

A community effort to assess and improve drug sensitivity prediction algorithms

Ga-Young La

Molecular Cell and Developmental Biology Lab

Department of Biological Science

Konkuk University

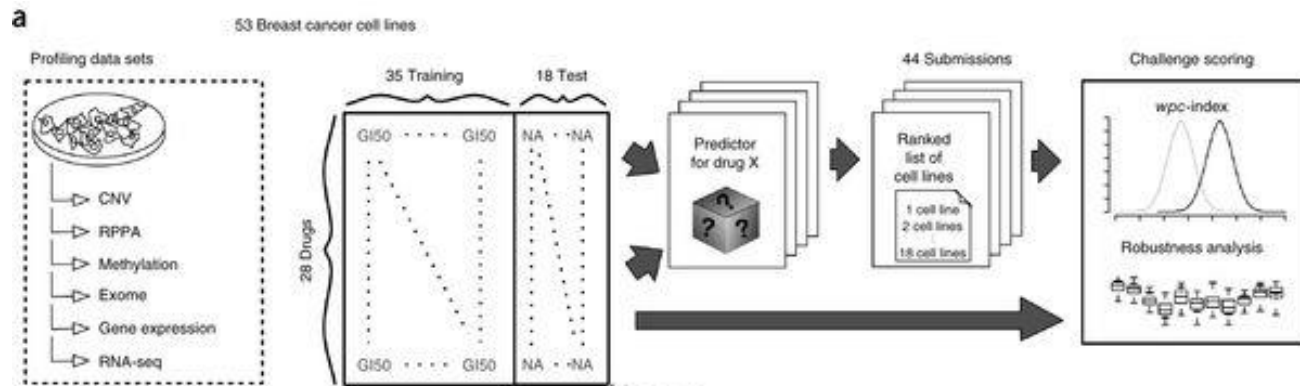
Abstract

Predicting the best treatment strategy from genomic information is a core goal of precision medicine.

They focus on predicting drug response based on a cohort of genomic, epigenomic and proteomic profiling data sets measured in human breast cancer cell lines.

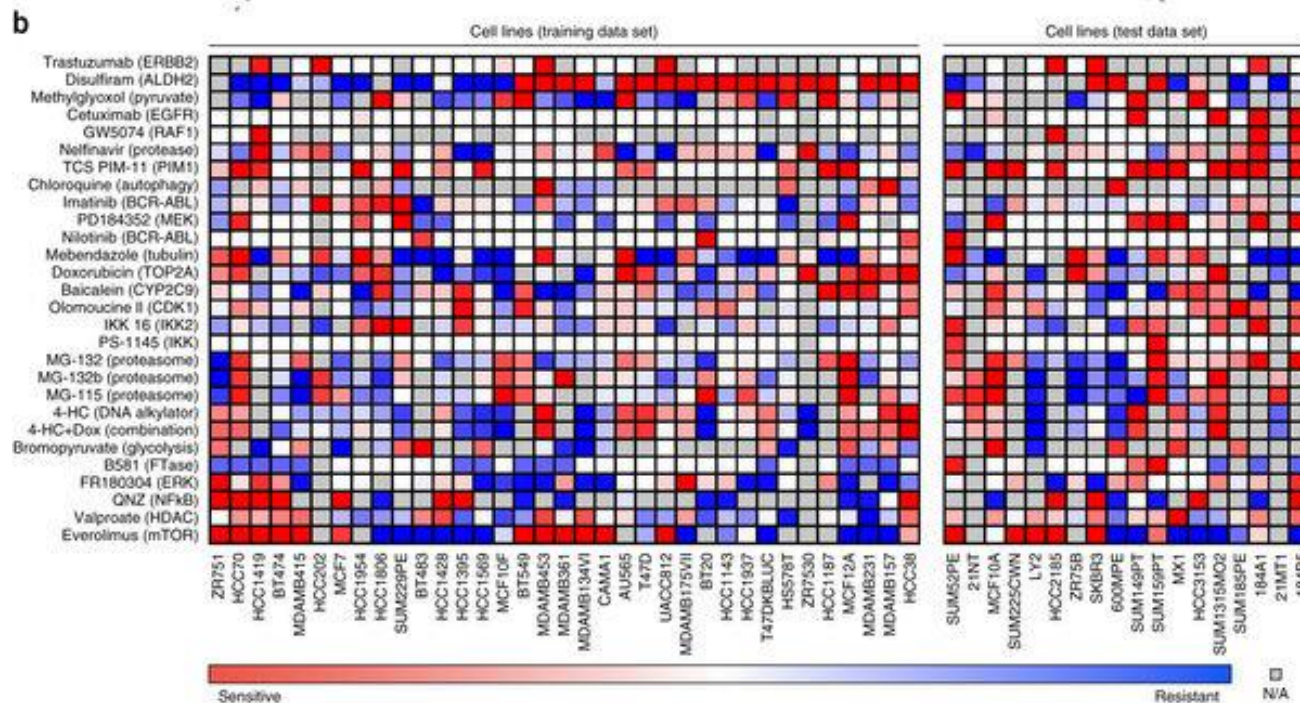
Through a collaborative effort between the **National Cancer Institute(NCI)** and the **Dialogue on Reverse Engineering Assessment and Methods(DREAM)** project, they analyzed a total of 44 drug sensitivity algorithms.

