Data Query

Functions to query data from UCSC Xena data hubs in single molecular level. The functionality of functions below may overlap, we rank them by importance.

query_molecule_value()	Get Molecule or Signature Data Values from Dense (Genomic) Matrix Dataset of UCSC Xena Data Hubs
query_pancan_value()	Query Single Identifier or Signature Value from Pan-cancer Database
get_ccle_cn_value() get_ccle_gene_value() get_ccle_protein_value() get_ccle_mutation_status() get_pancan_value() get_pancan_gene_value() get_pancan_transcript_value() get_pancan_protein_value() get_pancan_mutation_status() get_pancan_cn_value() get_pancan_methylation_value() get_pancan_miRNA_value()	Fetch Identifier Value from Pan-cancer Dataset
query_toil_value_df() query_toil_value_df()	Obtain ToilHub Info for Single Molecule
available_hosts()	Show Available Hosts

Analysis and Visualiation for Pan-Cancer Study

Functions to analyze and visualize data mainly from TCGA and CCLE

tcga_surv_get() tcga_surv_plot()	TCGA Survival Analysis
analyze_gene_drug_response_asso()	Analyze Association between Gene (Signature) and Drug Response with CCLE Data
vis_gene_drug_response_asso()	Visualize Gene and Drug-Target Association with CCLE Data
analyze_gene_drug_response_diff()	Analyze Difference of Drug Response (IC50 Value (uM)) between Gene (Signature) High and Low Expression with CCLE Data
vis_gene_drug_response_diff()	Visualize Gene and Drug Response Difference with CCLE Data
vis_ccle_tpm()	Visualize CCLE Gene Expression
vis_ccle_gene_cor()	Visualize CCLE Gene Expression Correlation
vis_gene_TIL_cor()	Heatmap for Correlation between Gene and Tumor Immune Infiltration (TIL)
vis_gene_immune_cor()	Heatmap for Correlation between Gene and Immune Signatures
vis_gene_cor()	Visualize Gene-Gene Correlation in TCGA
vis_gene_cor_cancer()	Visualize Gene-Gene Correlation in a TCGA Cancer Type
vis_gene_stemness_cor()	Visualize Correlation between Gene and Tumor Stemness
vis_gene_tmb_cor() vis_gene_msi_cor()	Visualize Correlation between Gene and TMB (Tumor Mutation Burden) Visualize Correlation between Gene and MSI (Microsatellite instability)
vis_pancan_anatomy()	Visualize Single Gene Expression in Anatomy Location
vis_toil_TvsN()	Visualize Pan-cancer TPM (tumor (TCGA) vs Normal (TCGA & GTEx))
vis_toil_TvsN_cancer()	Visualize Gene TPM in Single Cancer Type (Tumor (TCGA) vs Normal (TCGA & GTEx))
vis_unicox_tree()	Visualize Single Gene Univariable Cox Result from Toil Data Hub

General Analysis and Visualiation

Functions to analyze and visualize general data, i.e., any valid data from UCSC Xena or users

vis_identifier_cor()	Visualize Identifier-Identifier Correlation
vis_identifier_multi_cor()	Visualize Correlation for Multiple Identifiers
vis_identifier_grp_comparison()	Visualize Comparison of an Molecule Identifier between Groups
vis_identifier_grp_surv()	Visualize Identifier Group Survival Difference
ezcor()	Run Correlation between Two Variables and Support Group by a Variable
ezcor_batch()	Run correlation between two variables in a batch mode and support group by a variable
ezcor_partial_cor()	Run partial correlation

Table S1. Major functions provided by *UCSCXenaShiny*. All public data and functions provided by the tool is available at https://openbiox.github.io/UCSCXenaShiny/reference/index.html.

