two sets of genomic data: the **CCLE** database contains high-quality mutational data for a limited set of 1,651 oncogenes, while the **NIC60** database comprises mutations and SNPs for more than 15,000 genes.



Once the analysis has been selected, the user is redirected to the gene selection form, where it is possible to enter either the symbol of a gene of interest (or of a set of genes) or the dbSNP ID. In this tutorial, we will identify drugs that are active in cell lines bearing mutations in the **BRAF** gene using **CCLE** genomic data.

1. The *gene/SNP ID selection* form provides an input bar (red line) where the user can type the gene symbols or the dbSNP IDs of interest. Once selected the gene name or the dbSNP ID from the drop-down menu, click the "+" button (red circle) to confirm the gene selection. The gene symbol (or the dbSNP ID) will then be displayed at the center of the page. Multiple genes (or SNP IDs) can be added repeating the selection procedure.



After the gene selection, press the **Confirm** button, at the bottom right of the page, to proceed to the next input form.

- The tissue selection form provides the list of all tissues in which the mutational status of the selected gene is available. Different genes or gene combinations may yield different list of tissues for which the genomic information is available. For the purpose of this tutorial, we will select all available tissues (as for GDA default), but the list of selected tissues can be easily modified through the **checkbox panel**.

GDA Home Contacts Help & Tutorial

Genes / Tissues

UNSELECT ALL

Selected Genes

BRAF

Select Tissues from the list

- KIDNEY
- SOFT TISSUE
- LUNG
- BONE
- BREAST
- LARGE INTESTINE
- PROSTATE
- OVARY
- HAEMATOPOIETIC AND LYMPHOID TISSUE
- SKIN
- CENTRAL NERVOUS SYSTEM

← BACK

This site is optimized for Google Chrome, Safari and Mozilla Firefox (up to 38.0) and for a resolution of 1680*1050 or above.

CONFIRM →

Once selected the tissues, press the **Confirm** button, at the bottom right of the page, to proceed to the next input form.

- The variant classification form allows selecting one or more type of mutations for the investigated genes. In this tutorial, we will select **Missense Mutation** from the list of all possible mutations. As for the tissue list, the list of available mutations may change depending on the input genes and/or selected tissues. Different combination of genes and tissues will generate different lists of mutation variants.

GDA Home Contacts Help & Tutorial

Hugo Symbols / Tissues / Variant Classifications

SELECT ALL

Selected Genes

BRAF

Selected Tissues

KIDNEY
SOFT TISSUE
LUNG
BONE
BREAST
LARGE INTESTINE
PROSTATE
OVARY
HAEMATOPOIETIC AND LYMPHOID TISSUE
SKIN
CENTRAL NERVOUS SYSTEM

Select Variant Classifications from the list

- Intron
- Silent
- Missense Mutation

← BACK

This site is optimized for Google Chrome, Safari and Mozilla Firefox (up to 38.0) and for a resolution of 1680*1050 or above.

CONFIRM →

Once the type of variant has been selected, press the **Confirm** button, at the bottom right of the page, to proceed to the next form.

- After inputting genes, tissues, and variant type, GDA returns a *query recap* window summarizing the selected **genes and gene variants**, the **tissue types**, and of the **number of cell lines** available for the analysis given the selected combination of gene mutations and tissues.

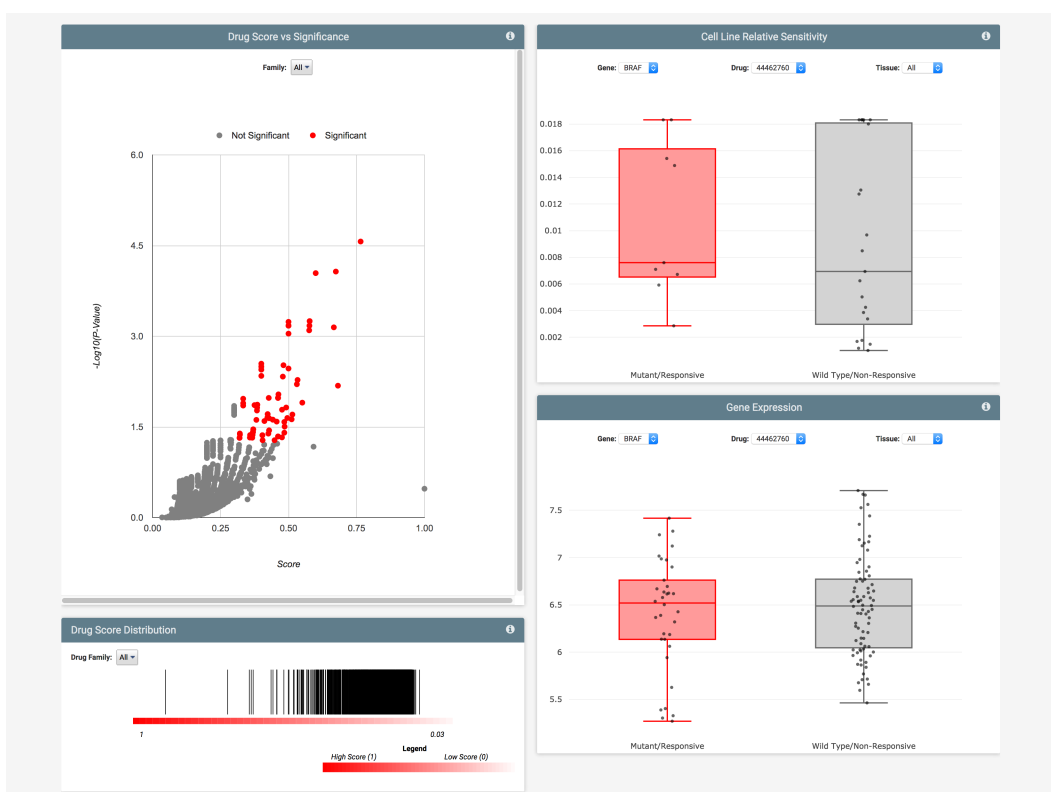
Once reviewed all input information, proceed to the analysis pressing the **Run analysis** button, at the bottom right of the page.

- The *result page* will be loaded once the analysis is completed. The result page contains two main tables: i) the left table “Drugs” lists all compounds that have been tested on the selected cancer cell lines bearing the selected type of mutation; ii) the right “Drug Families” table lists the drug families that are significantly enriched.

In the “Drug” table, drugs are identified in terms of compound ID (linking to Pubchem), name, drug family, mechanism of action (MoA), score, and statistical significance. Compounds are ordered based on the score (default) and, if the total number of compounds that have been tested on the selected tissues is larger than

500, only the first 500 are displayed. The table order can be easily changed acting on the column labels. Drugs that are significantly active on the selected combination of mutations and cell lines (i.e., with enrichment score ≥ 0.3 and p-value ≤ 0.05 ; see GDA help page and Taccioli et al., Oncotarget 2015 for details on the statistical analysis) are highlighted in cyan. Complete results for all drugs can be downloaded as an Excel spreadsheet using the “Download” button.

The “Drug Families” table lists the drug families that are significantly enriched, given the set of significantly active drugs (i.e. compounds highlighted in cyan in the “Drug” table). In this table, the cyan background highlights drug families comprising at least one compound with enrichment score ≥ 0.6 and p-value ≤ 0.05 .) Results are also visualized by a set of graphs. In this module, plots display (from the upper left to the lower right): i) the score as a function of the p-value for each drug and each drug family (with red dots highlighting significantly active drugs); ii) the distribution of the drug score for each drug in each drug family; iii) the distribution of relative sensitivities (RS) in cell lines bearing a given mutation and responsive to a given drug as compared to the distribution of relative sensitivities in wild type and non-responsive cell lines; and iv) the expression levels of each gene used in the query in cell lines with mutations in the gene and responsive to a given drug as compared to the expression levels of a gene used in the query in wild type and non-responsive cell lines.