This study establishes benchmarks for drug sensitivity prediction and identifies approaches that can be leveraged for the development of new methods.



Summary of data sets and challenge

The NCI-DREAM drug sensitivity challenge.



Six genomic, epigenomic, and proteomic profiling data sets were generated for 53 breast cancer cell lines, which were previously described.

Participants were supplied with all six profiling data sets and dose-response data for 35 cell lines and all 28 compounds (training set).

NCI–DREAM drug sensitivity prediction methods

The 44 team submissions were categorized according to their underlying methodology.

gene expression(e)

exome sequencing(x)

RNA seq(n)

methylation(m)

RPPA(r)

copy number variation(c)

Team	Synopsis	wpc-index (scaled)	FDR	Data
Kernel method				
1	Bayesian multitask MKL (see main text).	0.583(0.629)	2.6 × 10 ⁻⁵	exnmrc OI
2	A predefined number of features were selected using Pearson correlation, training and prediction was done using support vector regression (SVR; radial basis).	0.559(0.592)	1.0 × 10 ⁻³	enmrc
3	Separate normalizations were applied to each dataset, several support vector machine (SVM) classifiers were independently trained (varying kernels and input data), final predictions were made using a weighted average of all SVM outputs.	0.553(0.582)	2.7 × 10 ⁻³	exnmrc
4	Bidirectional search was used to select features, training and prediction was done using a SVM (radial basis).	0.549(0.575)	4.8 × 10 ⁻³	enmrc

NCI-DREAM drug sensitivity prediction methods

The 44 team submissions were categorized according to their underlying methodology.

gene expression(e)

exome sequencing(x)

RNA seq(n)

methylation(m)

RPPA(r)

copy number variation(c)

Nonlinear regression (regression trees)				
1	Features were randomly selected to built an ensemble of unpruned regression trees for each dataset, missing values were imputed, weights for the models were calculated, final predictions were made using a weighted sum of the individual models.	0.577(0.620)	7.2×10^{-5}	enm
2	Features were filtered based on their correlation to dose-response values, random forests were trained for each dataset, missing values were imputed, final rankings were based on a composite score from four individual dataset models (enrc).	0.569(0.607)	2.9×10^{-4}	enrc OI
3	Features were filtered based on their correlation to dose-response values, random forests were trained for each dataset, missing values were imputed, final rankings were based on a composite score from five individual dataset models (enmrc).	0.565(0.601)	5.1×10^{-4}	enmrc OI
4	Features were filtered based on their correlation to dose-response values, random forests were trained for each dataset, missing values were imputed, final rankings were based on a composite score from five individual dataset models (exnrc).	0.564(0.599)	5.1×10^{-4}	exnrc OI

NCI–DREAM drug sensitivity prediction methods

The 44 team submissions were categorized according to their underlying methodology.

gene expression(e)

exome sequencing(x)

RNA seq(n)

methylation(m)

RPPA(r)

copy number variation(c)