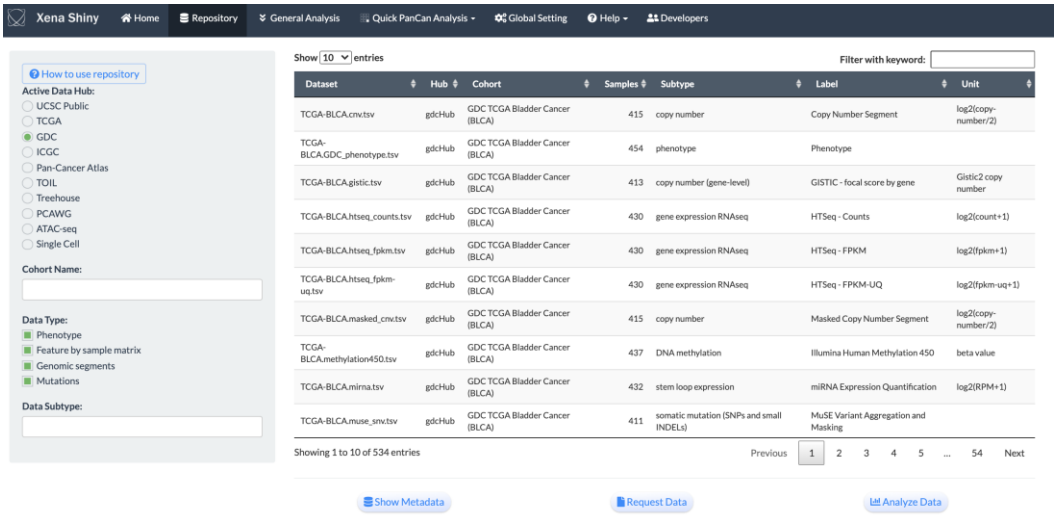
Table S2

Tool		UCSCXenaShiny	UCSC Xena Browser	UCSCXenaTools	xenaPython
	Visual download	Yes	Yes (limited)	No	No
Data download features	Any dataset download	Yes	Yes	Yes	No
	Any data subset download	Yes	Yes	Yes	Yes
	TCGA/CCLE data quick query	Yes	No	No	No
	Plot data download	Yes	Yes	No	No
General analysis features	Scatter correlation (for two variables)	Yes	Yes	No	No
	Correlation heatmap (for multiple variables)	Yes	No	No	No
	Group comparison	Yes	Yes	No	No
	Survival analysis	Yes	Yes	No	No
	Genomic signatures	Yes	Yes	No	No
	Sample filtering	Yes	Yes	No	No
	Custom data support	Yes	Yes	No	No
	Visual spreadsheet	No	Yes	No	No
Specified pan-cancer analysis features	TCGA/GTEx, PCAWG, CCLE molecular profile value distribution	Yes	No	No	No
	TCGA, PCAWG, CCLE molecular profile correlation	Yes	No	No	No
	CCLE drug sensitivity analysis	Yes	No	No	No
	TCGA/GTEx molecular profile anatomy	Yes	No	No	No
	TCGA, PCAWG molecular profile survival analysis	Yes	Yes (limited)	No	No
	TCGA molecular profile and immune signature/infiltration association	Yes	No	No	No
	TCGA, PCAWG molecular profile and tumor features (stemness/MSI/TMB) association	Yes	No	No	No
Generate publication ready plots		Yes	No	No	No
Has API		Yes (available in self-constructed R package)	Yes (available as UCSCXenaTools and xenaPython)	It's itself an API	It's itself an API
Target user type		User with and without programming background	User without programming background	User with programming background	User with programming background
URL		https://github.com/openbiox/UCSCXenaShiny	https://xenabrowser.net/	https://github.com/ropensci/UCSCXenaTools	https://github.com/ucscXena/xenaPython

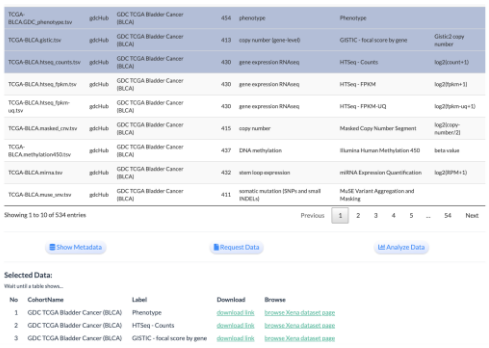
Table S2. A comparison of different tools for retrieving and analysis of UCSC Xena's data. Note: "API" means application programming interface. "URL" means uniform resource locator.