Directly from the result page, it is possible to access the **Drugs clustering** or, once a drug is selected, the **Differential expression** analysis modules.

- In the **drugs clustering** module, the structures of all significantly active drugs are used to retrieve groups of compounds that, in addition to be active, also share a comparable chemical structure. To perform this analysis, click on the orange button at the bottom of the drug table.

GDA Home Contacts Help & Tutorial

Hugo Symbols / Tissues / Variant Classification / Search Terms Recap / Results

Results for HUGO SYMBOLS: BRAF Tissues: ALL selected Variant Classifications: 1 out of 3 selected

CCLE

QUERY RECAP:
Gene/SNPs Searched: BRAF Variant Classifications selected: Missense Mutation Tissues Selected: KIDNEY, SOFT TISSUE, LUNG, BONE, BREAST, LARGE INTESTINE, PROSTATE, OVARY, HAEMATOPOIETIC AND LYMPHOID TISSUE, SKIN, CENTRAL NERVOUS SYSTEM

Drugs

Show 10 entries

Search:

SELECT	#	CID ID	DRUG NAME	DRUG FAMILY	MoA	SCORE	P-VALUE
<input type="radio"/>	1	44462760	Dabrafenib	Imatinib analog		0.765	0
<input type="radio"/>	2	2022	Acyclovir-NSC678518	DNA Polymerase Inhibitors		0.681	0.006
<input type="radio"/>	3	761586	Act1fx6f	Others		0.674	0
<input type="radio"/>	4	42611257	Vemurafenib	Imatinib analog		0.6	0
<input type="radio"/>	5	756456	St50162357	Others		0.577	0
<input type="radio"/>	6	756464	Mis000716013	Others		0.577	0
<input type="radio"/>	7	388024	NSC682449	Others		0.576	0
<input type="radio"/>	8	5477775	NSC354462	Others	tyrosine kinase inhibitor	0.576	0
<input type="radio"/>	9	389638	NSC685981	Others		0.55	0.012
<input type="radio"/>	10	10127622	Selumetinib	Imatinib analog	YK and Serine/threonine kinase inhibitor	0.533	0.005

Showing 1 to 10 of 487 entries

First Previous Next Last

[DOWNLOAD](#)
[DRUGS CLUSTERING](#)
[DIFFERENTIAL EXPRESSION](#)

Drug Families

DRUG FAMILY	# DRUGS		P-VALUE		
	TOTAL	SCORE > 0.3	SCORE > 0.6	DRUGS WITH SCORE > 0.3	DRUGS WITH SCORE > 0.6
DNA Polymerase Inhibitors	2	1	1	0.068	0.001
Imatinib analog	13	3	1	0.009	0.009
Vinca alkaloids	3	2	0	0.003	1
Sulfonate	35	2	0	0.349	1
Quinone/Quinoline	343	10	0	0.772	1
Antimetabolite	9	1	0	0.275	1
Neuropsychiatric agent	5	1	0	0.163	1

The output page of this analysis provides an interactive clustering tree of all the significant drugs from the previous analysis.

Each node of the tree can be selected resulting in the list of drugs reported in the table at the bottom of the clustering tree. The table is interactive and refreshes with the selection performed on the clustering tree, highlighting only drugs that have been selected through node clicking in the tree. The table does not refresh only in the case a terminal node (i.e., a node that does not split) is selected. The clustering tree can also be saved in a PDF document.

Once a node is selected, it is possible to perform the *Maximum Common Structure Analysis* to retrieve a common scaffold shared by all the compounds belonging to a group. The Maximum Common Structure analysis can be performed only if two or more drugs have been selected. In the case the user selects a single drug (e.g. a drug in a terminal node), the structure of the drug will be displayed but the analysis will not be performed. Although the Maximum Common Structure analysis can be performed on any number of clustered drugs, users should be aware that a large number of selected drugs will yield a very simple maximum common structure as output. Thus, we suggest performing this kind of analysis on a limited set of similar drugs, to highlight a potential common scaffold that can be exploited for drug discovery or for comparison against other known drugs not included in the NCI-60 panel.

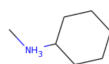
GDA Home Contacts Help & Tutorial

Results / Drugs Clustering

Drugs

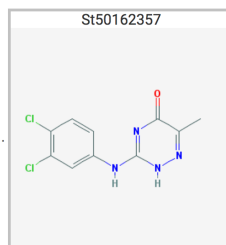
MCS	COMPOUND ID	NAME	DRUG FAMILY	MoA	SCORE	P-VALUE
<input checked="" type="checkbox"/>	395559	Chemb16116	Quinone/Quinoline		0.461	0.01
<input checked="" type="checkbox"/>	24203851	NSC729792	Others		0.456	0.025
<input checked="" type="checkbox"/>	5388227	NSC633942	Others		0.428	0.022
<input checked="" type="checkbox"/>	69462	Resorufin	Others		0.427	0.01
<input checked="" type="checkbox"/>	245772	2,4-Dimethylamino-Phenyl-Hydroxy-Methyl-Naphthalene-1,4-Dione	Others		0.411	0.025
<input checked="" type="checkbox"/>	760289	Act1flu4	Quinone/Quinoline		0.37	0.034
<input checked="" type="checkbox"/>	24202960	Chemb1516382	Others		0.369	0.038
<input checked="" type="checkbox"/>	324952	NSC292147	Others		0.368	0.042
<input checked="" type="checkbox"/>	81146	S-Benzoyloxy-1h-Indole-2-Carboxylic-Acid-NSC49096	Others		0.32	0.047
<input checked="" type="checkbox"/>	4713	Pd-98059	Others	serine/threonine kinase inhibitor	0.3	0.018
<input checked="" type="checkbox"/>	248466	NSC65537	Others		0.3	0.016
<input checked="" type="checkbox"/>	385176	Chemb1423845	Others		0.3	0.018
<input checked="" type="checkbox"/>	406532	5-5,4-Methoxy-2-Nitrophenyl-Furan-2-Yl-3,4-Methoxyphenyl-4,5-Dihydro-1h-Pyrazole	Others		0.3	0.02
<input checked="" type="checkbox"/>	16680184	2,4-Thiazinium-Chloride	Quinone/Quinoline		0.3	0.016

Maximum Common Structure Result



The Maximum Common Structure shares the SMILES sequence cNc1ccccc1

The compound more similar to the Maximum Common Structure is [St50162357](#)



b) The **Differential expression** module is activated selecting a drug using the check circle in the SELECT column. Once selected a drug (here **Vemurafenib**), press the **Differential expression** button to be directed to the comparison page.

The screenshot shows the GDA interface with the 'Drugs' table and 'Drug Families' summary table.

