DBPOM: a comprehensive database of

pharmaco-omics for cancer precision medicine

Yijun Liu¹, Fuhu Song¹, Zhi Li³, Liang Chen^{4,5}, Ying Xu⁶, Huiyan Sun^{1,2,*}, Yi Chang^{1,2,*}

¹ School of Artificial Intelligence, Jilin University, 2699 Qianjin Street, Changchun 130012, P. R. China

² International Center of Future Science, Jilin University, 2699 Qianjin Street, Changchun 130012, P. R. China

³ Medical Oncology Department, The First Hospital of China Medical University, Shenyang 110001, China

⁴ Department of Computer Science, College of Engineering, Shantou University, Shantou 515063, China

⁵ Key Laboratory of Intelligent Manufacturing Technology of Ministry of Education, Shantou University, Shantou 515063, China

⁶ Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, the University of Georgia, Athens, GA 30602, United States

* To whom correspondence should be addressed. Huiyan Sun, School of Artificial Intelligence, Jilin University, No.2699 Qianjin Street, Changchun, Jilin, 130012, China.Email: huiyansun@jlu.edu.cn. Yi Chang, School of Artificial Intelligence, Jilin University, No.2699 Qianjin Street, Changchun, Jilin, 130012, China.Email: yichang@jlu.edu.cn.

Abstract

During the course of cancer treatment, both efficacy and the adverse effects of drugs on patients should be taken into account. Although some public databases and modeling frameworks have been developed through studies on drug response, the negative effects of drugs are always neglected. Furthermore, most of them only considered the ramifications of the drug on the cell line, but the effects on the patient still require a huge amount of work to integrate data from various databases and calculations, especially in relation to precision treatment. In order to address these issues, we developed the DBPOM (http://www.dbpom.net/, a comprehensive database of pharmaco-omics for cancer precision medicine), which explores various drugs' efficacy levels by calculating their potency in reverse, or enhancing cancer-associated gene expression changes. When compared with existing databases, the DBPOM could estimate the effectiveness of a drug on individual patients through the mapping of various cell lines to each person according to their genetic mutation similarities. The DBPOM is an easy-to-use and one-stop database for clinicians and drug researchers to search and analyze the overall effect of a drug or a drug combination on cancer patients as well as the biological functions that they target. We anticipate that DBPOM will become an important resource and analysis platform for drug development, drug mechanism studies and the discovery of new therapies.

Keywords: cancer precision medicine, drug efficacy, adverse effect, mapping from cell to tissue, database.

Introduction

Patients of the same cancer type, or even those at the same stage, may respond differently to the same drug[1, 2]; hence, a comprehensive study of drug screens and the mechanism of drug actions on patients with specific pathological and/or physiological conditions will be needed to facilitate improved individualized therapy[3, 4]. Pharmaco-omics is the use of genomics and transcriptomics-based information to help guide individualized drug therapy to maximize drug

efficacy and minimize adverse drug reactions[5]. In the early stages of a drug's development, its efficacy, toxicity, and sensitivity are typically tested on cell lines[6]. Currently, the potential treatment outcomes of a drug on a cancer might be assessed by using certain indices like half-maximal inhibitory concentration (IC50). In addition, another increasingly accepted measurement is the reversed expressions of genes, which are oppositely regulated between drug treatment information and disease-state data[7]. However, the adverse effects of drugs are usually ignored, which is a major concern for both public health and the development of medicines as failing to identify these negative outcomes could lead to significant amounts of morbidity[8]. Consequently, fully considering both reversed expressions and adverse effects is a highly desirable strategy[9, 10]. What is more, as cancer cells may develop drug resistance, it has been suggested that treatments that employ drug combinations potentially enhance efficacy and reduce toxicity[11]. Therefore, the mechanism of synergy and new combination recommendations have acted as a catalyst for intensive studies by academic researchers and pharmaceutical enterprises.

Recently, with the development of high-throughput technology, large amounts of cancer multi-omics data and related therapy data have been released that facilitate studies. The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga)[12], molecularly characterizes over 20,000 primary cancer and matched normal samples, spanning 33 cancer types with detailed clinical information. Meanwhile, the Gene Expression Omnibus database (GEO, https://www.ncbi.nlm.nih.gov/geo/)[13] is a widely used public data repository of array and sequence-based multi-omics profiles, but it is not limited to cancer cells and samples. Also, the Cancer Cell Line Encyclopedia (CCLE, https://portals.broadinstitute.org/ccle)[14] collects and collates genomic, epigenomics data and transcriptomics data of more than 1000 cancer cell lines.

Some public databases and modeling frameworks have also been developed and widely used for studies of cancer precision medicine and drug response, such as CMap[15], GDSC[16], DSigDB[17], and CiDD[18]. While useful, most of these databases and computational tools only consider the effect of the drug on the cell line but not on cancer tissues, resulting in a huge amount of work to integrate data from various databases, and then calculation is still required to evaluate both the therapeutic and adverse effects on the patient, particularly during precision

treatment, which makes it extremely difficult for clinicians and drug researchers to utilize them[19]. Here, we propose the creation of an open source, comprehensive database of pharmaco-omics for cancer precision medicine that is supported by the collection and storage of drug information, and analysis of the outcomes of drug integrative effect at both a cellular and individual patient level.

In order to evaluate the effect of a drug on individual patients and not just at a cellular level, we quantify both reversed and adverse outcomes by calculating differential expressions of all the genes between the cancer sample and the normal sample, the cancer cell line with and without drug treatment, respectively, and the mapping between cell lines and cancer tissues based on similarities in genetic mutations. By defining the score function through the presentation of the integrated drug effect and applying it to FDA-approved medicines, we find that most FDA-approved drugs, and the ones in clinical trials, tend to have a low DBPOM-defined score, which suggests that the DBPOM provides an effective means of appraising the usefulness and safety of drugs. It is also promising for precision medicine and drug-combination recommendations.