Figure S1

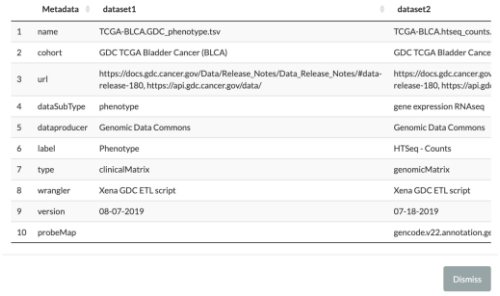
A



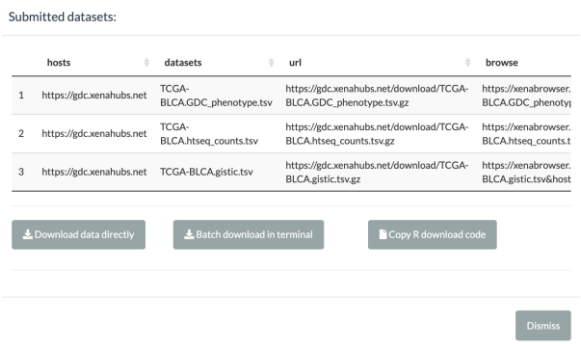
B



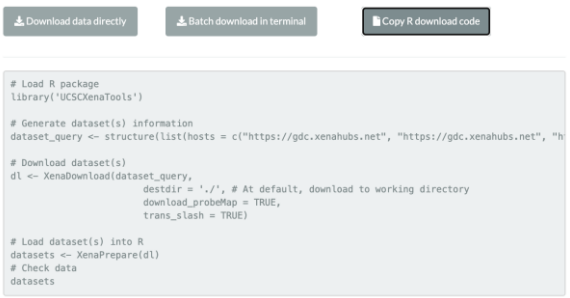
C



D



E



F

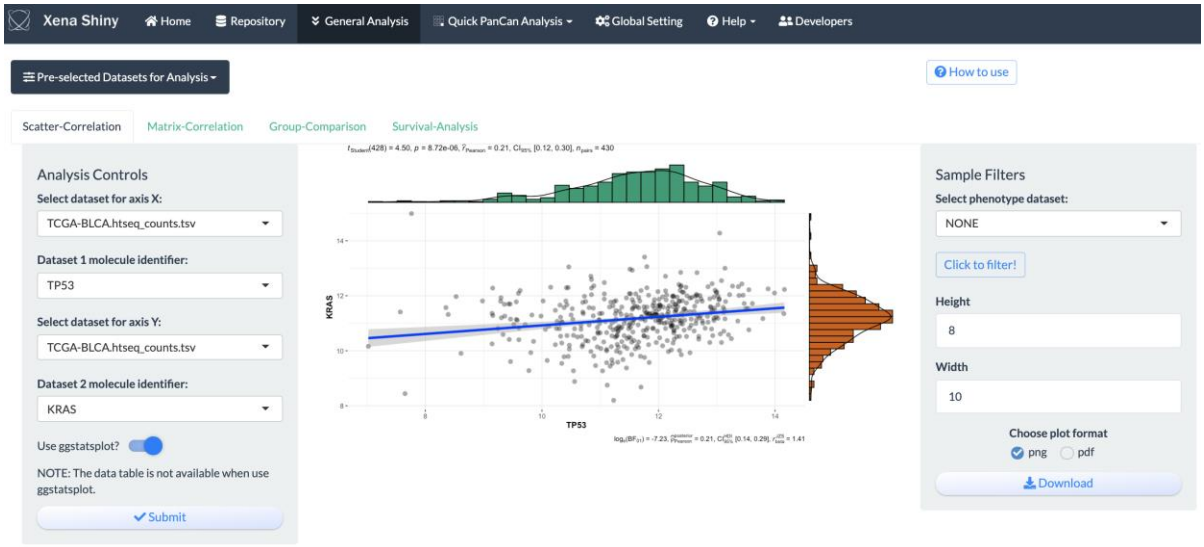
2021-05-01-commands.sh

```
#!/usr/bin/env bash
#Usage: run bash 2021-05-01-commands.sh in your terminal under a desired directory
wget -c https://gdc.xenahubs.net/download/TCGA-BLCA.GDC_phenotype.tsv.gz
wget -c https://gdc.xenahubs.net/download/TCGA-BLCA.htseq_counts.tsv.gz
wget -c https://gdc.xenahubs.net/download/TCGA-BLCA.gistic.tsv.gz
```

