D1184–D1199 *Nucleic Acids Research, 2022, Vol. 50, Database issue* *Published online 27 September 2021*
https://doi.org/10.1093/nar/gkab860

CTR-DB, an omnibus for patient-derived gene expression signatures correlated with cancer drug response

Zhongyang Liu[1,2,*,†], Jiale Liu[1,†], Xinyue Liu[1,†], Xun Wang[1], Qiaosheng Xie[3], Xinlei Zhang[4], Xiangya Kong[4], Mengqi He[1], Yuting Yang[5], Xinru Deng[1], Lele Yang[2], Yaning Qi[2], Jiajun Li[2], Yuan Liu[1], Liying Yuan[2], Lihong Diao[1], Fuchu He[1,*] and Dong Li[1,2,*]

[1]State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China, [2]College of Chemistry and Environmental Science, Hebei University, Baoding 071002, China, [3]Department of Radiation Oncology, China-Japan Friendship Hospital, Beijing 100029, China, [4]Beijing Geneworks Technology Co., Ltd., Beijing 100101, China and [5]Department of Immunology, Medical College of Qingdao University, Qingdao 266071, China

Received August 15, 2021; Revised September 08, 2021; Editorial Decision September 10, 2021; Accepted September 15, 2021

ABSTRACT

To date, only some cancer patients can benefit from chemotherapy and targeted therapy. Drug resistance continues to be a major and challenging problem facing current cancer research. Rapidly accumulated patient-derived clinical transcriptomic data with cancer drug response bring opportunities for exploring molecular determinants of drug response, but meanwhile pose challenges for data management, integration, and reuse. Here we present the Cancer Treatment Response gene signature DataBase (CTR-DB, http://ctrdb.ncpsb.org.cn/), a unique database for basic and clinical researchers to access, integrate, and reuse clinical transcriptomes with cancer drug response. CTR-DB has collected and uniformly reprocessed 83 patient-derived pre-treatment transcriptomic source datasets with manually curated cancer drug response information, involving 28 histological cancer types, 123 drugs, and 5139 patient samples. These data are browsable, searchable, and downloadable. Moreover, CTR-DB supports single-dataset exploration (including differential gene expression, receiver operating characteristic curve, functional enrichment, sensitizing drug search, and tumor microenvironment analyses), and multiple-dataset combination and comparison, as well as biomarker validation function, which provide insights into the drug resistance mechanism, predictive biomarker discovery and validation, drug combination, and resistance mechanism heterogeneity.

INTRODUCTION

Because of cancer heterogeneity, at present both for chemotherapy and targeted therapy, the treatment response rate of patients is still far below 100%. For example, a meta-analysis of phase II single-agent clinical studies (570 studies; 32 149 patients) has shown that the median response rate of chemotherapy is only 11.9%, and even for personalized targeted therapy, this rate is only 30% (1). Cancer drug resistance continues to be a major and challenging problem facing current cancer research (2). The key to solve this problem is to understand the underlying drug resistance mechanism, to identify predictive biomarkers for precise patient stratification, and even to develop combinational drugs for overcoming the drug resistance.

In the era of precision medicine, rapidly accumulated patient-derived clinical transcriptomic data with cancer drug response bring opportunities for solving the problem, but meanwhile propose computational needs for the data management, integration, and (re-) use. Previous studies have proved that patient-derived clinical transcriptomes with therapy response can help reveal drug resistance mechanism (3,4), and baseline (i.e. pre-treatment) transcriptomic signals of patients are promising biomarker candidates for predicting drug response (5,6). However, these data are always scattered and often mixed with various other types of

*To whom correspondence should be addressed. Tel: +86 106 177 7056; Fax: +86 106 177 7004; Email: liuzy1984@163.com
Correspondence may also be addressed to Dong Li. Tel: +86 106 177 7057; Fax: +86 106 177 7004; Email: lidong.bprc@foxmail.com
Correspondence may also be addressed to Fuchu He. Tel: +86 106 177 1001; Fax: +86 106 177 7004; Email: hefc@nic.bmi.ac.cn
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

data (such as cell line level data), and meanwhile they often use different terminologies, use inconsistent data processing pipelines, have non-uniform data formats, and usually have poor machine-readable metadata, which greatly hinders their access, integration, and reuse.

Great efforts have been devoted to integrating and reusing the transcriptomic data with cancer drug response, such as CellMinerCDB (7), GDA (8), Xeva (9), Borisov *et al*.'s work (10), ROCplot.org (11–13) and CDRgator (14). Both CellMinerCDB and GDA aim to integrate molecular and drug sensitivity data of cancer cell lines. Mer *et al*. developed Xeva, an open source software package, for integrative analysis of pharmacogenomics data of patient-derived xenografts (PDX) (9). Borisov *et al.* compiled a list of 26 clinical transcriptomic datasets with drug response, involving nine drugs and eight cancer types (available in the supplementary file of their paper) (10). ROCplot.org is a bioinformatics analysis tool used to validate candidate predictive biomarkers using patient-derived microarray data of four cancer types. CDRgator provides a resource of cancer drug resistance signatures (i.e. differentially expressed genes) extracted from transcriptomes of cancer cell lines and melanoma patient samples. All these resources have contributed greatly to the cancer drug resistance research, however, a database devoted to collecting the abundant and valuable patient-derived clinical transcriptomes with drug response covering multiple cancer types is still lacking.

Here, we present the Cancer Treatment Response gene signature DataBase (CTR-DB), which is a web-based, user-friendly, and interactive database, specially designed to comprehensively collect and uniformly reprocess patient-derived clinical transcriptomic data with cancer drug response information, and meanwhile to provide various data analysis functions facilitating the integration and reuse of these data. CTR-DB has collected 83 patient-derived baseline microarray or RNA-seq source datasets with manually curated cancer drug response information, involving 28 histological cancer types, 275 therapeutic regimens, 123 drugs (covering chemotherapy, targeted therapy, and immunotherapy), and 5139 patient samples. Moreover, CTR-DB supports single-dataset exploration (including differential gene expression analysis, receiver operating characteristic curve analysis, functional enrichment analysis, sensitizing drug discovery, and tumor microenvironment analysis), multiple-dataset combination and comparison, and the function of biomarker validation. In 'Results' section we introduced these functions in detail. Besides, we also provided a use case (in 'Use case' part) to demonstrate the usage and value of all these functions, aided by CTR-DB datasets related to anti-*PD1*/*PD-L1* therapy resistance.

MATERIALS AND METHODS

Data collection and curation

We collected and curated patient-derived clinical transcriptomic datasets with cancer drug treatment response from GEO (until 20201014) (15), ArrayExpress (until 20201214) (16), and TCGA (17).

Specifically, for GEO and ArrayExpress, firstly the potentially related datasets were identified by retrieval with the cancer-related keywords such as 'cancer', 'carcinoma', 'tumor', 'neoplasm' or 'malignancy', the drug-related keywords such as 'treatment', 'therapy' or 'drug', and the patient sample-related keywords such as 'clinical', 'patient' or 'sample' as well as by further filtering with '*Homo sapiens*'. Then among ∼15 000 potentially related datasets, according to the inclusion criteria described below, we manually collected qualified datasets. For the qualified dataset, we downloaded the expression data directly from the GEO/ArrayExpress website for the microarray data and from SRA (18) for the RNA-seq data, and manually curated the corresponding metadata from GEO/ArrayExpress records or the original references. The collected metadata mainly included sample ID, cancer subtype (of the minimum granularity), childhood cancer or not, therapeutic regimen, drug response status annotated by original authors, response status definitions (if necessary), data type, platform, source etc.

For TCGA data, firstly we downloaded the TCGA patient clinical information with the help of 'TCGAbiolinks' R package (19). Then according to the inclusion criteria described below, we manually picked qualified patients together with the corresponding drug response information, mainly based on 'days_to_drug_therapy_start', 'days_to_drug_therapy_end', 'days_to_sample_procurement', 'drug_name' and 'measure_of_response' fields. We manually recorded metadata of each qualified patient, with data fields the same as those of GEO/ArrayExpress data above. We downloaded expression profiles (count data) of qualified patients from UCSC Xena browser (20). For a patient with multiple sample expression profiles, the expression profile with sample ID containing '01A' was preferred.

Dataset inclusion criteria

1) Baseline (i.e. pre-treatment) expression profiles.
2) For GEO/ArrayExpress, source datasets with less than 10 samples were excluded.
3) In order to achieve the uniform expression data reprocessing, for GEO and ArrayExpress, only datasets providing CEL files (for microarray data) or FASTQ files (for RNA-seq data) were collected.
4) For microarray data, we only collected datasets produced from GPL96 [HG-U133A], GPL570 [HG-U133_Plus_2] and GPL571 [HG-U133A_2] platforms, because these three platforms are widely used and use the same probes to measure the same genes (11).
5) For TCGA samples, we only considered the first-round drug usage after the sample was obtained, together with its corresponding drug response. For the usage of multiple drugs in the same time period, we recorded them as a drug combination. And we deleted samples using combinational drugs with inconsistent drug response information.

Overall, CTR-DB has collected 83 patient-derived baseline microarray or RNA-seq source datasets with cancer drug response information (for TCGA, a TCGA Project, such as 'TCGA-BRCA', corresponds to a source dataset), involving 5139 patient samples.

Unified gene expression data reprocessing

To facilitate data integration and reuse, for each collected source dataset, gene expression data were reprocessed by a unified pipeline established by us, starting from CEL files for microarray data and FASTQ files for RNA-seq data.

Microarray data processing pipeline. Starting from raw CEL files, microarray data preprocessing was implemented with the help of the 'rma' function of the 'affy' R package (21). And the processing steps mainly included background correction, normalization, pm correction, summary expression value computation. After 'rma' processing, the expression data have already been log2 transformed. The probes were converted into gene symbols according to the platform-specific probe annotation file using the functions 'annPkgName' and 'aafSymbol' of the 'affy' R package (21). When multiple probes were mapped to a single gene symbol, the maximum expression value was used.

GEO/ArrayExpress RNA-seq data processing pipeline. All RNA-seq data were obtained in FASTQ format. FASTQ files were first processed through Trim Galore! (version: 0.6.6) (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), which is a Perl wrapper around the two tools FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Cutadapt (22), mainly used to trim off low-quality bases, then find and remove adapter sequences from the 3′ end of reads, and remove reads with sequence length shorter than 20 bp. Then STAR (version: 2.7.6a) (2-pass mode) was used for alignment to generate BAM files (23). After that, we used HTSeq (version: 0.12.4) to count the reads mapped to each gene (24). Read counts were normalized using the 'Fragments per Kilobase of transcript per Million mapped reads' (FPKM) method.