Sparse linear regression				
1	Features were simultaneously selected and a ranking model built for each drug by lasso regression.	0.564(0.600)	5.1×10^{-4}	en
2	Features were initially filtered based on linear regression to drug response, training and prediction were done using elastic nets.	0.564(0.600)	5.1×10^{-4}	exnmrc
3	Gene and pathway features were determined using a one-dimensional factor analysis, training and predictions were made with spike and slab multitask regression, drug dose-response values were recalculated from raw growth curves.	0.564(0.598)	5.1×10^{-4}	exnmrc OI
4	Missing features were imputed, combinations of datasets were enumerated and used to train elastic net regression models, for each drug, final predictions were made using the best-performing model.	0.551(0.579)	3.3×10^{-3}	exmrc
5	Gene and pathway features were determined using a one-dimension factor analysis, training and predictions were made with spike and slab multitask regression, drug dose-response values were recalculated from raw growth curves, Heiser <i>et al.</i> data were used to train the model.	0.539(0.560)	1.9×10^{-2}	exnmrc OI

NCI–DREAM drug sensitivity prediction methods

The 44 team submissions were categorized according to their underlying methodology.

gene expression(e)

exome sequencing(x)

RNA seq(n)

methylation(m)

RPPA(r)

copy number variation(c)

PLS or PC regression				
1	Removed lowly expressed and/or low variance features, features were selected based on correlation to drug response, multiple partial least squares regression models were trained and consensus determined for final prediction.	0.562(0.597)	5.5 × 10 ⁻⁴	en OI
2	Features were selected by using lasso regression and groups of genes predefined by core signaling pathways, predictions were made by linear regression of the reduced feature set to drug response, predictor datasets were merged in advance of drug response prediction, and responses were predicted simultaneously sharing information among drugs.	0.543(0.567)	1.0 × 10 ⁻²	exnmrc OI
3	Training and prediction were done using principal component regression for individual drugs.	0.535(0.554)	3.1 × 10 ⁻²	exnmrc
4	Statistically significant features were selected using correlation, models were fit using principal component regression, final predictions were made using a weighted average of models.	0.524(0.538)	9.2 × 10 ⁻²	en

NCI–DREAM drug sensitivity prediction methods

The 44 team submissions were categorized according to their underlying methodology.

gene expression(e)

exome sequencing(x)

RNA seq(n)

methylation(m)

RPPA(r)

copy number variation(c)

Ensemble/model selection				
1	Features were selected using correlation, dimensionality reduced using principal component analysis, lasso and ridge method, several regression models were trained for individual drugs and the top cross-validated model was selected to make final predictions for each drug.	0.562(0.597)	5.5×10^{-4}	exnmrc
2	Features were selected on outside information, missing values were imputed, predictions were made by aggregating results from an ensemble of machine-learning methods.	0.556(0.587)	1.6×10^{-3}	exnmrc
3	Features were selected using Spearman's rank correlation, missing values were imputed, predictions were made using the best-performance method (determined by cross-validation on the training set) among an ensemble of methods (random forest, support vector machine and linear regression).	0.554(0.583)	2.6×10^{-3}	exnmrc
4	Gene and pathway features were compiled using outside data, an ensemble of prediction models were trained, final predictions were based on a rank-aggregation of combined prediction models.	0.517(0.527)	1.7×10^{-1}	exnmrc OI
5	Features were selected using outside pathway and interaction data, missing values were imputed, individual drug predictions were made using the best model selected from an ensemble of methods.	0.506(0.509)	3.7×10^{-1}	e OI

NCI-DREAM drug sensitivity prediction methods

The 44 team submissions were categorized according to their underlying methodology.

gene expression(e)

exome sequencing(x)

RNA seq(n)