In summary, the DBPOM is an easy-to-use database for clinicians and drug researchers to search and analyze both the therapeutic and adverse effects of a drug/drug combination on cancer patients in accordance with the demand for personalized drug analysis, making it unlike similar databases that only focus on cell lines. By analyzing the biological function of reversed and adverse genes when different drug or drug combination treatments are being administered, the DBPOM can provide possible insights into studies of a drug's mechanism and play a guiding role in drug development and recommendations. Furthermore, since the DBPOM provides a patient-centric module that focuses on the changes to the molecular features of each patient, it is naturally useful for querying and designing more effective personalized treatments.

Methods

The overall design of the DBPOM is outlined in Figure 1.

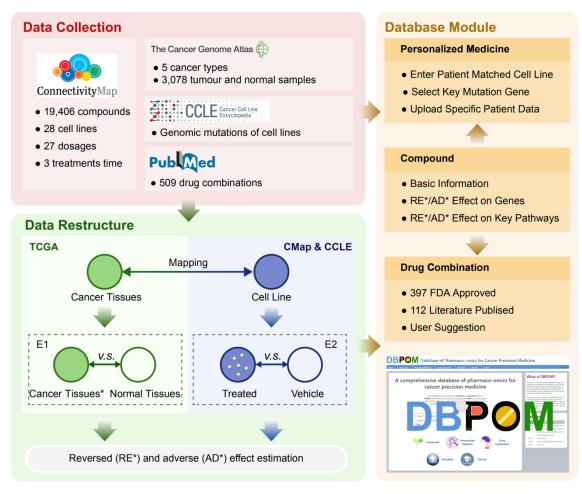


Figure 1. The procedures of the DBPOM database construction.

Data source and pre-processing of the DBPOM

We collected genomics and transcriptomics data of over 3,000 cancer and normal samples across five cancer types from TCGA, and 28 relative cancer-cell lines from CCLE, respectively. This was as well as gathering more than 14,000 compound-perturbated gene expression profiles from CMap, nearly 4,000 IC50 profiles, which is a quantitative measure of drug efficacy from the GDSC, and over 500 examples of drug combination information from the literature[20]. In addition, five cancer types that each has at least two forms of cell line have been used for precision medicine analyses. In total, 28 cell lines and 3,078 cancer samples have been selected from the largest intersection between the CMap, the CCLE and TCGA. After removing 18 pseudogenes, we retained 12,310 expressed genes and their expression under various conditions from CMap. 19,406 small molecule compounds and drugs were determined via the selected 28 cell lines, 27

dosages and three treatment durations. We manually annotated the Ensembl ID of the gene in TCGA and the gene name from the CMap. Detailed information can be downloaded in the download module. All data was collected before September 2019. **Table 1** summarizes the key information stored in the database.

Key information
19,406 Small molecular compounds and drugs
28 Cell lines
5 Cancer types
27 Dosages
3,078 TCGA cancer samples and normal samples
2,598 Specific mutations with high mutation frequency
397 FDA approved cancer drug combination pairs
112 Literature published cancer drug combination pairs
Total 376,486 files

Table 1. Statistical information about the DBPOM

Differential gene expression analysis and comparisons

Differential gene expression analyses have been performed on both cancer versus control tissues samples (E1) from TCGA and molecular-compound-treated cell lines versus DMSO-treated cell lines (E2) from the CMap database.

With regards to E1, to identify which genes are associated with carcinogenesis at the gene expression level, a DESeq2 package[21] was used for differential expression analysis between normal and cancer samples in TCGA. We set the threshold as p-value=0.001 and fold-change=2. For each cancer type, we defined the gene as up-regulated (down-regulated) and labelled it as 1 (-1) if its log2 fold-change is above 1 or less than -1 and the p-value is less than 0.001 when comparing cancer samples with normal samples; otherwise, it was labelled as 0.

In relation to E2, to evaluate the effect of compounds on the cell line at the gene level, the following formula was used for differential expression analysis between DMSO-treated and compound-treated cell lines in the CMap. We defined genes as significantly differentially expressed if their fold-change is above 15. For each cell line, we define the gene as up-regulated (down-regulated) and label it 1 (-1) if its fold-change is above 15 or less than -15 when comparing compounds with DMSO-treated cell lines; otherwise, it was labelled as 0.

$$FC(c^{j}) = \frac{T(c^{j}) - D(c^{j})}{Max\{|D(c^{j})|, 0.1\}}$$

Where $T(c^{j})$ and $D(c^{j})$ represent the mRNA expression value of the *j*-th gene in the compound treated group and the DMSO-treated group, respectively. $FC(c^{j})$ represents the fold-change value. In order to prevent overflow, we set 0.1 as the lowest expression value.

Disease gene expression signatures and assessment

Within the DBPOM, we have developed a tool for distinguishing the reversed and adverse genes of a specified compound, or drug on cancer samples, by comparing the data stemming from the two groups of differential gene expression results (E1 and E2). When a cell line is treated with a compound, if the effect on a gene is oppositional, meaning it is up-regulated in E1 and down-regulated in E2 or vice versa, we consider it to have a reversed effect on said gene. Otherwise, if a gene is marked as $1(E1) \sim 1(E2)$, $-1(E1) \sim -1(E2)$, $0(E1) \sim 1(E2)$ or $0(E1) \sim -1(E2)$, with 0 standing for no change, we consider the compound to have an adverse effect on this specific gene. For each compound, if it reverses the expression of a cancer-associated gene, it is thought to be effective on the gene. On the contrary, if the gene expression substantially deviates after treatment with the compound, we believe that the drug has had an adverse effect on the gene. It should be noted that the adverse effects defined here are different from the pharmacological adverse reactions in clinical trials.