Drugs Table:

SELECT	#	CID ID	DRUG NAME	DRUG FAMILY	MoA	SCORE	P-VALUE
<input type="radio"/>	1	44462760	Dabrafenib	Imatinib analog		0.765	0
<input type="radio"/>	2	2022	Acyclovir-NSC678518	DNA Polymerase Inhibitors		0.681	0.006
<input type="radio"/>	3	761586	Actixifexf	Others		0.674	0
<input checked="" type="radio"/>	4	42611257	Vemurafenib	Imatinib analog		0.6	0
<input type="radio"/>	5	756456	SIS0162357	Others		0.577	0
<input type="radio"/>	6	756454	Mis00716013	Others		0.577	0
<input type="radio"/>	7	388024	NSC682449	Others		0.576	0
<input type="radio"/>	8	547775	NSC354402	Others	tyrosine kinase inhibitor	0.576	0
<input type="radio"/>	9	389638	NSC685981	Others		0.55	0.012
<input type="radio"/>	10	10127622	Selumetinib	Imatinib analog	YK and Serine/threonine kinase inhibitor	0.533	0.005

Drug Families Summary Table:

DRUG FAMILY	# DRUGS			P-VALUE	
	TOTAL	SCORE > 0.3	SCORE > 0.6	DRUGS WITH SCORE > 0.3	DRUGS WITH SCORE > 0.6
DNA Polymerase Inhibitors	2	1	1	0.068	0.001
Imatinib analog	13	3	1	0.009	0.009
Vinca alkaloids	3	2	0	0.003	1
Sulfonate	35	2	0	0.349	1
Quinone/Quinoline	343	10	0	0.772	1
Antimetabolite	9	1	0	0.275	1
Neuropsychiatric agent	5	1	0	0.163	1

The comparison page allows selecting 3 different comparisons of the gene expression levels in Group A, i.e., in the cancer cell lines responsive to the selected drug (Vemurafenib) and bearing the selected genomic background (missense mutation on BRAF), against the gene expression levels of cell lines in Group B. Group B can be composed of cell lines i) non-responsive to the selected drug and lacking the mutation in the selected gene; ii) responsive to the selected drug and lacking the mutation in the selected gene; and iii) non-responsive to the selected drug even if bearing the mutation in the selected gene. Here, we compare cell lines **responsive** to Vemurafenib and **with** the missense mutations in BRAF (Group A) against cell lines **non-responsive** to Vemurafenib and **with** mutated BRAF e.g., searching for genes which likely interfere with the anticancer action of Vemurafenib. Thus, we select the **third option** for Group B and this unlocks the tissue selection in the left panel. It is worth noting that **not all tissues are selectable**, even though a comparison selection has been made. The tissue list contains **all** the potential comparable tissues for the 3 different comparisons, but each comparison may have a different number of available tissues, thus limiting a potential comparison. Here, we will select both **Melanoma** and **Breast** and press **confirm** to proceed.

The screenshot shows the 'Comparison Selection' page in GDA.

Group A: Cell lines **RESPONSIVE** to drug: 42611257 **with** Missense Mutation mutations in BRAF.

Group B:

- Cell lines **NON-RESPONSIVE** to drug: 42611257 **without** Missense Mutation mutations in BRAF.
- Cell lines **RESPONSIVE** to drug: 42611257 **without** Missense Mutation mutations in BRAF.
- Cell lines **NON-RESPONSIVE** to drug: 42611257 **with** Missense Mutation mutations in BRAF.

Perform differential gene expression analysis on:

Please select **first** the comparison analysis to perform. Tissues in **grey** are not available for the selected comparison analysis.

- Colon
- Melanoma
- Breast

Buttons: **← BACK** and **CONFIRM →**

On the selection process is completed, GDA returns a recap page, listing the cell lines included in each group, the number of samples for each cell line, and summarizing if the number of samples in each group is equal or higher than 3, i.e., the minimum number of samples requested for this type of analysis. It is now possible to run two different types of differential gene expression analysis, i.e. SAM analysis or T-Test analysis, based on the Significant Analysis of Microarray procedure and on the t-test, respectively. Click on **"Run SAM Analysis"** button to proceed.

The screenshot shows the GDA interface with a green header containing 'GDA Home Contacts Help & Tutorial' and a 'CCLE' logo. Below the header, there are navigation links: 'Results Comparison Selection / GE Terms Recap'. The main content area is divided into two panels:

- Cell lines RESPONSIVE to drug: 42611257 with mutated BRAF in Breast, Melanoma:**
 - LOXIMVI (4 samples)
 - MALME-3M (4 samples)
 - MDA-MB-231/ATCC (4 samples)
 - SK-MEL-28 (2 samples)
 - SK-MEL-5 (4 samples)
 - UACC-257 (4 samples)
 - UACC-62 (4 samples)
- Cell lines NON-RESPONSIVE to drug: 42611257 with mutated BRAF in Breast, Melanoma:**
 - MDA-MB-435 (4 samples)
 - ZR-75-30 (1 samples)

Below these panels is a 'Samples Number Check' box with the text: 'Each group has at least 3 samples. Thus, differential gene expression analysis can be performed.' At the bottom, there are two buttons: 'RUN SAM ANALYSIS' and 'RUN TTEST ANALYSIS'.

The output page provides the list of genes upregulated (left panel) and downregulated (right panel) in cancer cell lines of Group A. These lists can be directly functionally annotated using Enrichr, compared to results from L1000, used to generate gene signatures for further analysis in the *from signature to drug* module of GDA (see description of the *from signature to drug* module for details) or downloaded as an Excel file. Here, for the functional annotation in Enrichr, click Test Results on Enrichr and select $FDR < 0.05$ and $Fold\ Change < -2$ (Group B). We focus on the subset of 316 genes (compared to the total 1,043 differentially expressed genes) that are upregulated in cancer cell lines non-responsive to Vemurafenib (i.e., downregulated in Group A at $FDR > 0.05$ and $FC < -2$). In the Enrichr output page, select "Pathways" in the top layer and then select "KEGG 2016". Using these filters, the first result is PPAR signaling pathway, suggesting that cell lines bearing BRAF mutations but showing non-response towards the action of Vemurafenib presents an activated PPAR signaling pathway, which, from literature, has been linked to the activation of STAT3. This result suggests that a strategy to overcome defective response to Vemurafenib in these cell lines could be adding an additional treatment with STAT3 directed inhibitors, as demonstrated by Wang et al. Biol Cell 2009.

The screenshot shows the GDA interface with a green header containing 'GDA Home Contacts Help & Tutorial' and a 'CCLE' logo. Below the header, there are navigation links: 'Home / Results / Comparison Selection / GE Terms Recap / GE Results'. The main content area is titled 'Results for: 42611257' and includes a 'Query Recap' section:

- Group A:** cell lines RESPONSIVE to drug: 42611257 with Missense Mutation mutations in BRAF.
- Group B:** cell lines NON-RESPONSIVE to drug: 42611257 with Missense Mutation mutations in BRAF.

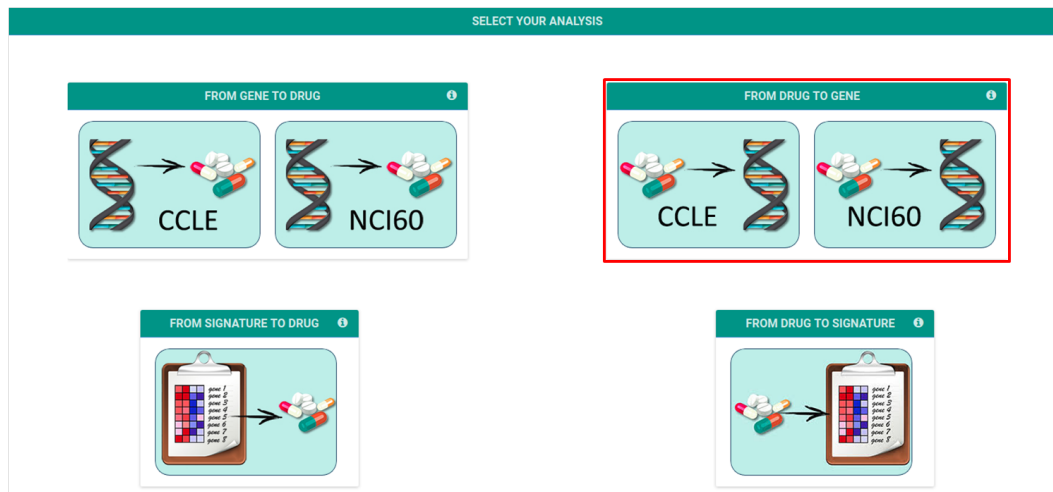
Below the recap are two panels:

- Genes Up-Regulated in Group A:** A table with columns: #, Gene Name, Fold Change, False Discovery Rate. It shows 10 entries, with the first being LIXN (Fold Change: 36.16, FDR: 0).
- Genes Down-Regulated in Group A (Up in Group B):** A table with columns: #, Gene Name, Fold Change, False Discovery Rate. It shows 10 entries, with the first being AZGP1 (Fold Change: -22.6, FDR: 0).