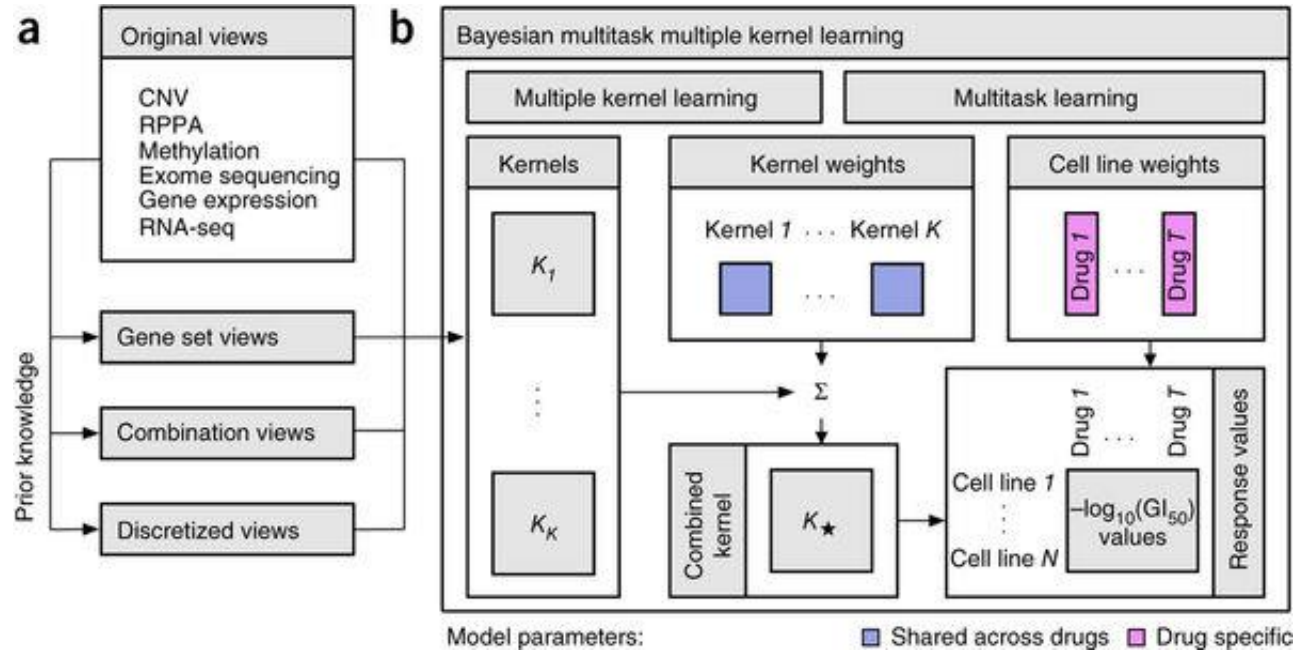
methylation(m)

RPPA(r)

copy number variation(c)

Other				
1	Features were weighted based on Pearson's correlation to drug response, predictions were made using the correlation of the weighted features.	0.570(0.608)	2.9×10^{-4}	enr
2	Gene features showing strong survival from the METABRIC dataset were selected, then hierarchically clustered, a linear model was built to fit gene clusters to drug response, predictions were made using a regression model.	0.553(0.582)	2.6×10^{-3}	e OI
3	Missing features were imputed, signatures were extracted for each dataset, predictions were made using 1-nearest-neighbor to training cell lines via Pearson's correlation between signatures for each data type, final predictions are the weighted sum of the individual datasets.	0.553(0.581)	2.7×10^{-3}	exnmrc
4	Features were selected using dataset-specific criteria, missing values were imputed, predictions were made using KNN.	0.531(0.549)	4.7×10^{-2}	exnmrc
5	Features were filtered using dataset-specific criteria, an ensemble of Cox regression models were constructed using random sampling from top-performing features, final prediction is the average of all models.	0.528(0.543)	6.5×10^{-2}	nmc
6	Features were selected using the concordance index, predictions were made using an integrated voting strategy based on each feature's ability to predict the order of pairs of cell lines.	0.521(0.532)	1.3×10^{-1}	enmrc

The method implemented by the best performing team



The top-performing team from Aalto University and the University for Helsinki developed a machine-learning method.

They present all dose-response values as $-\log_{10}(GI_{50})$, where GI_{50} is the concentration that inhibited cell growth by 50% after 72 hours of treatment.

The clustering of sensitive, resistant, and ambiguous

This table represents the clustering of sensitive(encoded as 1), resistant(encoded as 3), and ambiguous(encoded as 2).

Clustering of the cell lines was done on a drug-by-drug basis.