TCGA RNA-seq data processing pipeline. Because GEO/ArrayExpress RNA-seq data preprocessing described above used the pipeline consistent with that of TCGA (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/), we directly downloaded $\log_2$(count + 1) data provided by UCSC Xena (20) (https://xenabrowser.net/datapages/?hub=https://gdc.xenahubs.net:443). These data were reverted into original counts. Read count normalization used the same way as stated above.

Batch effect removal. For microarray data, batch effect removal was done by the 'ComBat' function of the 'sva' R package (25); For RNA-seq data, by the 'ComBat_seq' function of the 'sva' R package (26).

'CTR-DB dataset' definition

The processed source datasets were further divided into 'CTR-DB datasets', each of which was composed of samples with the same therapeutic regimen and cancer subtype (of the minimum granularity). Overall, 83 source datasets were divided into 626 CTR-DB datasets. Further, a CTR-DB dataset was subdivided into subsets, each of which was composed of samples with the same original drug response state annotated by original authors (e.g. complete response or stable disease). These CTR-DB sub-datasets were prepared to be used for the 'Combine' function of CTR-DB.

Predefined response and non-response grouping

For each CTR-DB dataset, we divided samples into response and non-response groups, so that we can perform various pre-analyses on the dataset under this predefined grouping.

Generally, if original authors annotated the response status of samples with four groups (complete response, CR; partial response, PR; stable disease, SD; progressive disease, PD), we divided the samples with CR and PR into the response group, and those with SD and PD into the non-response group. In some cases, original authors classified the samples into three or five groups, and then we further divided them into response and non-response groups using a criterion imitating the 4-group classification principle stated above. If the source reference annotated the samples with only two drug response groups, our grouping was consistent with the original reference. The grouping standard for each CTR-DB dataset together with the corresponding definitions of the original response states curated from the original reference is provided on the detailed annotation page of the CTR-DB dataset. Users can change the response/non-response grouping standard by themselves to re-analyze the CTR-DB datasets, by 'Combine' function of CTR-DB.

Terminology standardization

Drugs were harmonized by the IDs and names of DrugBank (27), ChEMBL (28) and PubChem (29). Cancers were manually harmonized based on the Disease Ontology (version: March 2021) (30) and the WHO Classification of Tumours Online (https://tumourclassification.iarc.who.int/) (accessed from January to April 2021). Genes were harmonized by the Entrez Gene IDs (31) and HGNC gene symbols (32).

Annotation information integration

In CTR-DB, we also integrated various annotations for drugs, cancers and genes from external databases. For drugs, target annotations were from DrugBank (version: 20201004), and drug types (chemotherapy, targeted therapy and immunotherapy) were from TCGA clinical information annotations or were manually curated. For cancers, annotations were from Disease Ontology (version: March 2021). For genes, gene set annotations (including KEGG pathway (33), Reactome pathway (34), WikiPathways (35), hallmark gene set (36), microRNA target and transcription factor target) were from the Molecular Signatures Database (MSigDB, v7.4) (36). Gene functional categories (including G protein-coupled receptor, GPCR; transcription factor, TF; kinase; ion channel; transporter; nuclear hormone receptor; and whether the gene contains signal peptides or transmembrane regions), Enzyme Commission (EC) number, subcellular location, interacting genes etc. were from POPPIT (http://poppit.ncpsb.org.cn/). Data on drugs/compounds that can inhibit the target genes were from DrugBank (version: 20201004) and DGIdb (version: 2021-January) (37).

Differential gene expression analysis

In order to reveal drug resistance-related molecules and discover candidate predictive biomarkers for drug response, CTR-DB supports differential gene expression analysis and receiver operating characteristic (ROC) curve analysis (38) between non-responders and responders. The obtained differentially expressed genes (DEG) constitute the drug resistance signature. The differential gene expression analysis was implemented by 'limma' R package (39) for microarray data and by 'DESeq2' R package (40) for RNA-seq data. The ROC AUC was calculated by 'pROC' R package (41) and we used one sample *t* test to examine its statistic difference from 0.5. Adjusted *P*-values were computed based on Benjamini-Hochberg (BH) multiple testing correction method (42).

Functional enrichment analysis

To describe the functional characteristics of a drug resistance signature, CTR-DB supports over-representation analysis (ORA) and gene set enrichment analysis (GSEA) (43). Here both ORA and GSEA support six classes of gene sets, including 186 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, 50 hallmark gene sets, 1604 Reactome pathways, 615 WikiPathways, target sets for 2598 microRNAs, and target sets for 1133 transcription factors from MSigDB (v7.4). ORA was performed based on statistically significantly DEGs (satisfying some |log FC| and adjusted *P*-value cutoffs). GSEA was performed based on all genes ranked by log FCs. In ORA, enrichment ratio was computed as the ratio of the proportion of gene set member genes among significantly DEGs to that among the whole genome, measuring the enrichment degree of a gene set. In GSEA, the enrichment score (ES) measures the degree to which a gene set is over-represented at the top or bottom of the ranked genes, and the normalized enrichment score (NES) further considers differences of gene set size (43). Here ORA and GSEA as well as the result visualization were implemented with the help of 'clusterProfiler' R package (44). Adjusted *P*-values were obtained based on BH multiple testing correction method.

L1000CDS^2 analysis

L1000CDS^2 analysis was used to search candidate drugs that can reverse the drug resistance signature. This function was implemented with the help of the API of the L1000CDS^2 search engine (45). When up-regulated and down-regulated genes are submitted, the search engine compares them to the differentially expressed genes computed from the LINCS L1000 small-molecule disturbance profiles (on a certain cell line, with a certain drug dose and sampling time) (46), and the top 50 matched opposite drug signatures are returned.

Tumor microenvironment analysis

The purpose of this analysis is to explore tumor microenvironment (TME) factors correlated to the drug resistance. Firstly, based on the sample gene expression profile, we used 'Estimation of STromal and Immune cells in MAlignant Tumours using Expression data' (ESTIMATE) method to infer the fraction of stromal and immune cells in a patient sample as well as tumor purity (measured by the 'ESTIMATE score', a normalized version of the 'TumorPurity', with [0, 1] range) with the help of 'estimate' R package (47), and used Microenvironment Cell Populations-counter (MCP-counter) method to predict the abundance of 10 cell populations, including $CD3^+$ T cells, $CD8^+$ T cells, cytotoxic lymphocytes, natural killer (NK) cells, B lymphocytes, cells originating from monocytes (monocytic lineage), myeloid dendritic cells, neutrophils, as well as endothelial cells and fibroblasts, with the help of 'MCPcounter' R package (48). Then we used the two-sided *t* test and ROC curve analysis to analyze the difference and search TME indexes that can discriminate between non-responders and responders. The adjusted *P*-value was computed by BH multiple testing correction method.

Meta-analysis

We used sumz method in the 'metap' R package (https://CRAN.R-project.org/package=metap) to integrate *P*-values of individual CTR-DB datasets into a meta-*P*-value, and used BH method to adjust the *P*-value. Items with small meta-*P*-values have strong and consistent associations with drug resistance across CTR-DB datasets (49).

Database implementation

The bottom of CTR-DB was a MongoDB database (https://docs.mongodb.com/). Above this database, the analysis application was written in R and Python. The web presentation application was implemented in Vue.js (https://vuejs.org/). The doughnut chart on the homepage was drawn using Apache ECharts (https://echarts.apache.org/en/index.html), the volcano plot and barplot using 'ggplot2' R package (50), the heatmap using 'ComplexHeatmap' R package (51), the boxplot using 'ggpubr' R package (https://CRAN.R-project.org/package=ggpubr), and the ROC curve plot using 'pROC' R package.

RESULTS

Overview of CTR-DB

CTR-DB is a user-friendly, interactive, and comprehensive database for patient-derived gene expression signatures correlated with cancer drug response (Figure 1). The core patient-derived clinical transcriptomic datasets and the corresponding cancer drug response information were manually collected and curated from GEO, ArrayExpress and TCGA. All transcriptomic data were re-processed using the uniform data processing pipeline, starting from raw CEL files for microarray data and FASTQ files for RNA-seq data. The terminologies of drugs, cancers and genes were harmonized, respectively. The source datasets were further divided into CTR-DB datasets, each of which was composed of samples with the same therapeutic regimen and cancer subtype (of the minimum granularity). For each CTR-DB dataset, we performed the uniform data analysis under a predefined responder and non-responder grouping,

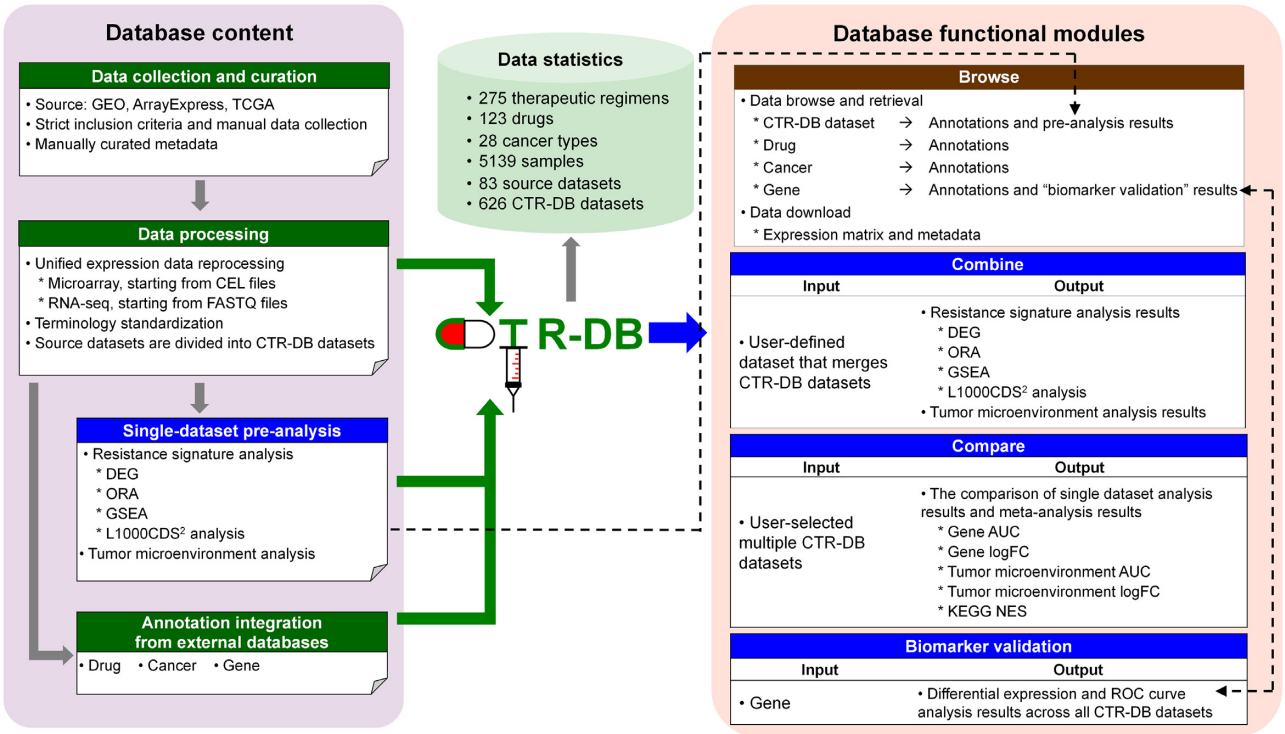

Figure 1. Overview of CTR-DB. CTR-DB is a comprehensive database, designed to collect patient-derived clinical transcriptomes with cancer drug response and meanwhile to provide various analysis functions to facilitate data integration and (re-) use. 'Combine' function is designed for CTR-DB dataset combination analysis, and 'Compare' aims for multiple-dataset comparison.