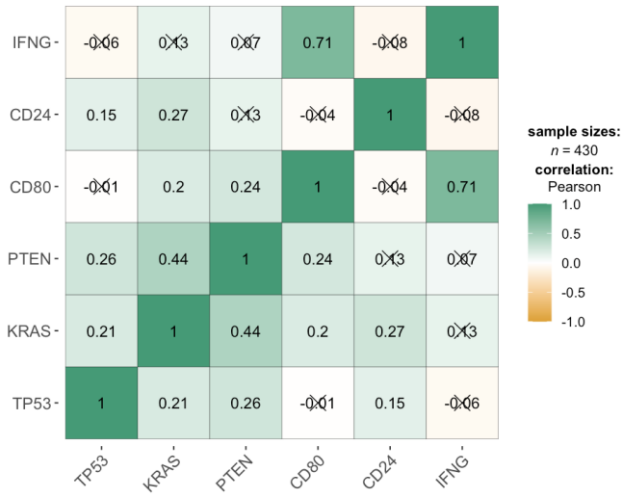
Figure S1. Major features provided by “Repository” page of Shiny application of *UCSCXenaShiny*. (A) Interface of “Repository” page. The table on the main panel shows current available datasets. The filters on the left panel provide functionality to filter datasets. The three buttons shown on the bottom of the main panel are used to operate datasets selected by users. (B) How to select specified datasets. The rows in dataset table are clickable. When a row is selected, metadata of the corresponding dataset will be queried from UCSC Xena data hubs, and the download link and the link to observe the dataset in UCSC Xena browser are provided. (C) When the button “Show Metadata” is clicked, all important metadata of datasets are shown. (D) When the button “Request Data” is clicked, links of datasets are shown and three buttons are provided to download datasets. (E) When the button “Copy R download code” is clicked, the code will be shown to tell users how to download the selected datasets by R. (F) When the button “Batch download in terminal” is clicked, a Shell script containing download code with wget command is automatically downloaded by the user browser so that the user can run this script in terminal. See **Figure S4** for analyzing data after clicking the button “Analyze Data”.