(table1)

Cell line	4-HC (DNA)	4-HC+Dox (1)	Baicalein (C)	Bromopyruv	Cetuximab (Chloroquine	Disulfiram (A	Doxorubicin	FR180304 (E	Everolimus (B581 (FTPas	GW5074 (R)	Trastuzuma	IKK 16 (IKK2
HCC1954	2	2	3	NA	3	NA	3	2	3	NA	2	NA	3	1
AU565	1	1	2	NA	3	NA	1	2	3	NA	2	3	3	2
HCC1937	1	2	2	NA	3	NA	1	3	3	2	2	3	NA	2
CAMA1	2	2	2	2	3	3	2	3	3	1	2	NA	3	2
T47DKBLUC	1	1	3	1	NA	NA	1	2	3	3	2	NA	NA	3
UACC812	1	2	2	2	3	NA	1	3	2	1	2	NA	2	3
HCC1569	3	3	2	1	3	NA	3	3	3	3	2	NA	3	2
MCF12A	1	1	1	1	3	3	2	2	3	3	2	3	3	3
HCC1187	NA	NA	1	NA	3	NA	1	NA	1	NA	2	NA	NA	3
HCC38	1	1	1	2	3	3	2	2	3	2	2	NA	3	2
SUM229PE	3	3	2	1	3	3	3	3	3	3	2	NA	NA	1
ZR751	1	1	2	1	3	3	NA	2	1	1	2	3	NA	3
BT483	NA	NA	2	1	3	NA	3	NA	3	2	2	NA	3	NA
T47D	1	1	1	NA	3	NA	1	2	3	2	2	NA	NA	2
ZR7530	3	3	NA	NA	NA	NA	1	1	NA	NA	NA	NA	NA	NA
BT549	NA	NA	1	NA	3	NA	1	NA	3	1	2	3	3	2
MDAMB231	2	1	1	2	3	3	2	2	NA	3	2	NA	3	2
MDAMB453	1	1	3	2	3	2	1	3	3	1	2	3	2	2
MCF10F	3	3	3	NA	3	NA	3	3	3	2	2	NA	3	3
MDAMB157	1	1	2	1	3	2	1	2	3	3	2	3	NA	1
HCC1428	1	2	2	NA	3	NA	2	3	2	3	2	NA	3	2

Classification of cell lines by resistance

This table represents the clustering results.

Clustering of the cell lines was done on a drug-by-drug basis.

(table2)

	4-HC (DNA alkylator)	Doxorubicin (TOP2A)	Cetuximab (EGFR)	FR180304 (ERK)	Everolimus (mTOR)	MG-132 (Proteasome)	QNZ (NfκB)	MG-132b (Proteasome)	MG-115 (Proteasome)
sensitive cell line	HCC70 UACC812 MDAMB453 HCC38 T47D MCF10A HCC1937 SUM1315MO2 SUM149PT MDAMB175VII HCC1428 MCF12A MDAMB157 T47DKBLUC MX1 AU565 21NT	ZR7530	184B5	184B5 SUM1315MO2 SUM149PT 184A1 ZR751 MDAMB175VII BT474 HCC1143 600MPE HCC1419 MDAMB361 HCC1187 SUM225CWN	ZR751 BT474 HCC1419 MDAMB361 SUM225CWN HCC70 UACC812 SKBR3 MDAMB453 CAMA1 MDAMB415 SUM52PE MDAMB134VI BT549 HCC2185 LY2	SUM225CWN HCC70 MDAMB415 SUM52PE BT549 184B5 184A1 T47D MCF10A MCF10F SUM159PT SUM185PE MCF12A SUM229PE HCC1187 AU565	HCC70 184B5 BT474 HCC1419 HCC2185 HCC1395 ZR751 HCC38 HCC1428	HCC70 HCC3153 MCF7 BT549 MCF10A SUM159PT MCF12A MDAMB157 BT20 SUM52PE MCF10F HCC1937 SUM1315MO2 HCC202 21NT 21MT1	HCC70 MCF10A SUM159PT MCF12A BT20 MCF10F SUM1315MO2 HCC202 21NT MDAMB361
resistant cell line	MDAMB134VI LY2 MCF10F 600MPE HCC1395 HCC1569 SUM229PE BT20 ZR7530 21MT1	UACC812 MDAMB453 HCC1937 SUM149PT HCC1428 MX1 21NT BT474 CAMA1 MDAMB415 SUM52PE HCC1143 MCF7 HCC3153 HS578T HCC202 MDAMB134VI	ZR751 HCC70 HCC38 T47D MCF10A MCF12A MDAMB157 AU565 SKBR3 MDAMB157 AU565 SKBR3 HCC1937 MDAMB231 HCC1806 ZR758 HCC1954 UACC812 MDAMB453 HCC1937	HCC38 T47D MCF10A MCF12A MDAMB157 AU565 SKBR3 HCC1954 MDAMB453 HCC1937 CAMA1 MDAMB415 SUM52PE MCF7 HCC3153 HS578T MDAMB134VI	SUM1315MO2 SUM149PT MDAMB175VII 600MPE HCC1428 HCC1395 MCF12A MDAMB157 HCC3153 HCC1569 SUM229PE T47DKBLUC MX1	ZR751 SKBR3 MDAMB453 LY2 HCC38 MCF7 600MPE HCC1428 MDAMB231 HCC1806 ZR758	BT549 184A1 MCF10A SUM159PT MCF12A AU565 MDAMB361 HCC1143 SUM149PT MDAMB157 600MPE MDAMB231	SUM149PT MDAMB415 ZR758	MCF7 ZR751 SKBR3 600MPE HCC1806 UACC812 T47DKBLUC MX1 LY2 SUM149PT MDAMB415 ZR758