The formulaic expression is as follows:

Reversed effect

$$\begin{cases}
In E1 : FC(c^{j}) > 2 \text{ and In E2} : FC(c^{j}) < -2 \\
In E1 : FC(c^{j}) < -2 \text{ and In E2} : FC(c^{j}) > 2
\end{cases}$$
Reversed effect

$$\begin{cases}
In E1 : FC(c^{j}) = 0 \text{ and In E2} : |FC(c^{j})| > 15 \\
In E1 : FC(c^{j}) > 2 \text{ and In E2} : FC(c^{j}) > 15 \\
In E1 : FC(c^{j}) < -2 \text{ and In E2} : FC(c^{j}) < -15
\end{cases}$$

In order to estimate the overall effect of a drug on each patient, we integrated both the reversed and adverse effects as described above of all of the genes. Moreover, we defined the score as the reversed and adverse of drug effect (RADE) to indicate the safety and effectiveness of each compound. RADE comprehensively considered the number of reversed genes and adverse genes after being treated by the compound. The lower the score, the higher the effectiveness and safety. The formulaic expression is as follows:

$$RADB_{c}^{k} = \frac{\#Ad^{k}(c)}{\#Re^{k}(c)} * \frac{\#Mu(k)}{\#Re^{k}(c)}$$

The need to explain pound sign (#) represents the number of elements in the set. In the above function, $RADB_c^k$ signifies the efficacy score of compound treatment on cell line *k*. $Ad^k(c)$ and $Re^k(c)$ are the set of adverse and reversed genes in cell line *k* with the treatment of the compound, respectively. Mu(k) is the gene set that should be reversed according to previous TCGA analyses.

Mapping individual mutation phenotype to cancer cell line

Genetic mutations of different cancer cell lines of the same cancer type vary, which results in them responding divergently to the same drug[22]. Meanwhile, as an important reference for molecular subtypes, genomic mutation could reflect the individual characteristics of different patients with the same type of cancer. Therefore, through the estimation of the cell line's response to drugs and by establishing the association between cancer cell lines and the similarities between cancer patient-based genetic mutations, it is helpful for recommending personalized medicine.

We collected the genomic mutation information of each cell line and each cancer patients from the CCLE and TCGA database, respectively, and defined the similarity score of each cancer cell line and cancer tissue of the same cancer type in the following manner:

$$Match_{k}(s) = \frac{\#DE(s) \cap Mutation(cellline_{k})}{\#Mutation(cellline_{k})}$$

$$L(s) = \arg\max_{k} Match_{k}(s)$$

 $Match_k(s)$ is the matching score of the sample on cell line *k* and the sample is from corresponding cancer type provided by TCGA. DE(s) is the set of mutation genes in the sample and $Mutation(cellline_k)$ is the set of mutation genes of cell line *k*. L(s) represents the sample from TCGA, which is roughly labelled as a cancer type in TCGA, and it has been chosen for re-labelling as the cell line with the highest matching score.

There are three reference options for mapping according to the samples selected from E1 and the mapping method we proposed:

(i) TCGA versus CMap: in E1, all of the cancer samples of the selected cancer type in TCGA are utilized. While this is the general strategy in current research, it does not consider individual differences. We provide this option in case some users need it.

(ii) subTCGA-mutaion versus CMap: in E1, the subset of cancer samples in TCGA which have high matching scores with selected cell line are utilized. We provide this strategy as overall mutation pattern is an important criterion for distinguishing different cell lines. So, we provide users with this option to suit diverse requirements. In practice, for each cancer patient, he or she is firstly classified into the most suitable sample group according to mutation similarity, and then to analyze effect of drugs on him or her.

(iii) subTCGA-specific versus CMap: E1 utilizes a subset of cancer samples with the specific mutation that user concerns from TCGA, which match the specific mutation that user concerns of the selected cell line in E2. We proposed this strategy and expected to get a reasonable evaluation of drug effectiveness that could then assist precision medicine.

Pathway enrichment analyses

To better disclose the real effects of drugs on cancer samples, we performed function analysis on the adverse and reversed genes of each perturbation experiment to find the main biological pathways that the compound potentially targets. Additionally, the DBPOM provides special enrichment analysis of 11 key pathways that concern researchers the most, which are closely related to cell proliferation, death and apoptosis. The gene sets of all the biological processes and pathways were collected from GSEA[23], and the statistical significance of the enrichment analyses was obtained based on the hypergeometric test below:

$$P = 1 - \sum_{i=0}^{m-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

Here, N is the number of all genes in the transcriptome, and M is the number of genes in the gene set to be detected. Meanwhile, n is the number of genes intersecting with the adverse/reversed gene set in N, m is the number of adverse/reversed genes in the pathway, and P represents the enrichment significance of adverse/reversed genes on the pathway.

Results

The DBPOM is designed to evaluate the effect of treatment when using each compound on every cancer patient by integrating the gene signatures of different perturbation experiments on cell lines, particularly cancer types or subtypes associated with genes. The database contains three modules, which are the compound effect on cancer samples, patient-centric precision medicine and drug effect comparison of known/unknown drug combination pairs.

Search by compound

This module is compound-centric and contains response information of each compound at different dosages and treatment lengths when applied to all cell lines. As shown in **Figure 2.A**, when a user selects a compound, the response information has two levels. One is the whole cell level. Users can bring up the detail page for a more in-depth understanding and pathway

enrichment analyses of KEGG, GO, PID and REACTOME, which are provided for the purpose of analyzing the functions of reversed and adverse genes. The other stratum is the pathway level. The database provides a panel focusing on the treatment effects of drugs or compounds on the genes involved in cell proliferation, cell death and 11 other key pathways of high attention for clinicians and drug researchers, including KEGG_CELL_CYCLE and apoptosis regulation. The module also offers information about the compound like its IC50 and structure from other databases. Users can choose a cell line in isolation to ascertain the responses of all compounds to this cell line at different dosages and treatment durations. According to our calculation formula, a drug with a higher score is thought to have a lower possibility of being effective on a certain cell line.