and the pre-analysis results can be browsed by the detailed annotation page of the corresponding CTR-DB dataset. The analysis results mainly include differential gene expression analysis, receiver operating characteristic (ROC) curve analysis, over-representation analysis (ORA), gene set enrichment analysis (GSEA), L1000CDS^2 analysis, and tumor microenvironment (TME) analysis, aiming to reveal the drug resistance mechanism and to discover candidate predictive biomarkers and candidate combinational drugs that can overcome the drug resistance. In addition, in CTR-DB we also integrated cancer/drug/gene-related various annotations from external databases.

The main functional modules of CTR-DB include 'Browse', 'Combine', 'Compare' and 'Biomarker validation' (Figure 1). 'Browse' supports basic database browse, retrieve and download, by which the detailed annotation pages of each CTR-DB dataset and each gene can be accessed, and CTR-DB datasets can also be downloaded. Other modules are designed to facilitate the data integration and reuse. 'Combine' supports user-customized CTR-DB dataset combination analysis, typically used for the analysis for a drug class or a cancer type of a coarser granularity. 'Compare' facilitates the comparison of analysis results of multiple datasets and implements the meta-analysis across CTR-DB datasets. Finally, by 'Biomarker validation' users can validate the interested candidate predictive biomarkers using transcriptomes of CTR-DB patient cohorts, which cover various drugs and cancer types. All analysis results are visually presented in the form of heatmap, ROC curve plot, volcano plot, barplot, boxplot etc. and various result tables, and all of them are downloadable.

Data statistics of CTR-DB

CTR-DB has comprehensively collected 83 patient-derived baseline transcriptomic datasets with manually curated cancer drug response information. These source datasets were divided into 626 CTR-DB datasets (Figure 2A), each of which was composed of samples with the same therapeutic regimen and cancer subtype. These CTR-DB datasets involve 28 histological cancer types, 275 therapeutic regimens (Figure 2B), 123 drugs covering chemotherapy, targeted therapy and immunotherapy (Figure 2C), and 5139 patient samples (Figure 2D). In our data, breast cancer has the largest sample size (Figure 2D), and breast cancer, skin cancer, stomach cancer, lung cancer, and colorectal cancer have relatively more therapeutic regimens and drugs (Figure 2E). From Figure 2F, we see that most CTR-DB datasets have a relatively small number of samples. These CTR-DB datasets with small sample size are mainly from TCGA. In TCGA data, among patients with the same cancer subtype (of minimum granularity), on average, only three patients shared the same therapeutic regimen, while 29 and 24 for GEO-derived and ArrayExpress-derived data, respectively. There are 76 CTR-DB datasets that have at least five responders and five non-responders with total sample size greater than 10, covering 3628 samples, 17 histological cancer types, 58 therapeutic regimens, and 48 drugs. However, the CTR-DB datasets with small sample size are also valuable, such

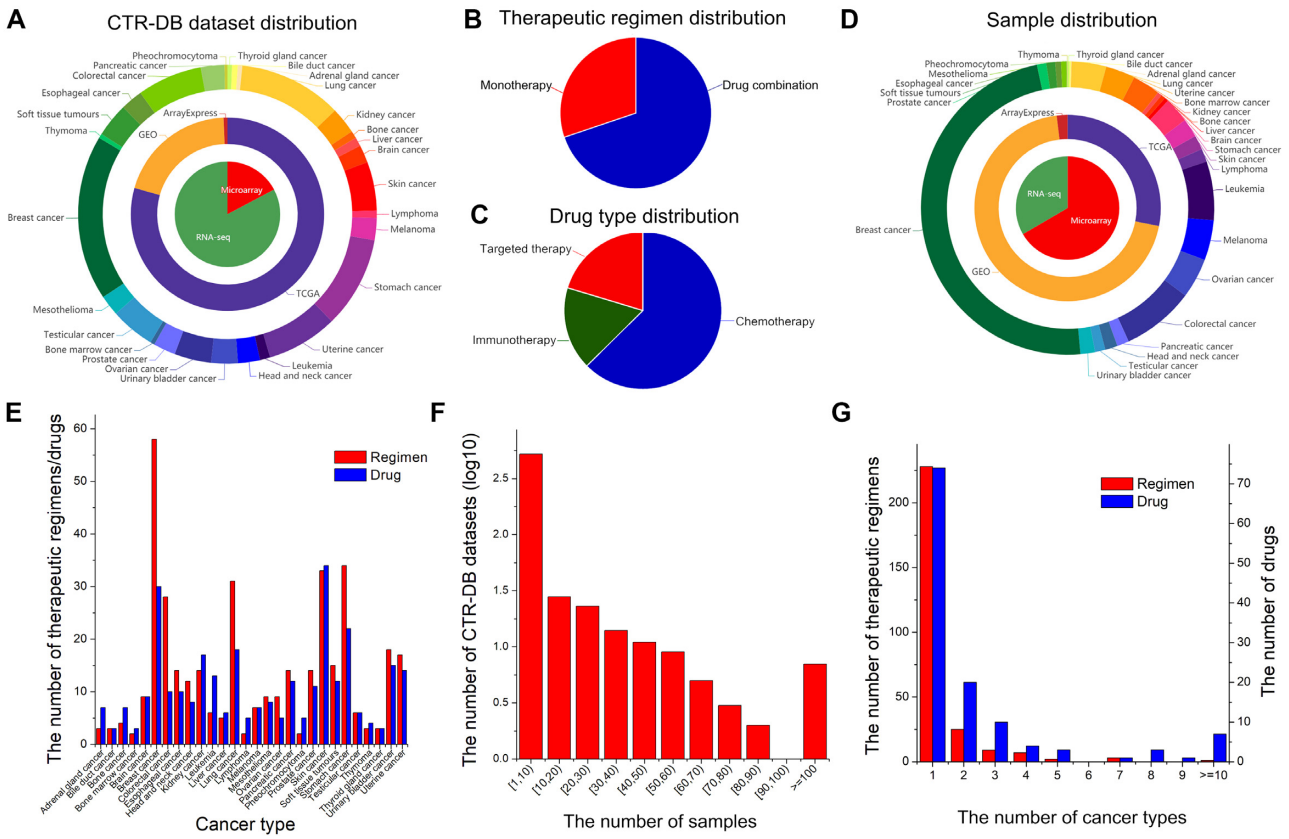

Figure 2. Data statistics of CTR-DB. (**A**) CTR-DB dataset distributions across data types (the inner ring), sources (the middle ring), and histological cancer types (the outer ring). (**B**) Therapeutic regimen distribution on monotherapy and drug combination. (**C**) Drug type distribution on chemotherapy, targeted therapy, and immunotherapy. (**D**) Sample distributions across data types (the inner ring), sources (the middle ring), and cancer types (the outer ring). (**E**) The number of therapeutic regimens/drugs of different cancer types. (**F**) The sample size statistics of CTR-DB datasets. (**G**) The number statistics of histological cancer types treated by therapeutic regimens/drugs.

as used to be merged with other datasets into a larger combined dataset for a certain cancer (sub-) type or a certain drug class of a coarser particle size. Finally, in our data we also see that most drugs/therapeutic regimens are applied to only one histological cancer type, and drugs/therapeutic regimens used in multiple cancer types are in the minority (Figure 2G).

CTR-DB browse

By the 'Browse' of CTR-DB, users can browse and retrieve CTR-DB datasets, drugs, cancers, and genes, and further access the detailed annotation pages of each CTR-DB dataset and each gene. Moreover CTR-DB datasets can also be downloaded here (Figure 3).

On the dataset browse page (Figure 3A), all CTR-DB datasets together with their simplified metadata are shown in the table on the right. Users can use dataset filtering function on the left to identify the interested ones, according to drug type (chemotherapy, targeted therapy, and immunotherapy), drug name, cancer type, dataset sample size, and data type (microarray and RNA-seq). Clicking on the CTR-DB ID in the table will lead to the detailed annotation page of the CTR-DB dataset (Figure 3B), on which besides the detailed annotations about the dataset (Figure 3B(i)), the pre-analysis results of the CTR-DB dataset under the default responder and non-responder grouping will also be presented (Figure 3B(ii)) (see 'Single CTR-DB dataset exploration' section below). The detailed annotations for a CTR-DB dataset mainly include cancer subtype, pediatric cancer or not, drug and drug annotations, source, data type, platform, sample number, responder number, non-responder number, the default response/non-response grouping standard etc. In the dataset browse table, users can select one or multiple CTR-DB dataset(s) and then use the 'Download selected files' button to download the uniformly reprocessed expression matrix ($\log_2$ expression value for microarray data and count for RNA-seq data) and clinical information files, or download all CTR-DB datasets in the table by the 'Download All' button (Figure 3A).