At the bottom, there are three buttons: 'Test Results on Enrichr', 'Test Results on L1000', and 'Test Results on GDA'. A 'HOME' button is on the left and a 'DOWNLOAD' button is on the right.

The *from drug to gene* analysis module

The *from drug to gene* query allows identifying gene mutations characterizing cancer cell lines that are responsive to a selected compound. This query can be run selecting either **CCLE** or **NCI60** from the *from drug to gene* button of GDA home page (red box). The **CCLE** and **NCI60** buttons refers to two different sets of genomic data: the **CCLE** database contains high-quality mutational data for a limited set of 1,651 oncogenes, while the **NCI60** database comprises mutation and SNPs for more than 15,000 genes. In this tutorial we will select **CCLE** genomic data.

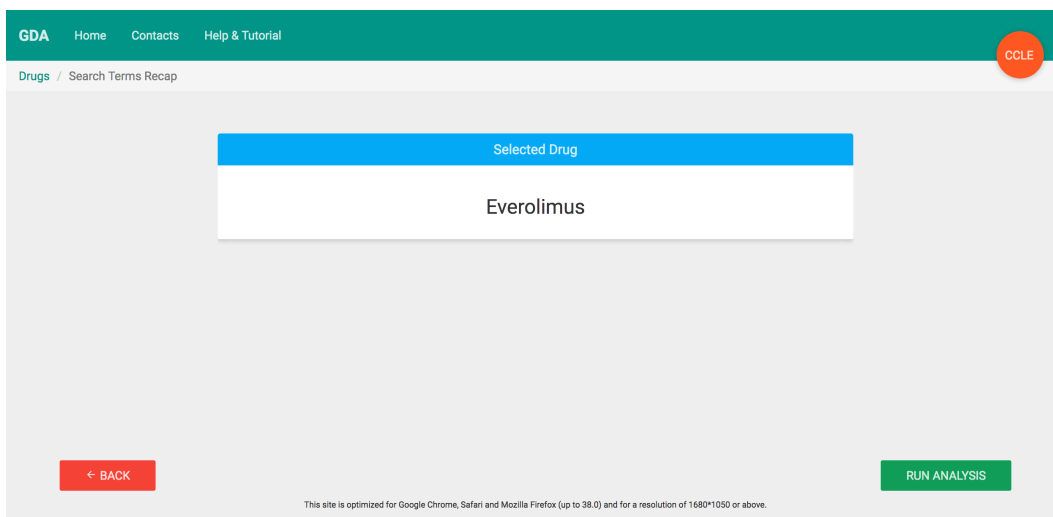


Once the analysis has been selected, the user will be redirected to the drug selection form, where it is possible to enter the name of a **single drug** of interest (i.e., **multiple drug names are not allowed**). In this tutorial, we will identify genomic mutations that are present in cell lines responsive to **Everolimus**, a derivative of Sirolimus and Temsirolimus, two inhibitors of the mammalian target of Rapamycin (mTOR).

1. The *drug name/NSC ID selection* form provides an input bar where the user can type the name or the NSC ID of the drug of interest. Once the drug name is typed in the drop-down menu, click the “+” button to confirm the drug selection, which will appear at the center of the page. Multiple drug names (or NSC ID) are not allowed.

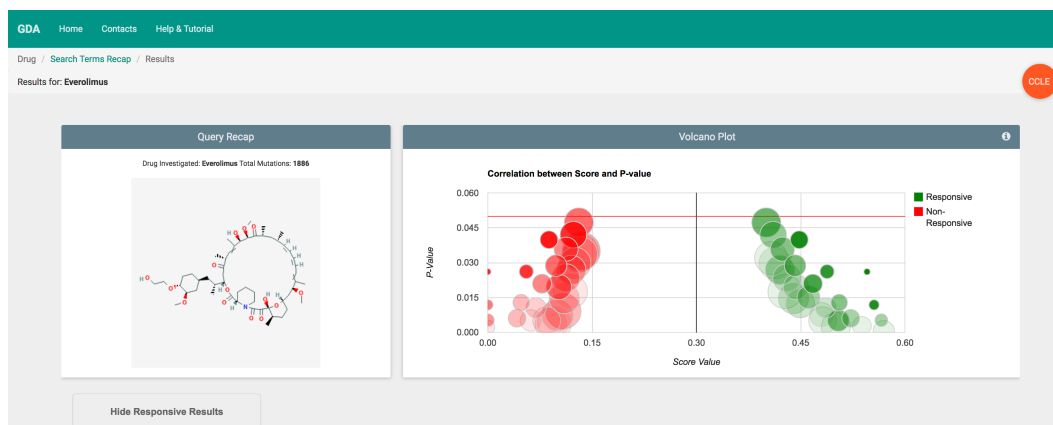
After the drug selection, press the **Confirm** button, at the bottom right of the page, to proceed to the next form.

2. After inputting the drug, GDA returns a *query recap* window containing **the selected input** for the subsequent analysis.

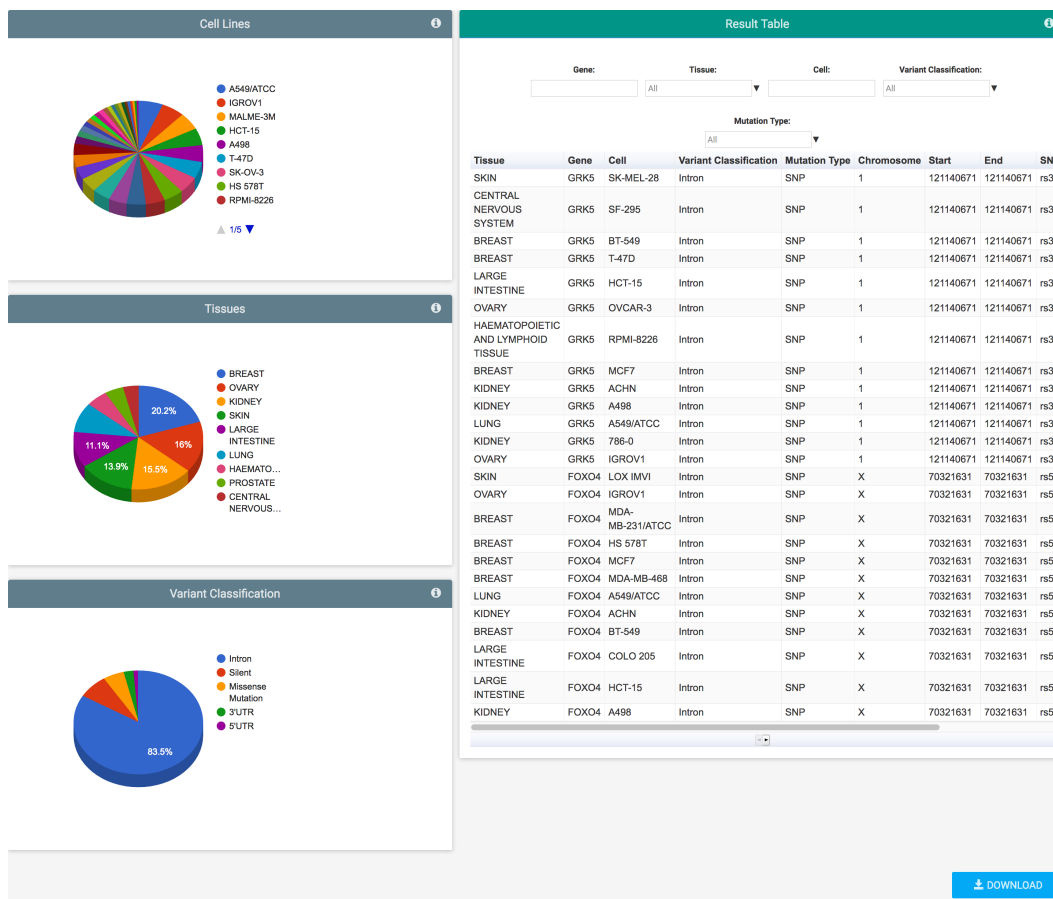


Once reviewed all input information, proceed to the analysis pressing the **run analysis** button, at the bottom right of the page.