Classification of cell lines by eIF2 α phosphatase

This table represents the clustering results.

Clustering of the cell lines was done on a drug-by-drug basis.

(table3)

	4-HC (DNA alkylator)	Doxorubicin (TOP2A)	Cetuximab (EGFR)	FR180304 (ERK)	Everolimus (mTOR)	MG-132 (Proteasome)	QNZ (NFkB)	MG-132b (Proteasome)	MG-115 (Proteasome)
sensitive cell line	HCC70 UACC812 MDAMB453 HCC38 T47D MCF10A HCC1937 SUM1315MO2 SUM149PT MDAMB175VII HCC1428 MCF12A MDAMB157 T47DKBLUC MX1 AU565 21NT	ZR7530	184B5	184B5 SUM1315MO2 SUM149PT 184A1 ZR751 MDAMB175VII BT474 HCC1143 600MPE HCC1419 MDAMB361 HCC1187 SUM225CWN	ZR751 BT474 HCC1419 MDAMB361 SUM225CWN HCC70 UACC812 SKBR3 MDAMB453 CAMA1 MDAMB415 SUM52PE MDAMB134VI BT549 HCC2185 LY2	SUM225CWN HCC70 MDAMB415 SUM52PE BT549 184B5 HCC70 184A1 T47D MCF10A MCF10F SUM159PT SUM185PE MCF12A SUM229PE HCC1187 AU565	HCC70 184B5 BT474 HCC1419 HCC2185 HCC1395 ZR751 HCC38 HCC1428	HCC70 HCC3153 MCF7 BT549 MCF10A SUM159PT MCF12A MDAMB157 BT20 SUM52PE MCF10F HCC1937 SUM1315MO2 HCC202 21NT 21MT1	HCC70 MCF10A SUM159PT MCF12A BT20 MCF10F SUM1315MO2 HCC202 21NT MDAMB361
resistant cell line	MDAMB134VI LY2 MCF10F 600MPE HCC1395 HCC1569 SUM229PE BT20 ZR7530 21MT1	UACC812 MDAMB453 HCC1937 SUM149PT HCC1428 MX1 21NT BT474 CAMA1 MDAMB415 SUM52PE HCC1143 MCF7 HCC3153 HS578T HCC202 MDAMB134VI	ZR751 HCC70 HCC38 T47D MCF10A MCF12A MDAMB157 MDAMB175VII AU565 SKBR3 MDAMB157 AU565 SKBR3 HCC1937 MDAMB231 HCC1806 ZR75B HCC1954 UACC812 MDAMB453 HCC1937	HCC38 T47D MCF10A MCF12A MDAMB157 AU565 SKBR3 MDAMB157 AU565 SKBR3 HCC1937 CAMA1 MDAMB415 SUM52PE MCF7 HCC3153 HCC1569 SUM229PE T47DKBLUC MX1	SUM1315MO2 SUM149PT MDAMB175VII 600MPE HCC1428 HCC1395 MCF12A MDAMB157 HCC3153 HCC1569 SUM229PE T47DKBLUC MX1	ZR751 SKBR3 MDAMB453 LY2 HCC38 MCF7 600MPE HCC1428 MDAMB231 HCC1806 ZR75B	BT549 184A1 MCF10A SUM159PT MCF12A AU565 MDAMB361 HCC1143 SUM149PT MDAMB157 600MPE MDAMB231	SUM149PT MDAMB415 ZR75B	MCF7 ZR751 SKBR3 600MPE HCC1806 UACC812 T47DKBLUC MX1 LY2 SUM149PT MDAMB415 ZR75B

NCBI

Resources

How To

Sign in to NCBI


GEO DataSets

GEO DataSets

Search

Advanced

Help



GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. Enter search terms to locate experiments of interest. DataSet records contain additional resources including cluster tools and differential expression queries.