We found that several drugs that were FDA approved, and some that are being employed in clinical trials for cancer treatment, have low RADE values in relation to our database, which demonstrates that the score is reasonably successful at evaluating the drug's efficiency. For example, vemurafenib is an FDA-approved drug used for the treatment of metastatic melanoma with a mutation on the BRAF in the valine located in the exon 15 at codon 600[24] When combined with standard-of-care or novel-targeted therapies, it is also reported that vemurafenib is effective on colorectal cancer[25]. Via our database, we discovered that the RADE value of vemurafenib acting on COAD generally ranks extremely highly as well (Figure 2.B). As another example, curcumin is a drug that has been used in clinical trials, having been investigated for the treatment and supportive care of clinical conditions, including proteinuria, breast cancer, multiple myeloma, and non-small-cell lung cancer[26]. The RADE value of curcumin ranks very highly on the whole in our database (Figure 2.C). Moreover, NNC-55-0396 is a molecular compound still in the drug development phase and reportedly could prevent human cancer cell proliferation and induce cancer cell apoptosis as a result of its ability to inhibit the function of T-type Ca²⁺ channels[27]; the experimental results are also in good agreement with the value we proposed (Figure 2.D).

A Compound Detail					trichostatin-a							
Compound Detail					ID: BRD-A19037878 Dosage: 10 µM Time: 24 h Overview Reversed Gene Key Pathway							
		s have been performed on both tumor vs control tissu	es samples (E1) from TCGA and molecular compound treated cell	1.1				Sea	ch:			
ie vs DMSO tre hen a cell line onverselv, we c	eated cell line (E2) fro s is treated with a con consider it having a ro	Im the CMap database. spound, if the effect on a gene is opposite, i.e., one u versed effect on this gene. Otherwise. if a gene is me	p-regulated or 1 in E1 and the other down-regulated or -1 in E2, or trked as $1(E1) \sim 1(E2) - 1(E1) \sim -1(E2)$, $0(E1) \sim -1(E2)$		Show 10 T entries					CSV Excel		
Effects of the second					Genes * cano	Fold change of er samples vs. normal samples	Fo compound treated	ld change of I vs. DMSO trea	ted cell lines			
	a 878 Dosage: 10 µM 1	ime: 24 h			A2M	down -3.623		up 6.288				
	ersed Gene Adverse				AATK	down -3.273		up 23.099				
np to CMap				1 1	ABCA3	down -4.273		up 6.12				
wnload Per Da	ata				ABCA8	down -6.966 up 4.456		up 3.294 Jown -4.464				
					ABC04	up 28.835		own -4.404				
					ABCC3	up 5.83	down -16.095					
chostatin-					ABCG1	down -2.618	up 18.124					
	78 Dosage: 10 µM Tir				A8HD5 down -2.402			up 6.435				
Overview Reversed Gene Adverse Gene Kry Pathway					ABI38P	down -5.494		up 6.919				
cycle	KEGG_CELL	CYCLE	Submit		Showing 1 to 10 of 1,304 entries		Previous 1	2 3 4	5	131 N		
				1 1	Enrichment Result							
	Show 10 entries Search				GO							
	Genes *	Fold change of cancer samples vs. normal samples	Fold change of compound treated vs. DMSO treated cell lines		KEGG							
	BUB1	up_6.269	down18.186		REACTOME			Sea	du .			
	CCNA1	down2.32	up_6.519		Show 10 T entries		249		CSV Excel			
	CCNA2	up_7.421	down60.518			Pathway Name	* p.	value Re	rerse Num	Hit No		
	CCNB1	up_7.833	down5.301		ACTIN_C	YTOSKELETON_ORGANIZATION_AND_BIOGENESIS		'2e-01	7	92		
erse gene	CCN82	up_8.257	down14.45			ACTIN_FILAMENT_BASED_MOVEMENT		4e-01	0	10		
	CCND2	down2.247	up_2.955			ACTIN_FILAMENT_BASED_PROCESS		9e-01	7	102		
	CDC14A	down_+2.036	up_6.756			ACTIN_FILAMENT_BUNDLE_FORMATION		8e-01	0	11		
	CDC25A	up_5.068	down19.629			ACTIN_FILAMENT_ORGANIZATION		8e-01	1	21		
	CDC25C	up_12.66	down14.259			ACTIN_FILAMENT_POLYMERIZATION		7e-01	2	12		
	CDC45	up_8.914	down2.791		AL IIN.	ACTIVATION_OF_IMMUNE_RESPONSE		Be-01	2	13		
	Showing 1 to 10 of 32 entries Previous 1 2 3 4 Next					ACTIVATION_OF_JINK_ACTIVITY		9e-01	0	12		
				1 1		ACTIVATION_OF_MAPK_ACTIVITY		4e-01	3	32		
	Show 10 • entries		Search:		Showing 1 to 10 of 826 entries		Previous 1	2 3	5	83 1		
	Genes *	Fold change of cancer samples vs. normal samples	Fold change of compound treated vs. DMSO treated cell lines	1 1	· ·							
	ANAPC1	0_0	down27.521	1.17								
	ANAPC2	0_0	up_20.511	1 î -	C	Compound C	urcumin					
	ATR	0_0	down18.653		Compound							
	8U818	up_10.52	up_7.693		This module is compound centric and contain	n response information of each compound at different dosages and th ng Detail.	me points on all the cell lines. All th	e reversed genes, a	verse genes and	key pathways I		
verse gene	CCND1	0_0	down41.75		one compound largets will be shown by clicks In Dabawets, "sabTCGA vs. CMap (all matatian specific matation).	η' is abort for mutation pattern matched TCGA vs. CMap, and "sabTC		shart for mutation	pattern matched	FCGA vs. CMap		
	CCND3	0_0	up_19.065	1.1.1	Show 10 w anties October Determined	v				Expo		
	CCNE1	up_9.765	up_2.018		Fiter by Cancer: #Please selects ¥ Fiter by Co	ell Line («Please select» ¥) e.g. petiteramycis, aristinik, 8MS-						
	CCNE2	up_4.113	up_2.844		Filter by Compound Carcumin 536924 Click here to view the compound list							
	CDC27 CDK6	0_0	down15.633		Fitter	former and the second se	Number of 1	and the set				
		0_0	down16.279		Datasets sub7CGA vs. CMap (all mutation)	Center Type Cell Line Compound Dosage T PRAD VCAP outputnin 10 µM 24	4.5	528	DE 1 V	etail Dow		
	Showing 1 to 10 of 28	entries	Previous 1 2 3 Next		subTCGA vs. CMap (specific mutation) TCGA vs. CMap	COAD HT29 ourcumin 10 µM 24 LHC HEPG2 ourcumin 500 rM 6	5 h 1039	654 662	1.246 D	etail Dow		
_					subTCGA vs. CMap (specific mutation) TCGA vs. CMap	PRAD VCAP oursumin 10 µM 2x PRAD VCAP oursumin 10 µM 2x	4h 675	493		etail Dow		
					subTCGA vs. CMap (specific mutation) subTCGA vs. CMap (all mutation)	COAD HCT116 ourcumin 10 µM 6 LHC HEP02 ourcumin 500 nM 6	5 h 1018	673	1.450 D	etal Dow etal Dow etal Dow		
					subTCGA vs. CMap (specific mutation) TCGA vs. CMap	COAD HT115 Curcumin 50 µM 6 COAD HCT116 Curcumin 10 µM 6	5 h 1168	762	1.539	etal Dow		