- Once the analysis is complete, the *result page* is loaded. The result page reports the chemical structure of the selected compound (if available and linking to PubChem) with the number of mutations found in responsive and non-responsive cell lines and an interactive **volcano plot** showing the score and p-value of gene mutations and SNPs, present in cancer cell lines that are responsive (green bubbles) or non-responsive (red bubbles) to the selected compound. Clicking on a bubble will highlight the gene and SNP name, the score and p-value, and the number of cell lines bearing the specific mutation and being either responsive or non-responsive to the given drug.

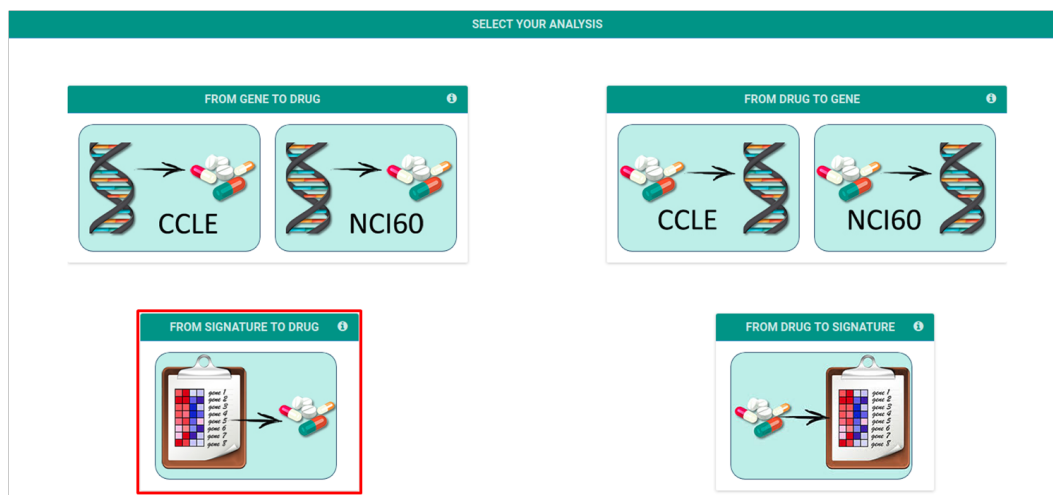


Results are also visualized by pie-charts and a table (default for responsive cell lines). The pie charts report: i) the distribution of responsive (or non-responsive) cell lines bearing mutations; ii) the distribution of tissues for responsive (or non-responsive) cell lines bearing mutations; and iii) the distribution of mutations types in cell lines responsive (or non-responsive) to the selected drug. The table lists tissue, gene symbol, cell line, variant classification, mutation type, chromosome coordinates, SNP ID, score and p-value of each mutation in cell lines responsive (or non-responsive) to the selected drug. Tables of results for responsive and non-responsive cell lines can be downloaded as Excel spreadsheets.



The from signature to drug analysis module

The *from signature to drug* query (red box) allows identifying drugs active in cancer cell lines bearing the activation of a specific gene signature, i.e. a pre-determined gene set. This query is run on the integrated gene expression profiles of **CCLE** and **NIC60** obtained using Affymetrix human gene expression arrays and the Brainarray custom chip definition files.



1. In the *input data* page, the user can type the gene signature in terms of official gene symbols or upload it as a text tab-delimited file. After imputing the gene signature, it is possible to specify if the signature of interest has to be up- or down-regulated and select the tissues of interest. By default, GDA searches for **upregulated** signatures. In this tutorial, we will upload a preset gene signature using the **Test Set** button and select **Upregulated** in all tissues. The test set contains genes upregulated upon knockdown of PTEN (Vivanco et al., Cancer Cell, 2007; PTEN_DN.V1_UP gene set of the Molecular

Signatures Database; <http://software.broadinstitute.org/gsea/msigdb/index.jsp>). Once all parameters have been selected, press the **Confirm** button.

2. GDA will then show a recap page, indicating those genes of the signature that are eventually not present in the gene expression matrix. The recap page will additionally show how many genes will be analyzed from the total input, the selected tissues, and the selected signature profile. Press **Run Analysis** to perform the analysis.

The output page of this analysis is identical of the result page of the from gene to drugs module and contains two main tables: i) the left table “Drugs” lists all compounds that have been tested on the selected cancer cell lines bearing the upregulation of the input gene signature; ii) the right “Drug Families” table lists the drug families that are significantly enriched. In the “Drug” table, drugs are identified in terms of compound ID (linking to Pubchem), name, drug family, mechanism of action (MoA), score, and statistical significance. Compounds are ordered based on the score (default) and, if the total number of compounds that have been tested on the selected tissues is larger than 500, only the first 500 are displayed. The table order can be easily changed acting on the column labels. Drugs that are significantly active on cell lines the up-regulation of the gene signature (i.e., with enrichment score \geq 0.3 and p-value \leq 0.05; see GDA help page and Taccioli