On the drug browse page (Figure 3C), drugs are classified into three types (chemotherapy, targeted therapy, and immunotherapy); and on the cancer browse page (Figure 3D), cancer types and subtypes are organized into a hierarchy tree of six levels. Drugs and cancer (sub-) types can be searched by name. Once a drug or a cancer (sub-) type is selected, its related annotations will be presented in the report table on the right, including drug type, cross-references, targets as well as the number of CTR-DB datasets related to the drug or the cancer type etc. Clicking on the 'Dataset

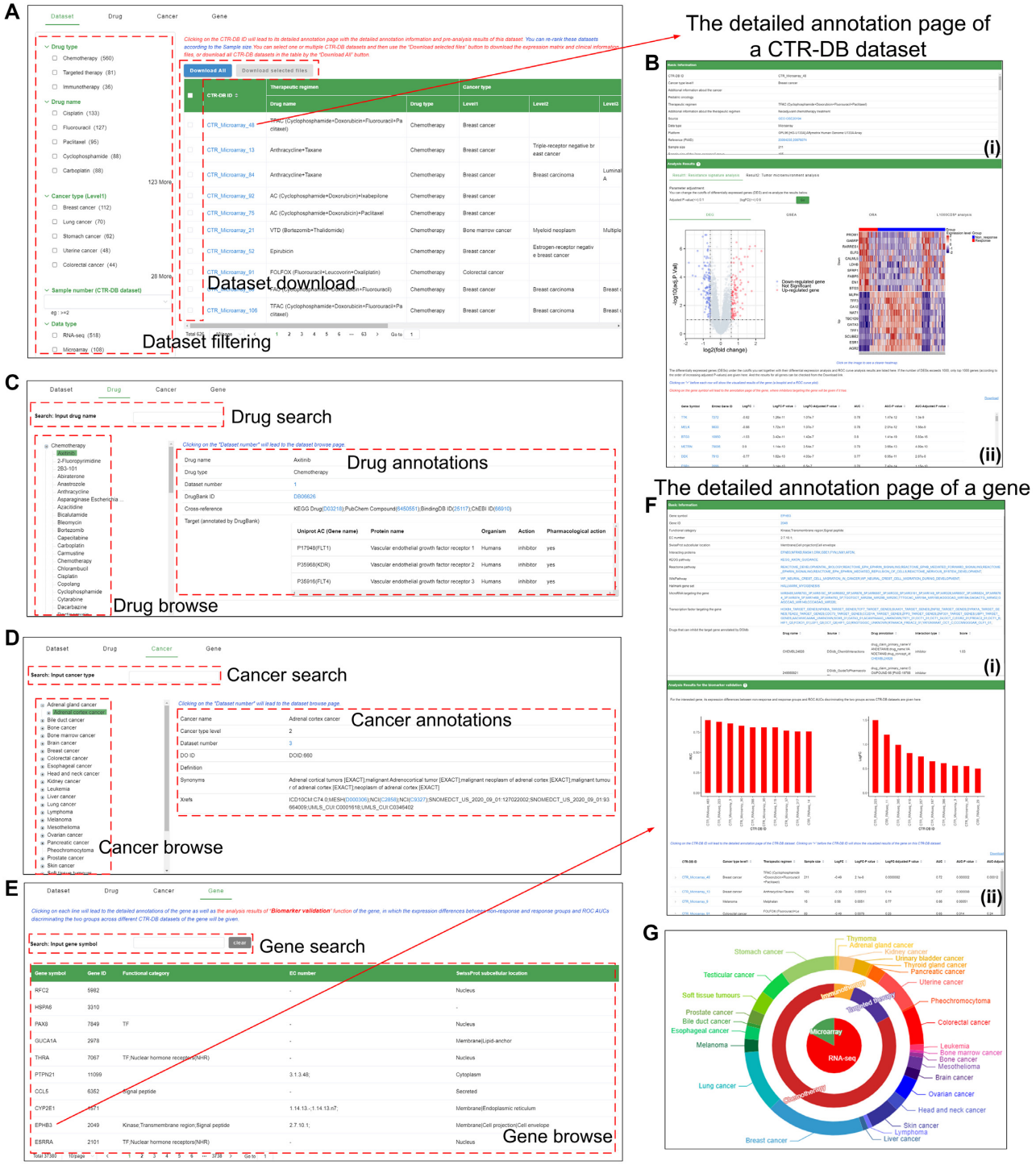

Figure 3. 'Browse' module of CTR-DB. (**A**) CTR-DB dataset browse page. Clicking on the CTR-DB dataset ID will lead to the detailed annotation page of the dataset. (**B**) The detailed annotation page of a CTR-DB dataset, on which besides the detailed annotations (i), the pre-analysis results of the dataset (ii) will also be presented. (**C**) Drug browse page. (**D**) Cancer browse page. (**E**) Gene browse page. Clicking on the interested gene line in the gene browse table will lead to the detailed annotation page of the gene. (**F**) The detailed annotation page of a gene, on which the detailed gene annotations (i) and the 'biomarker validation' results of the gene (ii) are shown. (**G**) The doughnut of data statistics on the homepage, by which CTR-DB datasets can also be browsed through clicking on each sector on the graph.

number' in the table will lead to the dataset browse page presenting the related CTR-DB datasets.

On the gene browse page (Figure 3E), genes can be searched by gene symbol. Clicking on the interested gene line in the gene browse table will lead to the detailed annotation page of the gene (Figure 3F), on which we will give the detailed annotations of the gene (Figure 3F(i)) as well as the analysis results of 'Biomarker validation' function of the gene (Figure 3F(ii)) (see 'Biomarker validation' section below). The gene annotations mainly include functional categories (including TF, GPCR, kinase, transporter, ion channel, whether the gene contains signal peptides or transmembrane regions etc.), subcellular location, interacting proteins, gene set annotations (including KEGG pathway, Reactome pathway, hallmark gene set, WikiPathways, microRNA target and TF target) etc., as well as drugs/compounds that can inhibit the gene.

In addition, users can also browse CTR-DB datasets through the doughnut of data statistics on the CTR-DB homepage (Figure 3G). The outer ring of the doughnut shows histological cancer types, the middle ring shows drug types, and the inner ring shows data types. Sectorial area is proportional to the number of CTR-DB datasets related to the item. Clicking on each area will lead to the dataset browse page presenting the related CTR-DB datasets.

Single CTR-DB dataset exploration

For each CTR-DB dataset, we performed the uniform data analysis under a predefined responder and non-responder grouping, and the analysis results can be browsed on the detailed annotation page of the corresponding CTR-DB dataset. The analyses include two sections: resistance signature analysis and tumor microenvironment analysis. All result figures and tables can be downloaded.

Resistance signature analysis. The drug resistance signature is referred to as the significantly differentially expressed genes (DEG) between non-responders and responders in a dataset. The resistance signature analysis, including 'DEG', 'GSEA', 'ORA' and 'L1000CDS^2 analysis', may help reveal drug resistance mechanism, discover predictive biomarkers, and even discover sensitizing drugs. The significantly DEGs are defined as those with the absolute values of '$\log_2$(fold change)' (i.e. |log FC|) larger than a certain cutoff and meanwhile the adjusted *P*-values smaller than a certain cutoff. Users can change the cutoffs on the resistance signature analysis result page (Figure 4A). Once the cutoffs are updated, the DEG, ORA and L1000CDS^2 analysis results will be correspondingly updated.

The DEG tab gives the analysis results of DEGs, including a volcano plot, a heatmap, and a detailed result table of significantly DEGs together with a boxplot and a ROC curve plot for each DEG (Figure 4A). These results help reveal drug resistance-related molecules and candidate predictive biomarkers. The volcano plot visualizes the log FCs and adjusted *P*-values of all genes, highlighting the significantly DEGs. The heatmap is drawn based on top 10 significantly up-regulated genes and top 10 significantly down-regulated genes, according to the order of log FCs, presenting the gene expression levels of these genes across resistant and sensitive samples in the dataset. In the result table, for each gene, the presented results include log FC, *P*-value, and adjusted *P*-value from the differential gene expression analysis, and AUC, *P*-value and adjusted *P*-value from the ROC curve analysis. Genes can be re-ranked based on these fields. The gene with a large |log FC| and a small *P*-value may play an important role for drug resistance. ROC AUC can measure the ability of a gene discriminating between responders and non-responders, and genes with large AUCs are potential predictive biomarkers for drug response. Clicking on '>' before the gene symbol in the table will show the visualized results of the gene, including a boxplot presenting the expression level of the gene across resistant and sensitive samples, and a ROC curve plot. Further clicking on the gene symbol in the table will lead to the detailed annotation page for the gene, on which known drugs that can inhibit the gene will be listed (Figure 4B). This inhibitor annotation is particularly useful for searching drugs that can inhibit the drug resistance-related significantly up-regulated genes, and these drugs might be candidates that can help overcome the drug resistance.

In the ORA tab, over-representation analysis (ORA) results of significantly DEGs are shown, describing the functional characteristics of the drug resistance signature (Figure 4C). Here, six classes of functional gene sets are supported, including KEGG pathway, hallmark gene set, Reactome pathway, WikiPathways, microRNA target, and transcription factor target. This analysis helps discover the candidate biological pathways, microRNAs, and TFs that are important for drug resistance. For each class of gene sets, a detailed result table and a barplot presenting the top 10 gene sets, according to the increasing order of the adjusted *P*-value will be shown. Generally, gene sets with small *P*-values are noteworthy, and enrichment ratio larger/smaller than 1 means that the gene set is enriched/depleted.

The GSEA tab presents the gene set enrichment analysis (GSEA) results based on all genes ranked by log FCs (Figure 4D). ORA qualitatively considers genes with large expression differences (i.e. significantly DEGs), but when for a set of functionally related genes, the expression differences are small but their changes are in a coordinated way, ORA doesn't work. Many relevant phenotypic differences are caused by small but consistent changes in a set of genes. GSEA addresses this limitation (43). Here, six classes of gene sets (the same as ORA) are supported. For each class of gene sets, the presented results include a result table and two barplots showing the top 10 up-regulated gene sets (NES >1) and the top 10 down-regulated gene sets (NES <-1), respectively, according to the increasing order of the adjusted *P*-value. Generally, gene sets with |NES| >1 and small *P*-values are noteworthy.