		Compound Vemur	afenib		subTCSA vs. CMap (all mutation)	COAD HT29 exerumin 10 µM 24	4 h 1268	818	1.618 D	etail Dow		
					Showing 1 to 10 of 412 entries				Trat Fice	NOS I NOS		
ompound	d				D	Compound NN	IC-55-039	6				
		yarne information of each compound at different desages and fame po betail.	airs an all the cell lines. All the reversed genes, adverse gones and key pathways that the		D	Compound NN	NC-55-039	6				
module is compo compound target	ound-centric and contains re ts will be shown by clicking i	yarne information of each compound at different desages and time po brief.			Compound					_		
module is compo compound targets atasets, "subTCG/ (fic mutation).	ound-centric and contains re ts will be shown by clicking i iA vs. CMap (ell mutotion)" i	s short for mutation pattern matched TCGA vs. CMap, and "subTCGA vs	. CMap (specific mutation)" is short for mutation pattern matched TCGA vs. CMap (cell line		Compound This module is compound centric and contain one compound largets will be shown by abbit	s respanse information of each compound at different docages and tis ng Dutali.	ime points on all the cell lines. All th	e reversed genes, o				
module is compo compound targets of overs, "subTCGs (fic mutation). = 10 v entries r by Cancer COAD	ound-costric and contains re ts will be shown by clicking I A vs. CMap (all mutation)" i Choose Datasets [al] D v] Fitter by Cel L	short for mutation pattern matched TCGA vs. CMap, and "subTCGA vs v) to [=Presse select= v]			Compound Dis models is compound centric and contain are compased targets will be shown by albida In Datasets, "subTCA vs. Chap (all motorial specific mutation).		ime points on all the cell lines. All th			TCGA vs. CMag		
module is compo compound target otravets, "subTCGs (fic mutation). = 10 v entries = by Cancer COAD = by Compound v	ound contric and contains re tr will be shown by clicking i A vs. CMap (off mutation)" i Choose Datasets [e] D v Fitter by Cel U Venuratienib a	s short for mutation pattern matched TCGA vs. CMap, and "subTCGA vs	. CMap (specific mutation)" is short for mutation pattern matched TCGA vs. CMap (cell line		Compound The sandake is compound centric and cost loss are composed togeth will be shown by dick! In Delevers, "sabTCA as, to Chap cell meterious specific meterion". Show (3 =) metrics — Docase Dataset (al. The bu Correct ("Please sector + V Herb to C	s regener loberation of each compound at different desages and lin og Desti. (* is short for exclusion patterns motivhed TCGA vs. CPAop, and "subTC vs.)	ime points on all the cell lines. All th	e reversed genes, o		TCGA vs. CMag		
module is compo imposed target laterate, "sub-FGG fic mutation). In 10 entries in by Cancer COAE in by Compound 22 24 is here to view the o	ound contric and contains re tr will be shown by clicking i A vs. CMap (off mutation)" i Choose Datasets [e] D v Fitter by Cel U Venuratienib a	short for mutation pattern matched TCGA vs. CMap, and "subTCGA vs v) to [=Presse select= v]	. CMap (specific mutation)" is short for mutation pattern matched TCGA vs. CMap (cell line		Compound The sandake is compound centric and cost loss are composed togeth will be shown by dick! In Delevers, "sabTCA as, to Chap cell meterious specific meterion". Show (3 =) metrics — Docase Dataset (al. The bu Correct ("Please sector + V Herb to C	s regener loberation of each compound at different desages and lin og Desti. (* is short for exclusion patterns motivhed TCGA vs. CPAop, and "subTC vs.)	ime points on all the cell lines. All th	e reversed genes, o		TCGA vs. CMag		
module is compo compound forget etranstr, "unbTCGs (fic mutation), a (10 v entries by Cancer COAD is by Compound (NA there to view the c er	eand centric and contains re tr will be shown by clicking i lik rs. Chap (all metation)" i Oncose Datasets (all D) (There by Cell U Venuryfanib) (a compound list Datasets	a sharf for material on pattern matched PCGA nr. OHign, and "subTCGA n v (rPrinor millers, v specific service), generation (b), generation generation (b), generation (Ong lyan(in mataling 's short for mataline patients matched TGE n. Ong ind for Engen ())		Compound This residue is responsed where is a statute and the statute of the statute of the statute is character, and a statute of the statute is character, and statute of the statute is character, and the statute of the statute is character, and the statute of the statute is character, and the statute of the statute is character of the statute of the statute is character of the statute of the statute St	n regense Johannation af rech compound at different desayes and to ny terest. ** a short for material partners methoded TCGA vs. (5%p. md "subT ** 0 Line (+19445 vs. vs.)	ime points on all the cell lines. All th	e reversed genes, o		TCGA vs. CMag		
module is compo compound forget otserts, "subTCG file mutation]. = 10 v entries by Concer COAD or by Compound w NA here to view the o er subTCGA vs.	vend-centric and contains re tr will be shown by clicking i dres. Chap (ell metorion)" i Decose Datasets (ell D v) (filter by Cell U Venurshenb is compound ist Datasets c. Chap (ell mutation)	a darf or metation pattern makhed FCGA nr. Offig. and "subTCGA n w (Person metation) p professionarysist anticide. (Eds):	One people matching is about for matchine matchine matchine fields in ONey and for Separation of the second		Compound The sandake is compound centric and cost loss are composed togeth will be shown by dick! In Delevers, "sabTCA as, to Chap cell meterious specific meterion". Show (3 =) metrics — Docase Dataset (al. The bu Correct ("Please sector + V Herb to C	a regular fabricultur of exh compound at different decayes and the optimizer. If a dist for material pattern matched PEGA is, CHop, and "subTC and the (<u>Channe strates</u>) (og getimmergen, mittenb, (bb))	ine points on all the cell lines. All th	e reversed genes, o	pattern matched	TCGA vs. CMag		