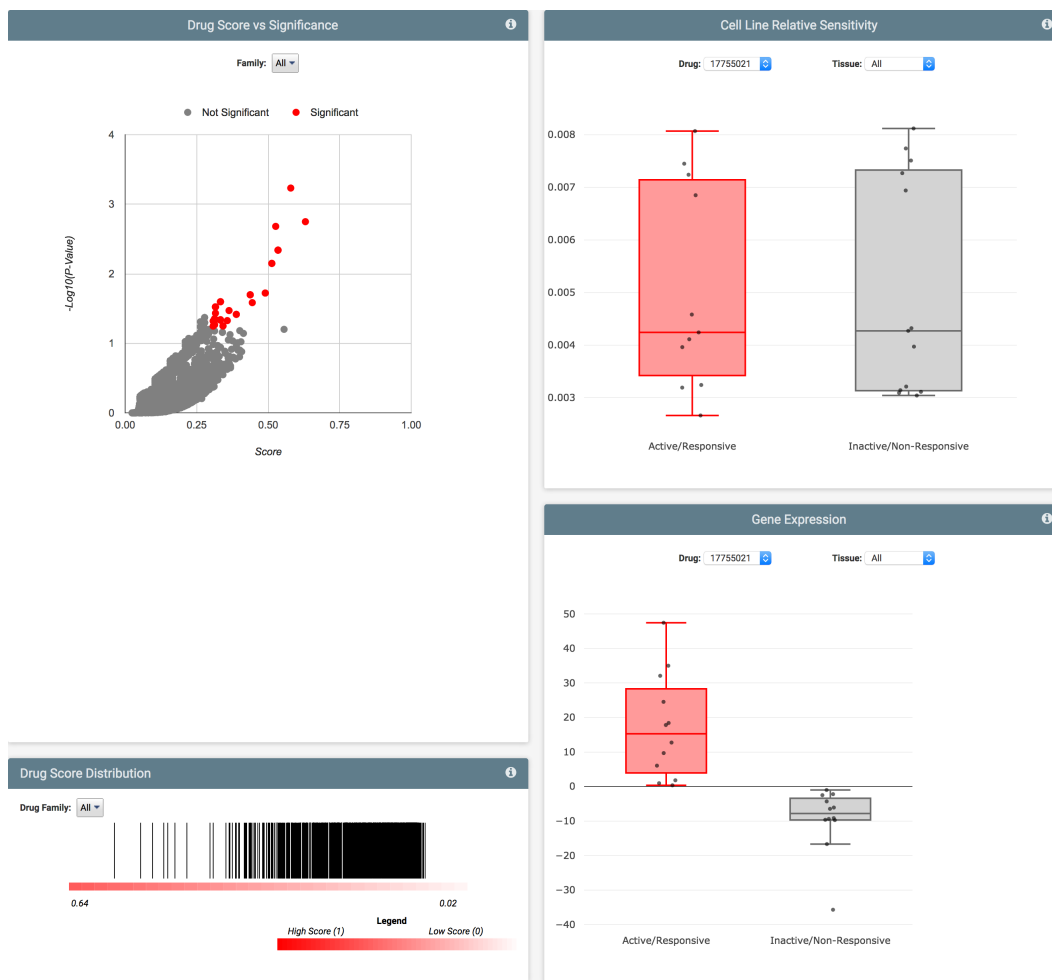
et al., Oncotarget 2015 for details on the statistical analysis) are highlighted in cyan. Complete results for all drugs can be downloaded as an Excel spreadsheet using the “Download” button. The “Drug Families” table lists the drug families that are significantly enriched, given the set of significantly active drugs (i.e. compounds highlighted in cyan in the “Drug” table). In this table, the cyan background highlights drug families comprising at least one compound with enrichment score ≥ 0.6 and p-value ≤ 0.05 .

The screenshot displays the GDA web interface. At the top, there is a navigation bar with 'GDA', 'Home', 'Contacts', and 'Help & Tutorial'. Below this, a breadcrumb trail shows 'Data Input / Data Input Recap / Results'. The main content area is titled 'Signature Analysis: UPREGULATED Tissues Selected: Breast, Central Nervous System, Colon, Leukemia, Lung, Melanoma, Ovarian, Renal'.

On the left, the 'Drugs Table' is shown. It has a search bar and a 'Show 10 entries' dropdown. The table lists 10 drugs with columns for SELECT, #, CID ID, NAME, DRUG FAMILY, MoA, SCORE, and P-VALUE. The first 10 rows are highlighted in cyan. Below the table, there are navigation buttons: 'First', 'Previous', 'Next', and 'Last'. At the bottom of the table, there are three buttons: 'DOWNLOAD', 'DRUGS CLUSTERING', and 'FROM DRUG TO GENE'.

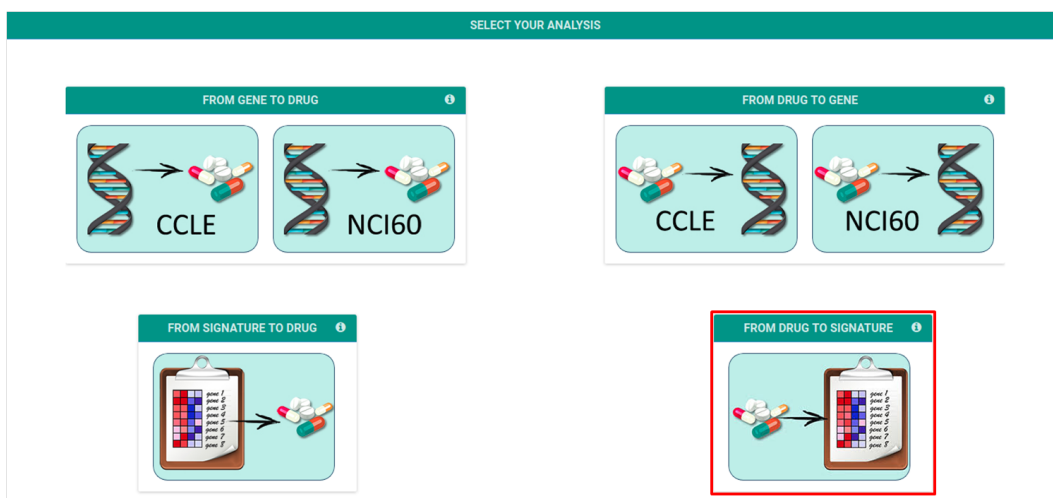
On the right, the 'Drug Families Table' is shown. It has a search bar and a 'Show 10 entries' dropdown. The table lists drug families with columns for DRUG FAMILY, # DRUGS (TOTAL, SCORE > 0.3, SCORE > 0.6), and P-VALUE (DRUGS WITH SCORE > 0.3, DRUGS WITH SCORE > 0.6). The first 10 rows are highlighted in cyan.

Results are also visualized by a set of graphs. In this module, plots display (from the upper left to the lower right): i) the scores as a function of the p-values for each drug and each drug family (with red dots highlighting significantly active drugs); ii) the distribution of the drug scores for each drug in each drug family; iii) the distribution of relative sensitivity (RS) in cell lines with an active (in this case upregulated) gene signature and responsive to a given drug (drug) as compared to the distribution of relative sensitivity in cell lines with an inactive (in this case downregulated) and non-responsive cell lines; and iv) the expression levels of the gene signature in cell lines with an active (in this case upregulated) gene signature and responsive to a given drug (drug) as compared to cell lines with an inactive (in this case downregulated) gene signature and non-responsive to a given drug. Directly from the result page, it is possible to access the **Drugs clustering** (see the *from gene to drugs* module for details) or, once a drug is selected, the *from drug to gene* analysis module (see the *from drug to gene* module for details).



The from drug to signature analysis module

The *from drug to signature* query (red box) allows identifying up- and down-regulated genes in cancer cell lines that are responsive to a specific compound. This query is run on the integrated gene expression profiles of **CCLE** and **NIC60** obtained using Affymetrix human gene expression arrays and Brainarray custom chip definition files.



Once the analysis has been selected, the user is redirected to the drug selection form, where it is possible to enter the name of a **single drug** of interest (i.e., **multiple drug names are not allowed**). In this tutorial, we will identify up- and down-regulated genes in cell lines responsive to **Oligomycin-A**, an inhibitor of

mitochondrial energy metabolism.

1. The drug name/NSC ID selection form provides an input bar where the user can type the name or the NSC ID of the drug of interest. Once the drug name is typed, click the “+” button to confirm the drug selection, which will appear at the center of the page. Multiple drug names (or NSC ID) are not allowed.

After the drug selection, press the **Confirm** button, at the bottom right of the page, to proceed to the next form.

2. After inputting the drug, it is possible to select the tissues of interest for the subsequent analysis. GDA default setting is to compare cell lines responsive against cell lines non-responsive to the selected compound across **all tissues**. Press the **Confirm** button to proceed to a recap page.

3. The recap page reports the list of cell lines that are responsive (left pane) and non-responsive (right pane) to the selected drug and indicates if each group has at least 3 samples, i.e. the minimum number of sample to perform a statistically acceptable differential gene expression analysis.