In the L1000CDS^2 analysis tab, top 50 drug signatures that can reverse the drug resistance signature are listed (Figure 4E), with the help of L1000CDS^2 search engine (45). The resistance signature that is obtained by comparing expression profiles between non-responders and responders to drug A can suggest the reasons of the drug A resistance, and then we assume that a drug B that can reverse the resistance signature might be able to restore the patient response to the treatment. That is, drug B is a candidate combinational drug that may sensitize drug A. In fact, this principle

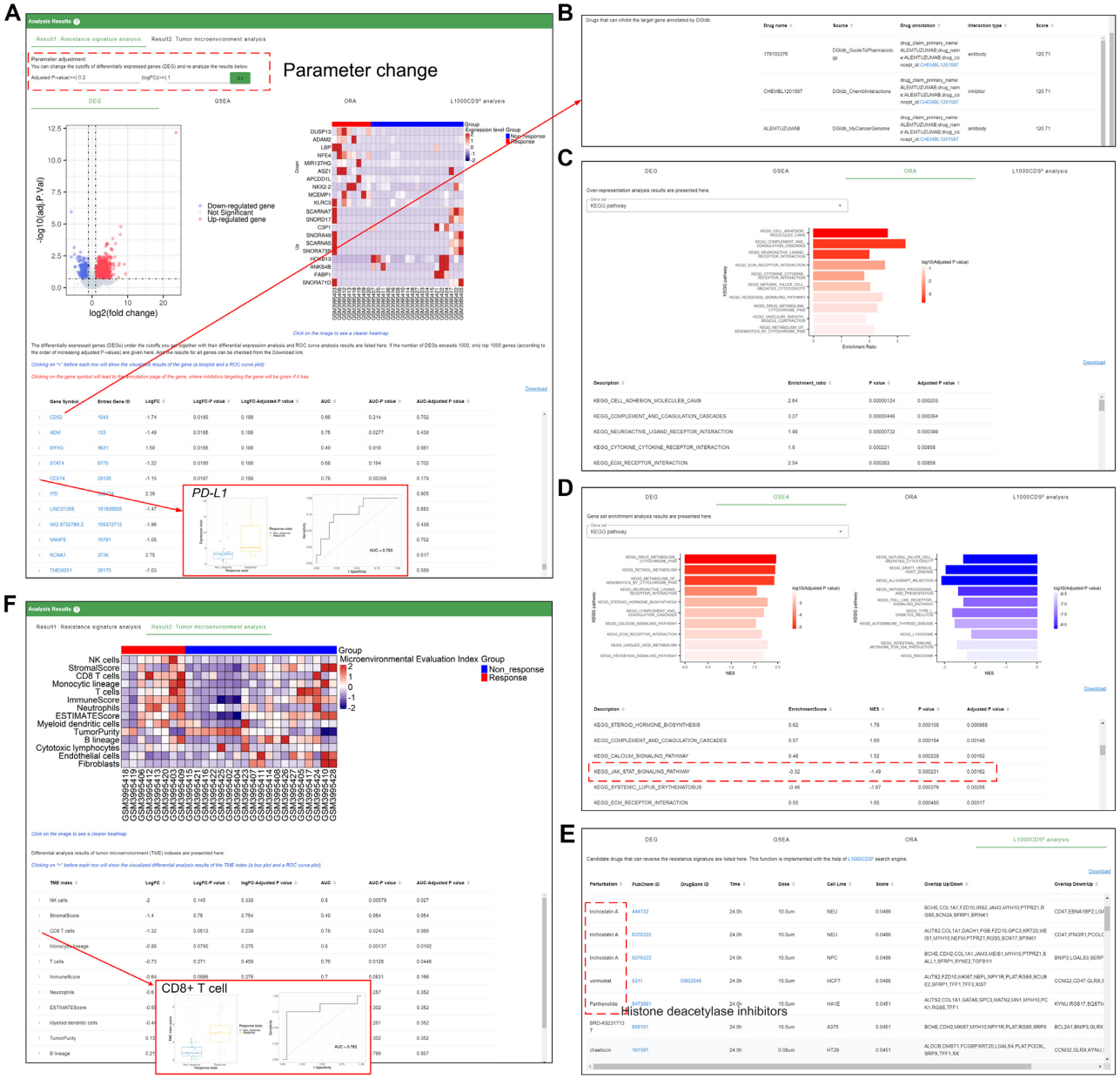

Figure 4. Single CTR-DB dataset analysis results. Here we use the results for the dataset of CTR_RNAseq_197 as the example. The results include 'Resistance signature analysis' (A, B, C, D and E) and 'Tumor microenvironment analysis' (F). (**A**) The resistance signature analysis result page and the results of the DEG tab. On this page, cutoffs to define significantly DEGs can be changed. And once the cutoffs are changed, the resistance signature analysis results will be updated. In the DEG tab, differential gene expression analysis results include a result table, a volcano plot, and a heatmap. Clicking on '>' before each gene in the result table will show a boxplot visualizing the differential expression analysis result and a ROC curve plot of the gene (shown by the embedded figure). Clicking on the gene symbol in the table will lead to the detailed annotation page of the gene, on which drugs/compounds known to inhibit the gene are given. (**B**) The inhibitor annotations on the detailed annotation page of gene *CD52*. (**C**) The ORA tab. Here shows the KEGG pathway ORA results, including a result table and a barplot visualizing the results of the top 10 KEGG pathways ranked by adjusted *P*-values. (**D**) The GSEA tab. Here shows the KEGG pathway GSEA results, including a result table and two barplots separately visualizing the results of the top 10 up-regulated (NES >1) and down-regulated (NES <−1) KEGG pathways ranked by adjusted *P*-values. (**E**) The L1000CDS^2 analysis tab. The result table presents the drug signatures that can reverse the drug resistance signature. Histone deacetylase inhibitors are highlighted on the graph. (**F**) Tumor microenvironment analysis result page, including a heatmap and a result table. Clicking on '>' before each TME index in the result table will show the visualized result plots of the differential analysis and ROC curve analysis of the TME index (shown by the embedded figure).

has been proved to be effective to identify possible combinational drugs (52). Therefore this function is designed to discover candidate combinational drugs that can overcome the drug resistance.

Tumor microenvironment analysis. Increasing evidence indicates that the tumor microenvironment (TME) is a crucial determinant of therapeutic resistance of many drugs, including chemotherapy, targeted therapy, and especially immunotherapy (53–55). One advantage of patient samples is that they can reflect the TME to some extent (47,48), and therefore this analysis is designed to reveal TME factors potentially associated with the drug resistance. Here TME indexes are firstly computed based on the gene expression profile of each sample. These indexes mainly reflect the abundance/fraction of various non-tumor cells and tumor purity in a patient sample. Then we perform the differential analysis of these indexes between non-response and response groups, to discover the drug resistance-related TME factors. The significantly differential TME factors help understand the mechanism of drug resistance, and even can be used as candidate predictive biomarkers. The presented analysis results include a heatmap and a result table (Figure 4F). The heatmap visualizes the TME indexes across responsive and non-responsive samples. The result table gives the log FC, *P*-value, and adjusted *P*-value of the differential analysis and AUC, *P*-value, and adjusted *P*-value of ROC curve analysis for each TME index. Clicking on '>' before each index will lead to the visualized results of the index, including a boxplot and a ROC curve plot.

CTR-DB dataset combination analysis

By the 'Combine' function of CTR-DB, users can select and combine existing CTR-DB datasets and specify the response/non-response grouping to perform the online analysis. The supported analysis functions are the same as the single-dataset analysis stated above. The purpose of this function is typically to allow users to define response and non-response groups according to their own needs; or to combine samples with the same treatment regimen and cancer subtype from different sources to obtain a larger sample size; or to merge CTR-DB datasets for a class of drugs (such as *PD1/PD-L1* inhibitors) or a cancer subtype of a coarser granularity etc.

To achieve the above purpose, each CTR-DB dataset (with a specific therapeutic regimen, cancer subtype, and source) was further divided into subsets, each of which was composed of samples with the same original drug response state. The original drug response states of samples in each CTR-DB dataset were provided by original authors. For example, some CTR-DB datasets have four original drug response states such as including complete response (CR), partial response (PR), stable disease (SD), and progressive disease (PD), some datasets have two states such as including pathological complete remission (pCR) and residual disease (RD), and some datasets have three or five original states. On the 'Combine' page, all such CTR-DB sub-datasets together with their key annotations are browsed, and can be further filtered by the dataset filtering function on the left. Users can specify which CTR-DB sub-datasets constitute the response group and which constitute the non-response group, and perform the further analysis.

During the dataset combination, users need to follow some rules. RNA-seq and microarray datasets cannot be combined. In addition, it is required that there is at least one sample in response group and meanwhile at least one sample in non-response group, which are from the same source dataset. Otherwise, we cannot distinguish whether the difference between responsive and non-responsive samples results from different batches or from the real biological difference. CTR-DB will check the submission, and the qualified one can be used for the subsequent analysis. In addition, users should be conscious that different CTR-DB datasets may have different definitions of original response states. Only samples with the same or similar original response states are suggested to constitute a group. The definitions of the original response states curated from the original reference can be checked on the detailed annotation page of each CTR-DB dataset, which we suggest users to check before dataset combination. The combination of adult and child samples (annotated in the 'Pediatric Oncology' field) should also be cautious.

CTR-DB dataset comparative analysis

'Compare' module implements the comparison of pre-analysis results of user-selected multiple CTR-DB datasets and their meta-analysis, aiming to explore the heterogeneity and homogeneity of drug resistance mechanism between different datasets (i.e. different patient cohorts) and even discover possible 'pan-dataset' (i.e. 'pan-cancer' or 'pan-drug') shared resistance mechanism and predictive biomarkers. The pre-analysis results that are supported to be compared include differential gene expression analysis ('Gene logFC'), gene ROC curve analysis ('Gene AUC'), TME index differential analysis ('Tumor microenvironment logFC'), TME index ROC curve analysis ('Tumor microenvironment AUC'), and KEGG pathway GSEA analysis ('KEGG NES').

On 'Compare' page, users can select CTR-DB datasets to be compared according to different needs. For example, users can select multiple datasets with the same therapeutic regimen and cancer type to study the heterogeneity and homogeneity of resistance mechanism across patient cohorts from different sources, or can select datasets using the same therapeutic regimen for different cancer (sub-) types, or select datasets for a drug class etc. Dataset filtering function on this page can facilitate the selection. Here we only consider CTR-DB datasets with sample size $\geq$10, responsive sample number >1, and meanwhile non-responsive sample number >1.