module is compo compound target latents, "subTCGs offic motations," is to compound in the compound in the subtract of the subTCGA vs. Cl subTCGA vs. Cl	sund certric and contains re tr will be shown by clicking" i d Choose Dataste (al over the contained of the contained over the contained of the contained were shown to be contained on the contained compound list Datasets to Chap (all mutation) Data peckin mutation) Data voc Chap	- Are for motifier pattern makker (FCA vs. CHay, and "ad-ICCA vs.	Other parallel methods in during multicle parallel methods if IGA as, Other parallel methods in the par		Component Service An address of surgering and address Andress	regener blomdin of esh respond at different designs and by a shark consistent prime and APTCH is Chap, and with a shark consistent prime and an analysis of the share of the share and a share of the share of the share of the share and a share of the share of the share of the share and a share of the share of the share of the share and a share of the share of the share of the share of the share and a share of the share of th	ine points an all the cell flows. All th GGA us. CPApp (pper(fit modallow)* lo Times	e reversed genes, a abort for mutation bornlog of trease to make 164	NDE TY 0.575 D	Espor		
module is compo- compound targets enterster, "unkTGG dif- file mutation). a 10 wents a 10 wents by Compound wents subTGGA vs. Cl subTGGA vs. Cl subTGGA vs. Cl subTGGA vs. Cl subTGGA vs. Cl	avand creatile and contribute re m will be alsona by chicking in Choose Datasets [at Occore Datasets [at] [bite by Cell 10] [bite by Cell 10	- Art for motion patient and half (CGA is, CGA, and "add)(CGA is) IIII	Other (people) metalation? In Acril people Margin (m) Margin (m) Margin (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m) Names (m)		Component Service An address of surgering and address Andress	regarder forwarder of one compared of different designs and the of a low-for mathematical model FEGI on CHips, and "suffic and the compared mathematical for the off of the second sec	ine polets on all the cell lines. All the Cell lines All the Cell lines are the Cell lines and the Cell lines are the Cell line	e reversed genes, o abort for matation borriby of 164 185 312	NDE T V 0.573 D 0.627 D 0.627 D	Espo Internal Down		
models is compo- compound forgets between, "white Code (contrasting), and Code (code), and the code of the code of the code (code), and the code of the code (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code), and (code	avand contrik, and contrikine pr m will be alsona by chicking in Choose Dataments [n]. Choose Dataments [n]. Que (n) [n] [n] [n] [n] [n] (n) [n] [n] [n] [n] [n] (n) [n] [n] [n] [n] [n] [n] (n) [n] [n] [n] [n] [n] [n] [n] [n] (n) [n]	A der for molatile patiers makkel FCGA is. CHay, and "ad-CGA is we characteristic and the second se	Other parallel matching is a lower for matchine matchine distribution of the low of the lower for the low		Compound An advantage of the second second second advantage of the second sec	angenet information of only compared of different designs and for a first sector of the sector of th	Ine points an all the cell lines. All 16 CGA on CPApp (gene)(is mutation) ^{an} is Time Research one and All States 24 h 319 24 h 319 24 h 540 24 h 540 24 h 540	e reversed genes, o abort for matation beening of 164 165 312 351 220	x06 *** 0.575 0 0.627 0 0.721 0 0.728 0 0.912 0	I TCGA ye. CMap Expor Istail Dow Istail Dow Istail Dow Istail Dow		
Interstry, "subTCGA offic methation). In To ", the entries w by Cancer (CoLA w by Cancer (CoLA and there to view the c subTCGA vs. subTCGA vs. subTCGA vs. TCC subTCGA vs. C1 subTCGA vs. C1 SubTCG	numé centrie and centries and centries en and te alcome by chicking in a conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation of the conservation	Open motion partner matched (COLA n. ORg., and "add/COLA Imit dimensioned Terminal and terminaland terminal and terminal and terminal and terminal and	Other process mutation with a contrast and the conttext and the contrast and the contrast and the cont		Compound An advantage of the second second second advantage of the second sec	angenet information of only compared of different designs and for a first sector of the sector of th	Ine pulses an all the cell lines. All th CA is, CMap Epsecific mutatloop" is the second secon	e reversed genes, a abort for mutation beendary of formation 164 185 212 251 251 250	x06	I FCGA vs. CMap Experi Ietai Down Ietai Down Ietai Down Ietai Down Ietai Down		
module is compo- remposed forget interest, "web/COS for moduling", 10 moduling, interest 10 moduling, interest	assist centric and centrics and centrics and the shows by children in the first sector of the shows by children in the sector of the shows by children in the sector of th	A dar for moletile patient matched FCGA es (Mag. and "ad-FCGA es and <u>Charan molecular</u> gradientempiles mission, BHS <u>Color</u> 4729 ensemptiles Datage <u>Tane</u> <u>Color</u> 4729 ensemptiles <u>Datage</u> <u>Color</u> 4749 ensemptiles <u>Color</u> 4749 ensemptile	Other paralle matching that part of matching that parallel matching that we define the second of the seco		Composed An address compared works and particular the destination of the composition of the composition of the compared of the composition of the compared the composition of the composition of the composition the composition of the composition of the composition the composition of the compos	Images before the only one of the composed of different decays and the off the composed of different decays and "walk" Image decays and the composed of different decays and "walk" Image decays and the composed of different decays and "walk" Image decays and the composed of different decays and "walk" Image decays and the composed of different decays and "walk" Image decays and the composed decays and	Ine pulses an all the cell lines. All th CA is, CMap Epsecific mutatloop" is the second secon	e reversed genes, a abort for mutation beendary of formation 164 185 212 251 251 250	NDE [**] 0.375 D 0.750 D 0.750 D 0.750 D 0.952 D	I TCGA ve. C I Ietail Ietail Ietail Ietail Ietail Ietail		