GDA Home Contacts Help & Tutorial

Drug Name / Tissues / GE Terms Recap

Cell lines **RESPONSIVE** to drug: Oligomycin-A in Non-Small Cell Lung, Colon, Leukemia, Breast, Ovarian, Prostate, Central Nervous System, Melanoma, Renal

- A549/ATCC (4 samples)
- COLO205 (3 samples)
- EKVX (3 samples)
- K-562 (4 samples)
- MCF7 (4 samples)
- NCI-H23 (4 samples)
- NCI-H522 (4 samples)
- OVCAR-8 (4 samples)
- PC-3 (4 samples)
- RPMI-8226 (4 samples)
- SF-295 (4 samples)
- SK-MEL-5 (4 samples)
- SNB-19 (4 samples)
- T-47D (4 samples)
- U251 (4 samples)
- UACC-257 (4 samples)
- UO-31 (3 samples)

Cell lines **NON-RESPONSIVE** to drug: Oligomycin-A in Non-Small Cell Lung, Colon, Leukemia, Breast, Ovarian, Prostate, Central Nervous System, Melanoma, Renal

- 786-0 (4 samples)
- A498 (4 samples)
- ACHN (4 samples)
- BT-549 (4 samples)
- CAKI-1 (3 samples)
- CCRF-CEM (4 samples)
- DU-145 (4 samples)
- HCC-2998 (3 samples)
- HCT-116 (4 samples)
- HCT-15 (4 samples)
- HOP-62 (3 samples)
- HOP-92 (3 samples)
- HS 578T (4 samples)
- HT29 (4 samples)
- IGROV1 (4 samples)
- KM12 (4 samples)
- LOX IMVI (4 samples)
- M14 (3 samples)
- MALME-3M (4 samples)
- MDA-MB-231/ATCC (4 samples)
- MDA-MB-435 (4 samples)
- MOLT-4 (4 samples)
- NCI-H322M (4 samples)
- NCI-H460 (4 samples)
- OVCAR-3 (4 samples)
- OVCAR-4 (3 samples)
- OVCAR-5 (4 samples)
- SF-268 (3 samples)
- SF-539 (3 samples)
- SK-MEL-2 (4 samples)
- SK-MEL-28 (3 samples)
- SK-OV-3 (4 samples)
- SN12C (3 samples)
- SNB-75 (3 samples)
- SR (4 samples)
- SW-620 (4 samples)
- TK-10 (3 samples)
- UACC-62 (4 samples)

Samples Number Check

Each group has at least 3 samples. Thus, differential gene expression analysis can be performed.

RUN SAM ANALYSIS
RUN T-TEST ANALYSIS

It is now possible to perform the differential expression analysis using either SAM (**Run SAM Analysis**) or the t-test (**Run t-test Analysis**)

- The [output page](#) provides the lists of genes upregulated in cancer cell lines of Group A (responsive; left panel) and in cancer cell lines of Group B (non-responsive; right panel). These lists can be directly functionally annotated using Enrichr, compared to results from L1000, used to generate gene signatures for further analysis in the [from signature to drug](#) module of GDA, or downloaded as an Excel file. Here, we use the functional annotation in Enrichr, clicking [Test Results on Enrichr](#) and selecting [FDR<0.05 \(Up in Group A\)](#). We focus on the subset of 1,540 genes that are up-regulated in cancer cell lines that are responsive to Oligomycin-A. In the Enrichr output page, select "Pathways" in the top layer to verify that sensitivity to mitochondrial inhibition is accompanied by the activation of oxidative phosphorylation, respiratory electron transport chain, and ATP synthesis pathways.

GDA Home Contacts Help & Tutorial

Drug Name / Tissues / GE Terms Recap / Results

Results for Oligomycin-A

Query Recap

Group A: cell lines RESPONSIVE to drug: Oligomycin-A in Non-Small Cell Lung, Colon, Leukemia, Breast, Ovarian, Prostate, Central Nervous System, Melanoma, Renal.
Group B: cell lines NON-RESPONSIVE to drug: Oligomycin-A in Non-Small Cell Lung, Colon, Leukemia, Breast, Ovarian, Prostate, Central Nervous System, Melanoma, Renal.

Genes Up-Regulated in Group A

Show 10 entries

#	Gene Name	Fold Change	False Discovery Rate
1	GJA1	3.64	0.0023
2	DKK1	3.55	0
3	PNMAL1	3.47	0
4	UCHL1	3.09	0.0013
5	TSPYL5	2.87	0
6	CYP24A1	2.69	0
7	LDOC1	2.68	0
8	LINC00461	2.67	0
9	TUBB2B	2.65	0
10	NEFL	2.6	0

Showing 1 to 10 of 1,540 entries

Genes Down-Regulated in Group A (Up in Group B)

Show 10 entries

#	Gene Name	Fold Change	False Discovery Rate
1	SCG5	-3.17	0
2	CAPG	-3.07	0
3	PRKDCBP	-3	0
4	AHR	-2.8	0
5	LGALS3BP	-2.7	0
6	RNF128	-2.68	0.0048
7	GBP1	-2.65	0
8	TM4SF1	-2.63	0.0016
9	FOXQ1	-2.56	0
10	ANXA3	-2.55	0.0079

Showing 1 to 10 of 602 entries

Test Results on Enrichr

Test Results on L1000

Test Results on GDA

HOME DOWNLOAD

No Results messages and negative cases with no results

GDA displays different *No Results* messages in case the analysis cannot be performed due to the absence of cell lines with a specific genomic background (as in the *from gene to drug* analysis) or of responsive cell lines to the drug selected in the query (as in the *from drug to gene* and *from drug to signature* analysis). Examples of *No Results* messages are as follows:

- a genomic background queried in the input forms of the *from gene to drug* module may result in no cell lines either in the mutant or in the wild-type group. In both cases, the *from gene to drug* analysis cannot be performed and GDA returns the following *No Results* messages:

case a. AMZ2, BRAF, DAPK3, EGFR, FAM159A, LDHA, ZMYND10 missense, missense, intron, silent, missense, intron, and 3'UTR mutations using CCLE genomic data on all tissues

Selected Hugo Symbols

AMZ2
Missense Mutation

BRAF
Missense Mutation

DAPK3
Intron

EGFR
Silent

FAM159A
Missense Mutation

LDHA
Intron

ZMYND10
3'UTR

Select Tissues from the list

HAEMATOPOIETIC AND LYMPHOID TISSUE

LARGE INTESTINE

Tissues and Cell Lines

For the selected genomic background, there are **no mutant samples** to perform the *from gene to drug* analysis

case b. ARHGEF12 intron mutations using CCLE genomic data on all tissues

- the NCI-60 Drug response data indicates that none of the NCI-60 cell lines are significantly responsive to the compound **Aspirin (NCS 27223)**. Indeed, the compound NSC 27223 is not included in the list of drugs and compounds with statistically significant activity used by the NCI-60 Analysis Tool CellMiner (https://discover.nci.nih.gov/cellminer/samples/IDENTIFIER_LIST.zip) and, when interrogating CellMiner (<https://discover.nci.nih.gov/cellminer/analysis.do>) for drug activity z scores of input NSC 27223, the result is the error message “None passed the range minimum of 1.2”. Similarly, when querying GDA with the drug Aspirin, the *from drug to gene* analysis retrieves no cell line in the responsive group and returns the following *No Results* message:

No Results

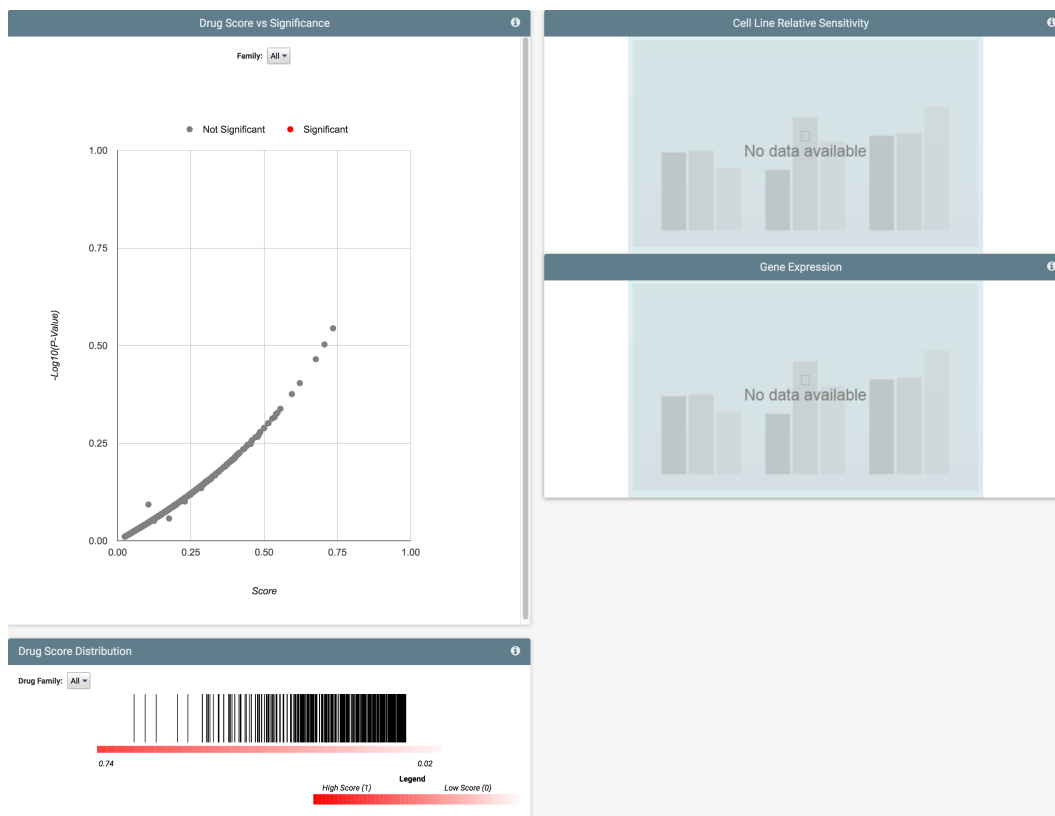
For the selected drug, the **responsive group** contains too few cell lines (i.e., ≤ 1) to perform the *from drug to gene* analysis

Although predisposed, a similar message will never appear for the non-responsive group as in the NCI-60 Drug response data (Sept 2014 release) there is no compound for which less than a single cell line is non-responsive.

In the case no drug is significantly active on the selected combination of mutations and cell lines (i.e., with enrichment score ≥ 0.3 and p-value ≤ 0.05) the results page of the *from gene to drug* module does not present the buttons to access the **Drugs clustering** or the **Differential expression** analysis modules and the Cell Line Relative Sensitivity and Gene Expression plots are empty:

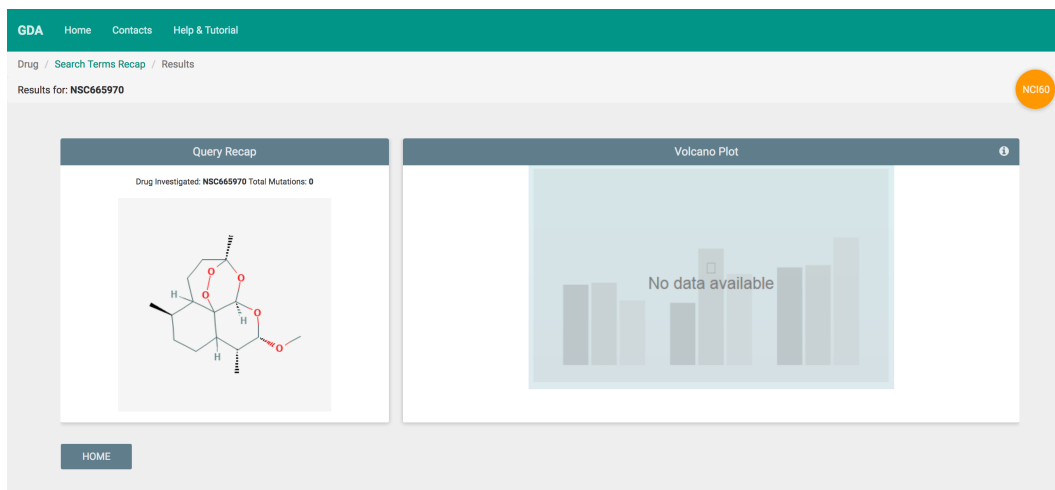
case c. MDM2 all mutations in all tissues using CCLE genomic data on all tissues

DRUG FAMILY	# DRUGS		P-VALUE	
	TOTAL	SCORE > 0.3	DRUGS WITH SCORE > 0.3	DRUGS WITH SCORE > 0.6
Quinone/Quinoline analogues	644	24	0	1
Purine/Pyrimidine analogues	21	4	0	1
Anthracyclines	21	1	0	0.262
DNA Polymerase Inhibitors	2	1	0	0.028
Sulfonate	60	2	0	0.213
Antimetabolite	9	4	0	1
Topoisomerase inhibitors	4	1	0	0.056
Imatinib analog	23	2	0	0.042

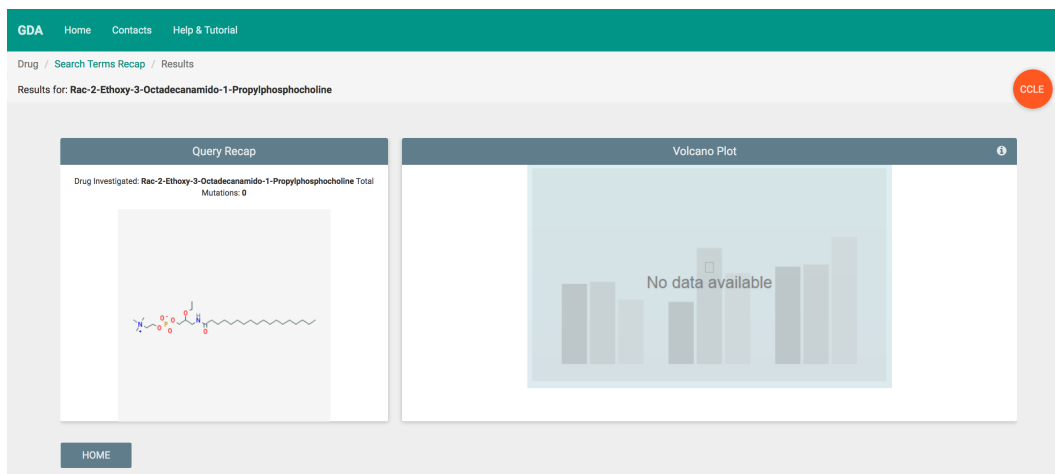


In the case there are no mutations significantly associated to a given compound in both responsive and non-responsive cancer cell lines, the results page of the *from drug to gene* module displays only the Query Recap panel (with the compound structure and total mutations equal to 0), while the Volcano Plot is empty and results for both the responsive and the non-responsive groups are not displayed:

case d. **Artemether (NSC 665970)** using NCI60 genomic data



case e. Rac-2-Ethoxy-3-Octadecanamido-1-Propylphosphocholine using either CCLE or NCI60 genomic data



Similarly, in the case no gene mutation is significantly associated to a given compound in only one of the two groups (either responsive or non-responsive), the results page of the *from drug to gene* module displays the results only for that group while the plots and the table for the other group are left empty:

case f. Yuanhuacin using NCI60 genomic data