The comparative analysis results are shown in five tabs, including 'Gene logFC', 'Gene AUC', 'Tumor microenvironment logFC', 'Tumor microenvironment AUC', and 'KEGG NES' (Figure 5). Each tab presents a detailed table of result comparison together with a heatmap. Taking 'Gene AUC' as an example, in the result table, each row is a gene, giving its ROC curve analysis results (AUCs and *P*-values) across user-selected CTR-DB datasets together with the meta-*P*-value and the adjusted meta-*P*-value. Meta-*P*-value was obtained by integrating the *P*-values computed

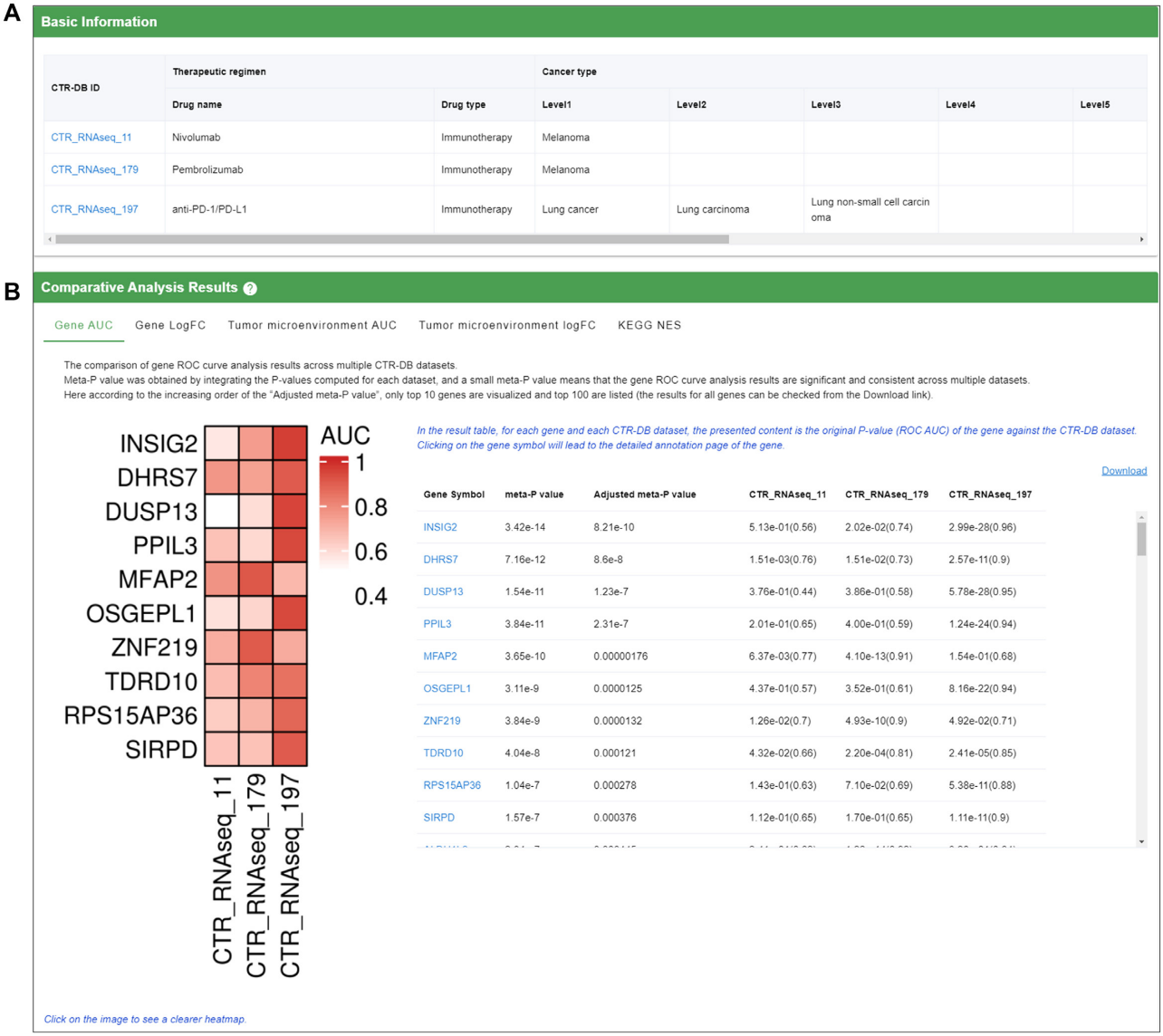

Figure 5. The result page of the 'Compare' module. Here, we use the comparison of three anti-*PD1*/*PD-L1* therapy-related datasets as the example. (**A**) Basic information for the user-selected CTR-DB datasets. Clicking on the CTR-DB ID in the table will lead to the detailed annotation page of the dataset. (**B**) Comparative results. The results include five tabs, including 'Gene AUC', 'Gene logFC', 'Tumor microenvironment AUC', 'Tumor microenvironment logFC', and 'KEGG NES'. In each tab, a heatmap and a result table will be presented. Clicking on the gene symbol in the table will lead to the detailed annotation page of the gene.

for each dataset, and a small meta-*P*-value means that the gene has a strong and consistent association with drug resistance across multiple datasets (49). For 'Gene AUC' result, a gene with a small meta-*P*-value might be a potential 'pan-dataset' predictive biomarker. For other results, small meta-*P*-values suggest potential 'pan-dataset' resistance mechanism etc. The heatmap presents the AUCs across user-selected CTR-DB datasets of top 10 genes, in the increasing order of adjusted meta-*P*-values.

Biomarker validation

Predictive biomarkers can predict the patient response to a drug, which are crucial for the cancer precise medicine (56). 'Biomarker validation' module mainly serves two purposes. One is that for the interested candidate predictive biomarker for a certain drug and a certain cancer type (e.g. supported by the cell line level evidence), users can validate its predictive ability using corresponding transcriptomes of CTR-DB patient cohorts. The other is that for the interested gene, users can check its correlation with drug resistances across various cancer subtypes and drugs, which can help identify the functional significance of the gene and design the following experiments.

This function can be accessed directly by 'Biomarker validation' in the navigation bar or by the gene browse page. Searching and further clicking on the interested gene, the analysis results for the biomarker validation will be shown on the detailed annotation page of the gene, including a result table and two barplots (Figure 6). The re-

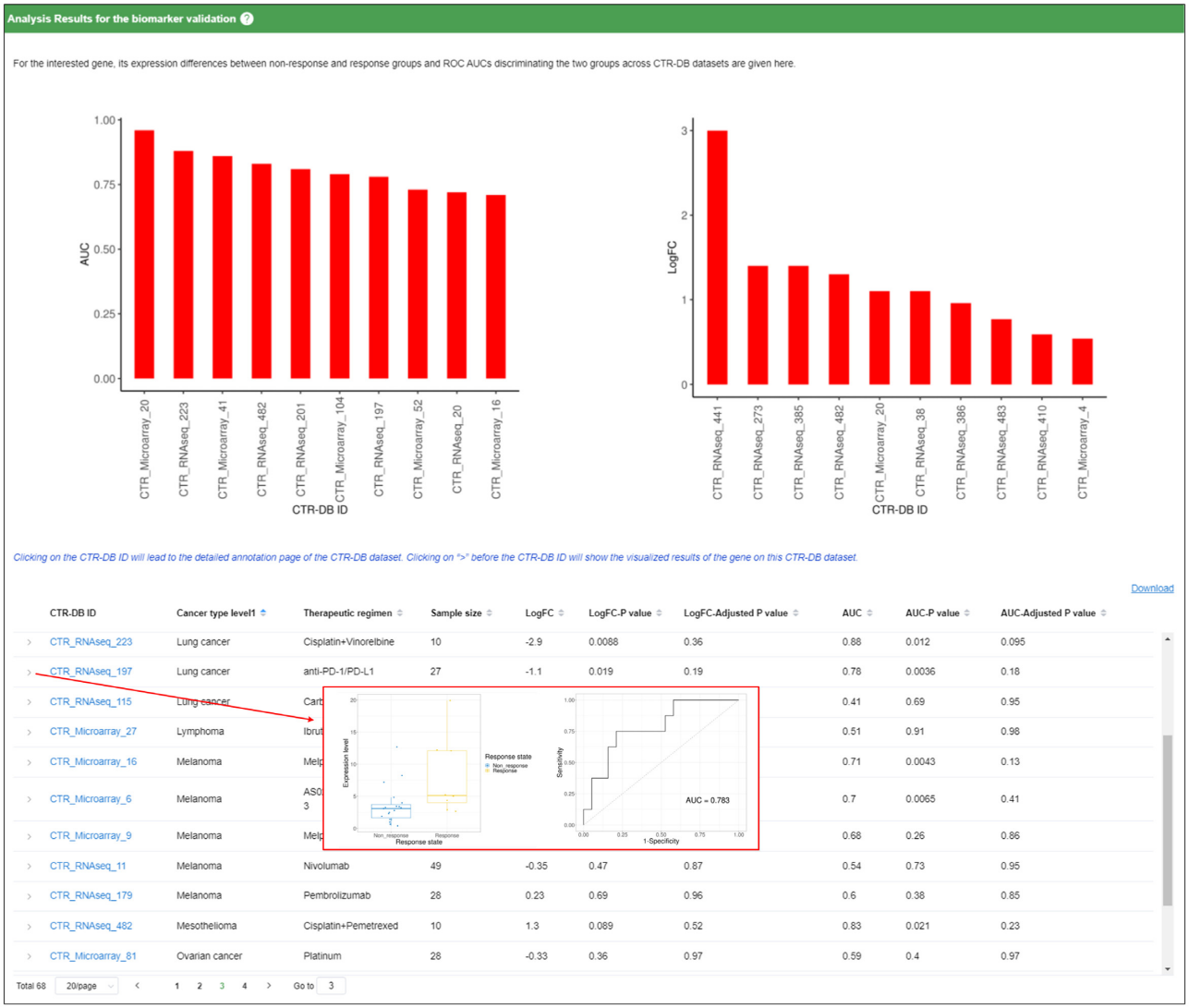

Figure 6. The result page of the 'Biomarker validation' module. Here we use *PD-L1* (i.e. *CD274*) as the example. The result table gives differential expression analysis and ROC curve analysis results of the gene across CTR-DB datasets. Clicking on '>' before each CTR-DB ID will show the visualized results of the gene on this CTR-DB dataset, including a boxplot and a ROC curve plot (shown by the embedded figure). And clicking on the CTR-DB ID will lead to the detailed annotation page of the CTR-DB dataset. Two barplots, respectively, visualize ROC curve analysis results and differential expression analysis results of the gene against top 10 most significant CTR-DB datasets.

sult table presents both the differential expression analysis results (log FC, *P*-value, adjusted *P*-value) and ROC curve analysis results (AUC, *P*-value, adjusted *P*-value) for the gene across CTR-DB datasets (a row corresponds to a dataset). Users can identify the interested CTR-DB patient dataset based on cancer type and therapeutic regimen given in the result table. Here the CTR-DB datasets meeting 'sample size $\geq$10, responsive sample size $>$1, and meanwhile non-responsive sample size $>$1' are considered. The barplot on the left presents ROC AUCs of the gene against top 10 CTR-DB datasets, in the decreasing order of AUCs. The right one presents log FCs of the gene against top 10 CTR-DB datasets, in the decreasing order of log FCs. The two barplots illustrate datasets (i.e. the drugs and cancer types) whose drug resistances are the most correlated to the expression of the interested gene.

Use case

To demonstrate the usage of CTR-DB, we use anti-*PD1*/*PD-L1* therapy resistance as an example. *PD1* is a famous immune checkpoint protein on T cells, and its binding to *PD-L1* on tumor cells promotes tumor immune escape. Over the past several years, *PD1*/*PD-L1* blockade therapy has been used in multiple solid tumors such as non-small-cell lung cancer (NSCLC) and melanoma, however, only a small proportion of patients show clinical response (57,58). Therefore, it is crucial to reveal molecular determinants of anti-*PD1*/*PD-L1* therapy response.