Figure 2. Outline and statistical analysis of Compound module. (A) The compound detail page includes concise information, details of reversed and adverse genes, and pathway enrichment analysis. (B–D) Three responsive tables of users' search results have been employed to prove the rationality of the RADE scores raised by the DBPOM.

Personalized medicine

This is a patient-centric module and contains three options for precision medicine. The first is for an optimal match between a cell line and cancer samples according to the similarity score of their overall mutation patterns. When a cell line is chosen, the module matches the referenced cancer samples with the cell, and performs the aforementioned analyses (**Figure 3.A**). The second example relates to the specific mutation that a user is interested in. When a certain mutation and

a drug are selected, this module compares both the reversed and adverse effects of this drug on cancer patients versus one without the mutation (**Figure 3.B**). For each cancer, we only list the genes which have high genetic mutation rate. Finally, the user can also submit his/her own data regarding a patient's mutation profile for analysis (**Figure 3.C**).

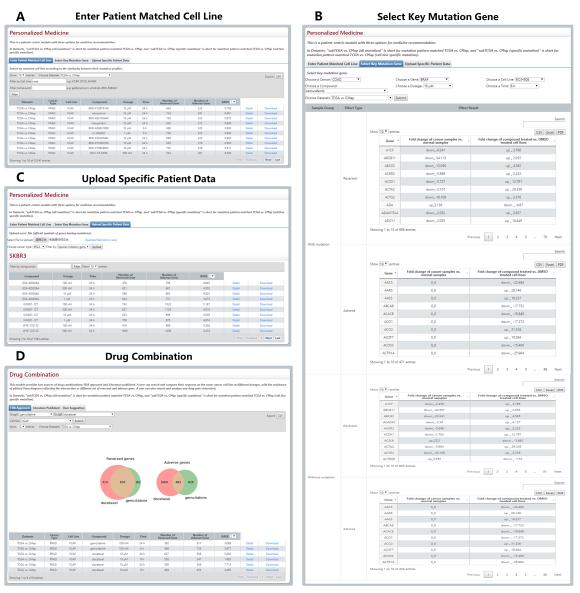


Figure 3. Results from Personalized Medicine and Drug Combination modules. (A–C)The three modules for personalized medicine. (D) The Venn plot and search results of gemcitabine and docetaxel.

Drug combination

This module provides FDA-approved, text-mining-validated drug pairs. For each pair, the user can search and compare their response on the same cancer cell line in a different physical environment with the assistance of a plotted Venn diagram that reflects the intersection or a different set of reversed and adverse genes. Furthermore, they can also search and investigate any drug pairs they are interested in. For instance, gemcitabine and docetaxel are an FDA-approved drug combination used for breast, ovarian, and non-small-cell lung cancer. Their reversed and adverse genes on VCAP cell line and the corresponding Venn plot are shown in **Figure 3.D**.

Easy Download

This module allows easy access to all the reversed and adverse effects of 19,406 small molecular compounds, existing drugs and 509 drug combinations on the 12,310 genes of cancer patients across five cancer types based on 28 cell lines and 3,078 cancer samples. In addition, all of the documents used in the database are available for download, which is also the case with the compounds, key pathways and drug combinations.