Getting Started

- [GEO Documentation](#)
- [GEO FAQ](#)
- [About GEO DataSets](#)
- [Construct a Query](#)
- [Download Options](#)

GEO Tools

- [Submit to GEO](#)
- [Advanced Search](#)
- [DataSet Browser](#)
- [Programmatic Access](#)
- [GEO2R](#)

More Resources

- [GEO Home](#)
- [GEO Profiles](#)
- [SRA](#)

Example Searches

Keywords and species	(smok* OR diet) AND (mammals[organism] NOT human[organism])
Study type	"expression profiling by high throughput sequencing"[DataSet Type]
Studies with CEL files	cel[Supplementary Files]
DataSets that have 'age' as an experimental variable	age[Subset Variable Type]
Studies with between 100 and 500 samples	100-500[Number of Samples]
Author	smith a[Author]

GEO DataSets

This database stores curated gene expression DataSets, as well as original Series and Platform records in the Gene Expression Omnibus(GEO) repository.

Enter search terms to locate experiments of interest.

About GEO2R

Background

- GEO2R is an interactive web tool that allows users to compare two of more groups of Samples in a GEO Series in order to identify genes that are differentially expressed across experimental conditions.
- GEO2R performs comparisons on original submitter-supplied processed data tables using the GEOquery and limma R packages from the Bioconductor project.

About GEO2R

How to use - Enter a Series accession number

- If you followed a link from a Series record, the GEO accession box will already be populated.
- Otherwise, enter a Series accession number in the box, e.g., GSE25724.

About GEO2R

How to use - Define Sample groups

- In the Sample panel, click ‘Define groups’ and enter names for the groups of Samples you plan to compare, e.g., *test* and *control*.
- Up to 10 groups can be defined.

About GEO2R

How to use - Assign Sample to each group

- To assign Samples to a group, highlight relevant Sample rows.
- Multiple rows may be highlighted either by dragging the cursor over contiguous Samples, or using Ctrl or Shift keys.

About GEO2R

How to use - Assign Sample to each group

▼ Samples

▼ Define groups

Selected 8 out of 8 samples

Enter a group name: [List](#)

☐ Cancel selection

☐ space flown (4 samples)

☐ control (4 samples)

Columns

Set

Group	Accession	Tissue	Source name	Strain	Tissue
space flown	GSM458594	Space-Flown Thymus (FLT)-4	Thymus mRNA extracted from space-flown mice	C57BL/6NTac	thymus
space flown	GSM458595	Space-Flown Thymus (FLT)-4	Thymus mRNA extracted from space-flown mice	C57BL/6NTac	thymus
space flown	GSM458596	Space-Flown Thymus (FLT)-4	Thymus mRNA extracted from space-flown mice	C57BL/6NTac	thymus
space flown	GSM458597	Space-Flown Thymus (FLT)-4	Thymus mRNA extracted from space-flown mice	C57BL/6NTac	thymus
control	GSM458598	Control Thymus (AEM)-1	Thymus mRNA extracted from ground-control mice	C57BL/6NTac	thymus
control	GSM458599	Control Thymus (AEM)-2	Thymus mRNA extracted from ground-control mice	C57BL/6NTac	thymus
control	GSM458600	Control Thymus (AEM)-3	Thymus mRNA extracted from ground-control mice	C57BL/6NTac	thymus
control	GSM458601	Control Thymus (AEM)-4	Thymus mRNA extracted from ground-control mice	C57BL/6NTac	thymus

About GEO2R

How to use - Perform the test

- After Samples have been assigned to groups, click [Top 250] to run the test with default parameters.
- Alternatively, you can use features in the other tabs to first assess the Sample value distributions, or edit default test parameters.