Resistance mechanism elucidation. The dataset of CTR_RNAseq_197 has 27 NSCLC patients receiving *PD1*/*PD-L1* blockade therapy, including 19 non-responders and 8 responders. Firstly by the 'Resistance

signature analysis', we find that in non-responders in this cohort *PD-L1* is significantly down-regulated (i.e. *CD274*, log FC = −1.15, *P*-value = 0.02) (Figure 4A), together with its upstream *JAK-STAT* signaling pathway (NES = −1.49, *P*-value = 2.31e−04) (Figure 4D) and its further upstream *INF-γ* (i.e. *IFNG*, log FC = −1.82, *P*-value = 0.03). And antigen processing and presentation pathway (NES = −2.55, *P*-value = 2.10e−10) is also significantly down-regulated. On the other hand, 'Tumor microenvironment analysis' results indicate that in non-responsive samples the fraction of immune cells ('ImmuneScore', log FC = −0.64, *P*-value = 0.10), especially the abundance of $CD8^+$ T cells (log FC = −1.32, *P*-value = 0.05) is apparently lower (Figure 4F). These two aspects of results are consistent with each other, because in antitumor immunity, $CD8^+$ T cells are key sources of *IFN-γ* production (59). These results suggest that low immune cell infiltration especially low $CD8^+$ T-cell infiltration in TME as well as *PD-L1* low expression induced by the down-regulated *INF-γ-JAK-STAT* cascade may be the mechanism of the anti-*PD1*/*PD-L1* therapy resistance. These findings are consistent with previous studies (57).

Sensitizing drug discovery. To further discover candidate drugs that can sensitize anti-*PD1*/*PD-L1* therapy, on CTR_RNAseq_197, we used 'Resistance signature analysis → $L1000CDS^2$ analysis' to search drugs that can reverse the resistance signature (|log FC| ≥1, adjusted *P*-value ≤0.05). In result, we find that 15 of the returned top 50 drug signatures are histone deacetylase (HDAC) inhibitors (Figure 4E), strongly suggesting HDAC inhibitors are promising sensitizing drug candidates for anti-*PD1*/*PD-L1* therapy resistance when immune infiltration degree is low and *PD-L1* is lowly expressed. In fact, previous studies have indicated that on the one hand, HDAC inhibitors can enhance antigen presentation and $CD8^+$ T cell and NK cell infiltration (60–62), but on the other hand, HDAC inhibitors can also stimulate *PD-L1* expression, dampening subsequent T-cell activation (63). Therefore the combination of HDAC inhibitors and *PD1*/*PD-L1* blockade can achieve the treatment sensitivity by recruiting immune cells and meanwhile avoiding immune escape. Indeed, we find 'HDAC inhibitor + *PD1*/*PD-L1* blockade' combination has already been in the clinical trials for NSCLC (64).

Dataset combination analysis. To confirm the above findings in a larger patient population, by CTR-DB dataset 'Combine' function, we combined all 143 samples receiving anti-*PD1*/*PD-L1* therapy in CTR-DB, including 47 responders and 96 non-responders. On this larger merged dataset, we obtained consistent results (Supplementary Figure S1). For example, the low infiltration level of immune cells ('ImmuneScore', log FC = −1.08, *P*-value = 3.17e−03), especially $CD8^+$ T cells (log FC = −1.13, *P*-value = 3.88e−03) in non-responders is more significant on the merged population that is larger than the previous single dataset (Supplementary Figure S1D).

Dataset comparative analysis. By 'Compare' function, we further compared the CTR-DB patient sets receiving *PD1*/*PD-L1* blockade therapy but from different sources. There are three related CTR-DB datasets with sample size ≥10, including CTR_RNA_197 for NSCLC, CTR_RNAseq_11 for melanoma and CTR_RNAseq_179 for metastatic melanoma. By comparison, we find that drug resistance mechanism shows high heterogeneity in different patient cohorts. For example, compared to responders, the non-responsive metastatic melanoma patients in CTR_RNAseq_179 have slightly higher immune cell infiltration tendency ('ImmuneScore', log FC = 0.11, *P*-value = 0.76; $CD8^+$ T cell, log FC = 0.99, *P*-value = 0.36), slightly up-regulated *JAK-STAT* pathway (NES = 0.9, *P*-value = 0.71) induced slightly higher *PD-L1* expression (log FC = 0.23, *P*-value = 0.69), up-regulated NK cell mediated cytotoxicity pathway (NES = 1.24, *P*-value = 0.08) etc., suggesting that these non-responders in CTR_RNAseq_179 have an apparently different resistance mechanism from CTR_RNAseq_197 patients stated above. Indeed, due to high tumor heterogeneity, drug resistance mechanism has been found to be of high heterogeneity and complexity (57).

Biomarker validation. *PD-L1* has been clinically widely used as a biomarker for identifying potential responders of anti-*PD1*/*PD-L1* therapy (57). Here we also validated its predictive ability using transcriptomic data of CTR-DB patients. We find that for all patients receiving anti-*PD1*/*PD-L1* therapy in CTR-DB, *PD-L1* expression indeed has some predictive power (ROC AUC = 0.64, *P*-value = 5.33e−03, Supplementary Figure S1A). However, further by 'biomarker validation' function we see that only on CTR_RNAseq_197 it can effectively predict treatment response (ROC AUC = 0.78, *P*-value = 3.60e−03), while on both CTR_RNAseq_11 and CTR_RNAseq_179 it has little predictive power (Figure 6). This result verifies the previous knowledge that the predictive accuracy of *PD-L1* expression is insufficient. *PD-L1* positive patients are not always responders and negative patients are not necessarily non-responsive (57). More effective combinational biomarkers are needed.

In summary, this use case shows the power of CTR-DB on generating and validating hypotheses on drug resistance mechanism, discovering sensitizing drugs, biomarker validation, and resistance mechanism heterogeneity exploration.

DISCUSSION

To date, only some cancer patients can benefit from drug treatment due to cancer heterogeneity. Drug resistance of cancer patients is one of the most important and challenging problems in the era of precision medicine (65). Rapidly accumulated patient-derived clinical transcriptomes with cancer drug response bring unprecedented opportunities for studying this issue, and moreover their integration and reuse may provide new insights. However, there is still no database systematically collecting and helping integrate and reuse these data. Therefore, we have developed CTR-DB. CTR-DB, as the first database for patient-derived gene expression signatures correlated with cancer drug response, has several advantages: (i) comprehensive patient-derived clinical transcriptomes with cancer drug response,

covering abundant drugs and cancer types; (ii) uniformly reprocessed transcriptomic data; (iii) manually-curated and standardized drug response information; (iv) multiple data analysis functions for a single dataset, multiple-dataset combination and comparison, and biomarker validation function to facilitate data mining, integration, and reuse; (v) and the user-friendly and interactive interface. All of these features enable CTR-DB to be a valuable resource, satisfying multiple needs from both basic and clinical cancer researchers.

In future, on the data aspect, we will regularly add new datasets into CTR-DB. And meanwhile we plan to expand data scope, such as including acquired resistance-related clinical transcriptomes, which can also inform the resistance mechanism and combinational drugs. On the database function aspect, we will support data submission, to enable the crowdsourcing data collection and curation. Other improvements include supporting combinational biomarker discovery, drug response prediction and clinical drug recommendation based on transcriptomic signals of patients, and online data analysis for users' own datasets etc.

DATA AVAILABILITY

CTR-DB can be accessed at http://ctrdb.ncpsb.org.cn/, and is compatible with Chrome, Firefox, and Opera browsers for Windows; and Safari, Chrome, Firefox, and Opera browsers for the Mac operating system.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Pengbo Cao, Yahui Wang, Kaikun Xu and Yuanfeng Li for the fruitful discussion. We also thank the bioinformatics platform at Phoenix Center for the strong and stable IT support.

FUNDING

National Key Research and Development Program of China [2020YFE0202200, 2017YFC1700105]; National Natural Science Foundation of China [32088101, 31871341]; State Key Laboratory of Proteomics of China [SKLPO202010]; Beijing Talents foundation [to D.L.]. Funding for open access charge: the National Key Research and Development Program of China [2020YFE0202200].
Conflict of interest statement. None declared.