Database and web interface service and implementation

The DBPOM website displays succinct information about the gene signatures of compounds and drugs. Detailed information can also be downloaded from the relevant pages. We provide partial/complete downloads of the entire database and its description files. Furthermore, the DBPOM website affords researchers an online analysis capability for compounds, and they can easily download their query or calculation results from any module. The data in the DBPOM is all stored and managed by MySQL, and the web interface is implemented using HTML and JavaScript. In addition, all of the data analysis work is completed with R and Python.

Discussion and future directions

Overall, DBPOM is an easy-to-use database, which offers users the chance to explore the mechanism of the effect of a potential drug on cancer patients via the use of genomics and transcriptomics-based information. Clinicians and drug researchers could search and analyze both the therapeutic and adverse effects of a drug or drug combination on cancer patients according to their needs for personalized drug analysis, which is unlike other similar databases that focus solely on cell lines. Among other advantages, the DBPOM provides a worthwhile way to compare the effect of different drugs through the prism of gene expression change, meaning it could assist in speeding up the process of drug development, facilitate new uses of old drugs, and act as a catalyst for the discovery of new therapies. In addition, the DBPOM is the first large-scale database to study adverse gene expression during drug treatment.

We will continue to collect more data relating to other cancer types, cell lines, compounds, and approved drug combinations to enrich our database. We will also develop and import more computational tools for reversed effects, adverse effects and their comprehensive effect estimation as well as extending the capabilities for precision medicine recommendations. We expect the DBPOM to become an important resource and analysis platform for drug development, drug mechanism studies, and new therapy discoveries.

Acknowledgments

The authors thank all members of the Zhi Li Laboratory at the First Hospital of China Medical University who provided us with critical comments and feedback.

Funding

National Natural Science Foundation of China (61902144, 62002212, 62072212) and 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG07D).

Reference

1. Jameson, J.L. and D.L. Longo, Precision medicine—personalized, problematic, and promising. Obstetrical & gynecological survey, 2015. 70(10): p. 612-614.

2. Ginsburg, G.S. and K.A. Phillips, Precision medicine: from science to value. Health Affairs, 2018. 37(5): p. 694-701.

3. Dietel, M. and C. Sers, Personalized medicine and development of targeted therapies: the upcoming challenge for diagnostic molecular pathology. A review. Virchows Archiv, 2006. 448(6): p. 744-755.

4. Dietel, M., et al., Predictive molecular pathology and its role in targeted cancer therapy: a review focussing on clinical relevance. Cancer gene therapy, 2013. 20(4): p. 211-221.

5. Weinshilboum, R.M. and L. Wang. Pharmacogenomics: precision medicine and drug response. in Mayo Clinic Proceedings. 2017. Elsevier.

6. Monks, A., et al., Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. JNCI: Journal of the National Cancer Institute, 1991. 83(11): p. 757-766.

7. Chen, B., et al., Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature communications, 2017. 8(1): p. 1-12.

8. Jing, Y., et al., Multi-omics prediction of immune-related adverse events during checkpoint immunotherapy. Nature communications, 2020. 11(1): p. 1-7.

9. Nguyen, D.A., C.H. Nguyen, and H. Mamitsuka, A survey on adverse drug reaction studies: data, tasks and machine learning methods. Briefings in Bioinformatics, 2019.

10. Vilar, S. and G. Hripcsak, The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug–drug interactions. Briefings in Bioinformatics, 2017. 18(4): p. 670-681.

11. Wu, Z., et al., Single-Cell Techniques and Deep Learning in Predicting Drug Response. Trends in Pharmacological Sciences, 2020.

12. Weinstein, J.N., et al., The cancer genome atlas pan-cancer analysis project. Nature genetics, 2013. 45(10): p. 1113.

13. Edgar, R., M. Domrachev, and A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research, 2002. 30(1): p. 207-210.

14. Barretina, J., et al., The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012. 483(7391): p. 603-607.

15. Subramanian, A., et al., A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. Cell, 2017. 171(6): p. 1437-1452. e17.

16. Yang, W., et al., Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res, 2013. 41(Database issue): p. D955-61.

17. Yoo, M., et al., DSigDB: drug signatures database for gene set analysis. Bioinformatics, 2015. 31(18): p. 3069-3071.

18. San Lucas, F.A., et al., Cancer in silico drug discovery: a systems biology tool for identifying candidate drugs to target specific molecular tumor subtypes. Molecular cancer therapeutics, 2014. 13(12): p. 3230-3240.

19. Wakaskar, R.R., Challenges pertaining to adverse effects of drugs. International Journal of Drug Development and Research, 2017. 9(3): p. 1-2.

20. Cheng, F., I.A. Kovács, and A.-L. Barabási, Network-based prediction of drug combinations. Nature communications, 2019. 10(1): p. 1-11.

21. Love, M.I., W. Huber, and S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 2014. 15(12): p. 550.

22. Zhao, W., et al., Large-scale Characterization of Drug Responses of Clinically Relevant Proteins in Cancer Cell Lines. Cancer Cell, 2020.

23. Subramanian, A., et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 2005. 102(43): p. 15545-15550.

24. Luke, J.J. and F.S. Hodi, Vemurafenib and BRAF inhibition: a new class of treatment for metastatic melanoma. Clinical Cancer Research, 2012. 18(1): p. 9-14.

25. Yang, H., et al., Antitumor activity of BRAF inhibitor vemurafenib in preclinical models of BRAF-mutant colorectal cancer. Cancer research, 2012. 72(3): p. 779-789.

26. Gupta, S.C., S. Patchva, and B.B. Aggarwal, Therapeutic roles of curcumin: lessons learned from clinical trials. The AAPS journal, 2013. 15(1): p. 195-218.

27. Huang, W., et al., T-type calcium channel antagonists, mibefradil and NNC-55-0396 inhibit cell proliferation and induce cell apoptosis in leukemia cell lines. Journal of Experimental & Clinical Cancer Research, 2015. 34(1): p. 54.