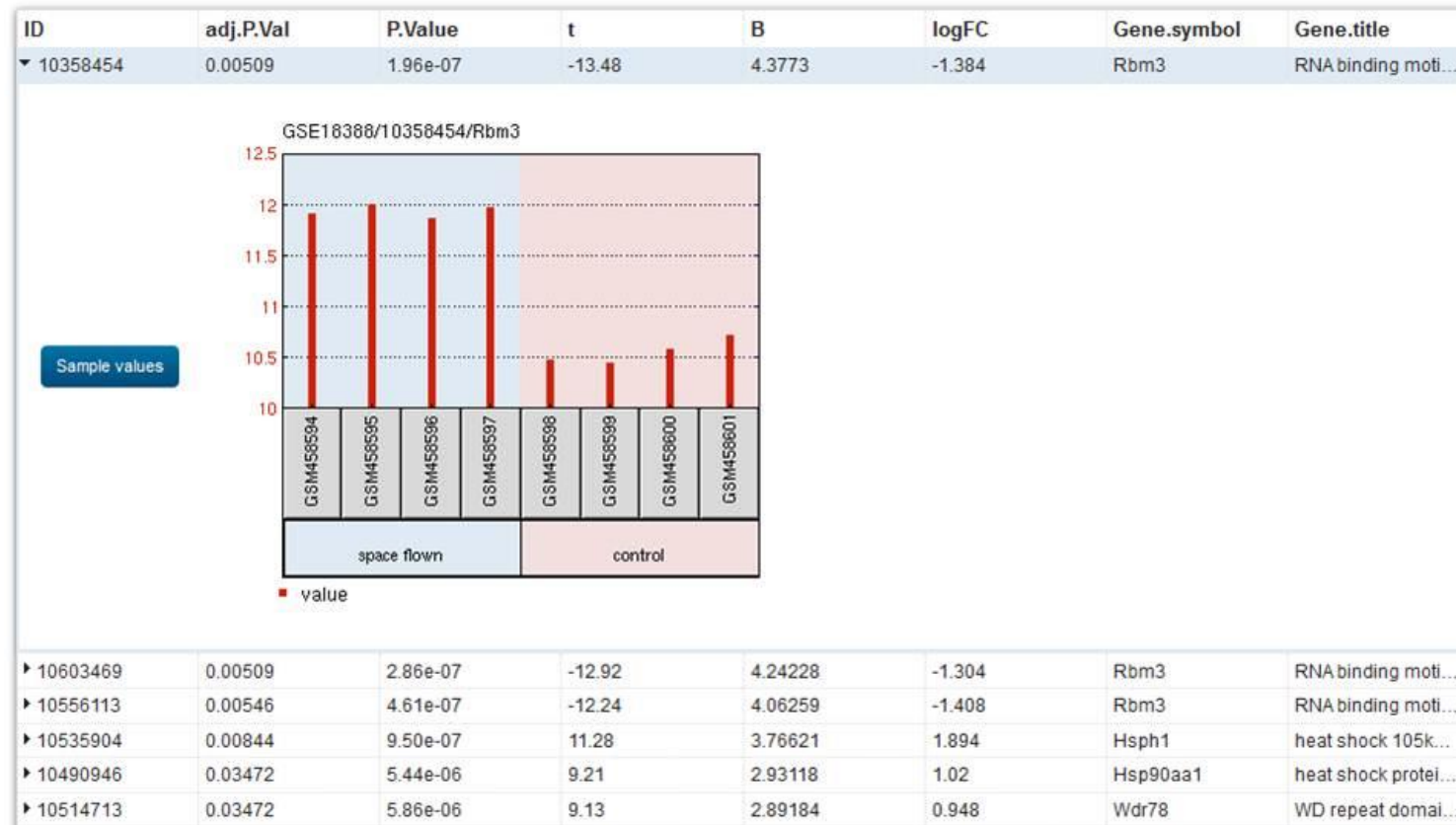
About GEO2R

How to use - Interpret the results table

- Results are presented in the browser as a table of the top 250 genes ranked by P-value.
- Genes with the smallest P-value are the most significant.
- Each red bar in the graph represents the expression measurement extracted from the value column of the original submitter-supplied Sample record.

About GEO2R

How to use - Interpret the results table



Thank you for
listening!