REFERENCES

1. Schwaederle,M., Zhao,M., Lee,J.J., Eggermont,A.M., Schilsky,R.L., Mendelsohn,J., Lazar,V. and Kurzrock,R. (2015) Impact of precision medicine in diverse cancers: a meta-analysis of phase II clinical trials. *J. Clin. Oncol.*, **33**, 3817–3825.
2. Holohan,C., Van Schaeybroeck,S., Longley,D.B. and Johnston,P.G. (2013) Cancer drug resistance: an evolving paradigm. *Nat. Rev. Cancer*, **13**, 714–726.
3. Hugo,W., Zaretsky,J.M., Sun,L., Song,C., Moreno,B.H., Hu-Lieskovan,S., Berent-Maoz,B., Pang,J., Chmielowski,B., Cherry,G. *et al.* (2016) Genomic and transcriptomic features of response to anti-PD-1 therapy in metastatic melanoma. *Cell*, **165**, 35–44.
4. Amato,C.M., Hintzsche,J.D., Wells,K., Applegate,A., Gorden,N.T., Vorwald,V.M., Tobin,R.P., Nassar,K., Shellman,Y.G., Kim,J. *et al.* (2020) Pre-treatment mutational and transcriptomic landscape of responding metastatic melanoma patients to anti-PD1 immunotherapy. *Cancers*, **12**, 1943.
5. Lee,J.S., Nair,N.U., Dinstag,G., Chapman,L., Chung,Y., Wang,K., Sinha,S., Cha,H., Kim,D., Schperberg,A.V. *et al.* (2021) Synthetic lethality-mediated precision oncology via the tumor transcriptome. *Cell*, **184**, 2487–2502.
6. Ayers,M., Lunceford,J., Nebozhyn,M., Murphy,E., Loboda,A., Kaufman,D.R., Albright,A., Cheng,J.D., Kang,S.P., Shankaran,V. *et al.* (2017) IFN-γ-related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.*, **127**, 2930–2940.
7. Luna,A., Elloumi,F., Varma,S., Wang,Y., Rajapakse,V.N., Aladjem,M.I., Robert,J., Sander,C., Pommier,Y. and Reinhold,W.C. (2021) CellMiner cross-database (CellMinerCDB) version 1.2: exploration of patient-derived cancer cell line pharmacogenomics. *Nucleic Acids Res.*, **49**, D1083–D1093.
8. Caroli,J., Sorrentino,G., Forcato,M., Del Sal,G. and Bicciato,S. (2018) GDA, a web-based tool for genomics and drugs integrated analysis. *Nucleic Acids Res.*, **46**, W148–W156.
9. Mer,A.S., Ba-Alawi,W., Smirnov,P., Wang,Y.X., Brew,B., Ortmann,J., Tsao,M.S., Cescon,D.W., Goldenberg,A. and Haibe-Kains,B. (2019) Integrative pharmacogenomics analysis of patient-derived xenografts. *Cancer Res.*, **79**, 4539–4550.
10. Borisov,N., Sorokin,M., Tkachev,V., Garazha,A. and Buzdin,A. (2020) Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments. *BMC Med. Genomics*, **13**, 111.
11. Fekete,J.T. and Győrffy,B. (2019) ROCplot.org: Validating predictive biomarkers of chemotherapy/hormonal therapy/anti-HER2 therapy using transcriptomic data of 3,104 breast cancer patients. *Int. J. Cancer*, **145**, 3140–3151.
12. Fekete,J.T., Osz,A., Pete,I., Nagy,G.R., Vereczkey,I. and Gyorffy,B. (2020) Predictive biomarkers of platinum and taxane resistance using the transcriptomic data of 1816 ovarian cancer patients. *Gynecol. Oncol.*, **156**, 654–661.
13. Menyhart,O., Fekete,J.T. and Gyorffy,B. (2021) Gene expression-based biomarkers designating glioblastomas resistant to multiple treatment strategies. *Carcinogenesis*, **42**, 804–813.
14. Jang,S.K., Yoon,B.H., Kang,S.M., Yoon,Y.G., Kim,S.Y. and Kim,W. (2019) CDRgator: an integrative navigator of cancer drug resistance gene signatures. *Mol. Cells*, **42**, 237–244.
15. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
16. Sarkans,U., Füllgrabe,A., Ali,A., Athar,A., Behrangi,E., Diaz,N., Fexova,S., George,N., Iqbal,H., Kurri,S. *et al.* (2021) From ArrayExpress to BioStudies. *Nucleic Acids Res.*, **49**, D1502–D1506.
17. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
18. Kodama,Y., Shumway,M., Leinonen,R. and International Nucleotide Sequence Database Collaboration. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
19. Mounir,M., Lucchetta,M., Silva,T.C., Olsen,C., Bontempi,G., Chen,X., Noushmehr,H., Colaprico,A. and Papaleo,E. (2019) New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput. Biol.*, **15**, e1006701.
20. Goldman,M.J., Craft,B., Hastie,M., Repecka,K., McDade,F., Kamath,A., Banerjee,A., Luo,Y., Rogers,D., Brooks,A.N. *et al.* (2020) Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.*, **38**, 675–678.
21. Gautier,L., Cope,L., Bolstad,B.M. and Irizarry,R.A. (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

22. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10–12.
23. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
24. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.
25. Leek,J.T., Johnson,W.E., Parker,H.S., Jaffe,A.E. and Storey,J.D. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.
26. Zhang,Y., Parmigiani,G. and Johnson,W.E. (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.*, **2**, lqaa078.
27. Wishart,D.S., Feunang,Y.D., Guo,A.C., Lo,E.J., Marcu,A., Grant,J.R., Sajed,T., Johnson,D., Li,C., Sayeeda,Z. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, **46**, D1074–D1082.
28. Mendez,D., Gaulton,A., Bento,A.P., Chambers,J., De Veij,M., Felix,E., Magarinos,M.P., Mosquera,J.F., Mutowo,P., Nowotka,M. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, **47**, D930–D940.
29. Kim,S., Chen,J., Cheng,T., Gindulyte,A., He,J., He,S., Li,Q., Shoemaker,B.A., Thiessen,P.A., Yu,B. *et al.* (2021) PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.*, **49**, D1388–D1395.
30. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R. *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.
31. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52–D57.
32. Tweedie,S., Braschi,B., Gray,K., Jones,T., Seal,R.L., Yates,B. and Bruford,E.A. (2021) Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res.*, **49**, D939–D946.
33. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
34. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
35. Martens,M., Ammar,A., Riutta,A., Waagmeester,A., Slenter,D.N., Hanspers,K., Miller,R.A., Digles,D., Lopes,E.N., Ehrhart,F. *et al.* (2021) WikiPathways: connecting communities. *Nucleic Acids Res.*, **49**, D613–D621.
36. Liberzon,A., Birger,C., Thorvaldsdottir,H., Ghandi,M., Mesirov,J.P. and Tamayo,P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.*, **1**, 417–425.
37. Freshour,S.L., Kiwala,S., Cotto,K.C., Coffman,A.C., McMichael,J.F., Song,J.J., Griffith,M., Griffith,O.L. and Wagner,A.H. (2021) Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsource efforts. *Nucleic Acids Res.*, **49**, D1144–D1151.
38. Fawcett,T. (2006) An introduction to ROC analysis. *Pattern Recognit. Lett.*, **27**, 861–874.
39. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
40. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
41. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.C. and Muller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.*, **12**, 77.
42. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.
43. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
44. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L. *et al.* (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, **2**, 100141.
45. Duan,Q., Reid,S.P., Clark,N.R., Wang,Z., Fernandez,N.F., Rouillard,A.D., Readhead,B., Tritsch,S.R., Hodos,R., Hafner,M. *et al.* (2016) L1000CDS^2: LINCS L1000 characteristic direction signatures search engine. *NPJ Syst. Biol. Appl.*, **2**, 16015.
46. Subramanian,A., Narayan,R., Corsello,S.M., Peck,D.D., Natoli,T.E., Lu,X., Gould,J., Davis,J.F., Tubelli,A.A., Asiedu,J.K. *et al.* (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437–1452.
47. Yoshihara,K., Shahmoradgoli,M., Martinez,E., Vegesna,R., Kim,H., Torres-Garcia,W., Trevino,V., Shen,H., Laird,P.W., Levine,D.A. *et al.* (2013) Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.*, **4**, 2612.
48. Becht,E., Giraldo,N.A., Lacroix,L., Buttard,B., Elarouci,N., Petitprez,F., Selves,J., Laurent-Puig,P., Sautes-Fridman,C., Fridman,W.H. *et al.* (2016) Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.*, **17**, 218.
49. Vasaikar,S.V., Straub,P., Wang,J. and Zhang,B. (2018) LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.*, **46**, D956–D963.
50. Wickham,H. (2016) In: *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, NY.
51. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.
52. Wei,G., Twomey,D., Lamb,J., Schlis,K., Agarwal,J., Stam,R.W., Opferman,J.T., Sallan,S.E., den Boer,M.L., Pieters,R. *et al.* (2006) Gene expression-based chemical genomics identifies rapamycin as a modulator of MCL1 and glucocorticoid resistance. *Cancer Cell*, **10**, 331–342.
53. Tredan,O., Galmarini,C.M., Patel,K. and Tannock,I.F. (2007) Drug resistance and the solid tumor microenvironment. *J. Natl. Cancer Inst.*, **99**, 1441–1454.
54. Sun,D., Wang,J., Han,Y., Dong,X., Ge,J., Zheng,R., Shi,X., Wang,B., Li,Z., Ren,P. *et al.* (2021) TISCH: a comprehensive web resource enabling interactive single-cell transcriptome visualization of tumor microenvironment. *Nucleic Acids Res.*, **49**, D1420–D1430.
55. Junttila,M.R. and de Sauvage,F.J. (2013) Influence of tumour micro-environment heterogeneity on therapeutic response. *Nature*, **501**, 346–354.
56. Chen,J.J., Lu,T.P., Chen,Y.C. and Lin,W.J. (2015) Predictive biomarkers for treatment selection: statistical considerations. *Biomarkers Med.*, **9**, 1121–1135.
57. Lei,Q., Wang,D., Sun,K., Wang,L. and Zhang,Y. (2020) Resistance mechanisms of anti-PD1/PDL1 therapy in solid tumors. *Front. Cell. Dev. Biol.*, **8**, 672.
58. Kim,J.Y., Choi,J.K. and Jung,H. (2020) Genome-wide methylation patterns predict clinical benefit of immunotherapy in lung cancer. *Clin. Epigenet.*, **12**, 119.
59. Gocher,A.M., Workman,C.J. and Vignali,D.A.A. (2021) Interferon-γ: teammate or opponent in the tumour microenvironment? *Nat. Rev. Immunol.*, https://www.nature.com/articles/s41577-021-00566-3.
60. Neuwelt,A.J., Kimball,A.K., Johnson,A.M., Arnold,B.W., Bullock,B.L., Kaspar,R.E., Kleczko,E.K., Kwak,J.W., Wu,M.H., Heasley,L.E. *et al.* (2020) Cancer cell-intrinsic expression of MHC II in lung cancer cell lines is actively restricted by MEK/ERK signaling and epigenetic mechanisms. *Immunother. Cancer*, **8**, e000441.
61. Adeegbe,D.O., Liu,Y., Lizotte,P.H., Kamihara,Y., Aref,A.R., Almonte,C., Dries,R., Li,Y., Liu,S., Wang,X. *et al.* (2017) Synergistic immunostimulatory effects and therapeutic benefit of combined histone deacetylase and bromodomain inhibition in non-small cell lung cancer. *Cancer Discov.*, **7**, 852–867.
62. Zhu,M., Huang,Y., Bender,M.E., Girard,L., Kollipara,R., Eglenen-Polat,B., Naito,Y., Savage,T.K., Huffman,K.E., Koyama,S. *et al.* (2021) Evasion of innate immunity contributes to small cell lung cancer progression and metastasis. *Cancer Res.*, **81**, 1813–1826.

63. Briere,D., Sudhakar,N., Woods,D.M., Hallin,J., Engstrom,L.D., Aranda,R., Chiang,H., Sodre,A.L., Olson,P., Weber,J.S. *et al.* (2018) The class I/IV HDAC inhibitor mocetinostat increases tumor antigen presentation, decreases immune suppressive cell types and augments checkpoint inhibitor therapy. *Cancer Immunol. Immunother.*, **67**, 381–392.
64. Gray,J.E., Saltos,A., Tanvetyanon,T., Haura,E.B., Creelan,B., Antonia,S.J., Shafique,M., Zheng,H., Dai,W., Saller,J.J. *et al.* (2019) Phase I/Ib study of pembrolizumab plus vorinostat in advanced/metastatic non-small cell lung cancer. *Clin. Cancer Res.*, **25**, 6623–6632.
65. Garraway,L.A. and Jänne,P.A. (2012) Circumventing cancer drug resistance in the era of personalized medicine. *Cancer Discov.*, **2**, 214–226.