Figure S2

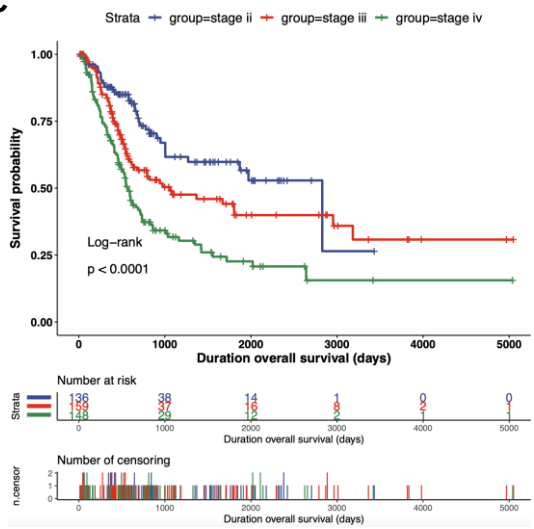
A



B



C



D

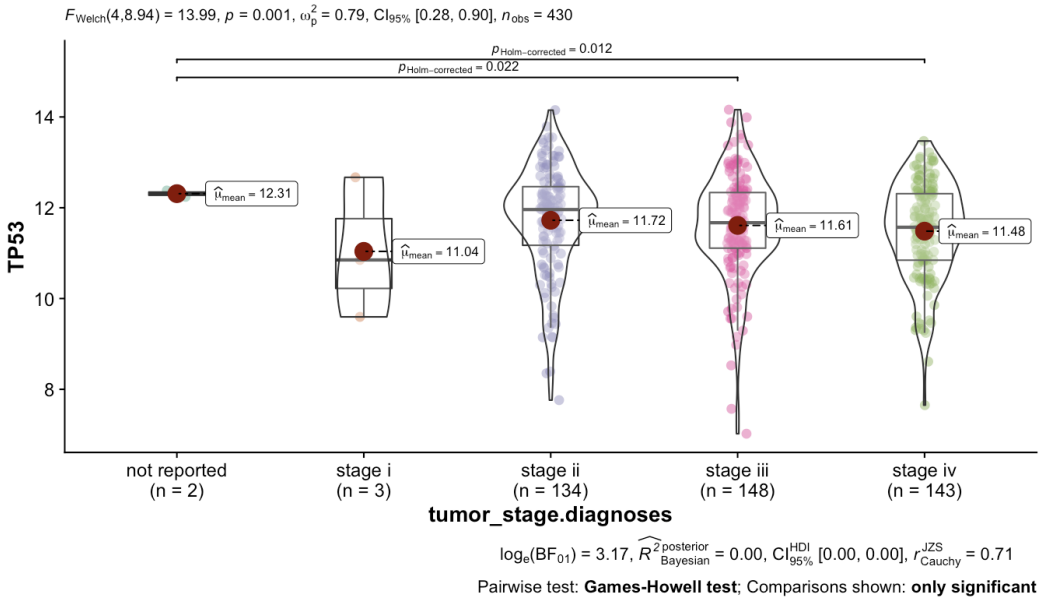


Figure S2. Features provided by “General Analysis” page of Shiny application of *UCSCXenaShiny*. Any dataset with proper format in UCSC Xena data hubs can be used for the current 4 supported analysis modules. (A) “Scatter-Correlation” module is used for exploring correlation between two continuous variables. The UI (user interface) widgets on the left and right sides of plot region are used for controlling analysis, visualization, sample filtering and result outputting. (B) “Matrix-Correlation” module is used for exploring correlation along multiple continuous variables. (C) “Survival-Analysis” module is used for exploring survival difference between sample groups. Here a continuous variable (e.g., mRNA expression of a gene) is supported to set sample groups by a threshold. (D) “Group-Comparison” module is used for exploring and visualizing value difference of a continuous variable between sample groups. The datasets used for demo analyses in this figure are selected shown as **Figure S3**.

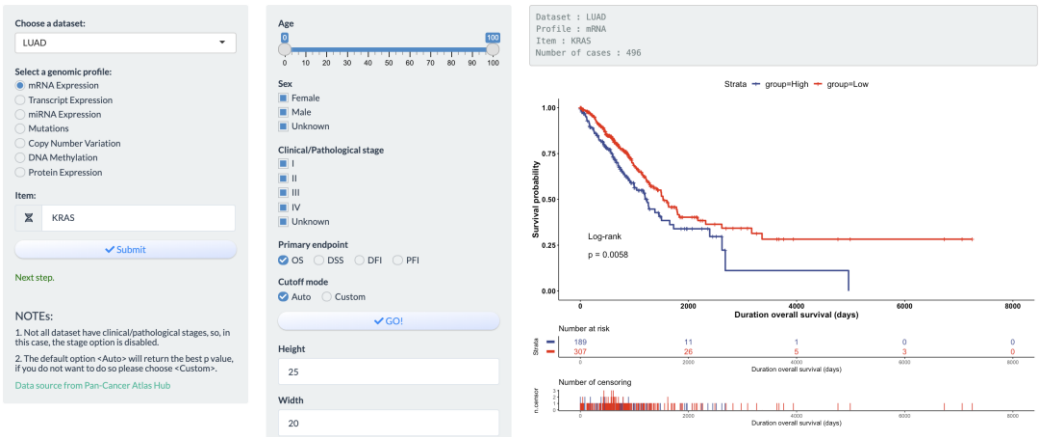
Figure S3


A
B
C
D
mRNA - TP53
TCGA: LUAD
Cor: 0.17 **
P53 protein
TP53 mRNA
TP53 mRNA stemness
Select a genomic profile:
mRNA Expression
Transcript Expression
DNA Methylation
Protein Expression
miRNA Expression
Copy Number Variation
Input a gene or formula (as signature)
TP53
Select a feature
stemness
Select Correlation method
spearman
Height
Width
Choose plot format
pdf
png
Download
NOTEs:
1. The data query may take some time based on your network. Wait until a plot shows
2. If a void plot shows, please check your input
3. Genomic profile data source
4. Check description of this link for TMB/Stemness/MSI data source

E


Select a genomic profile:
mRNA Expression
Transcript Expression
DNA Methylation
Protein Expression
miRNA Expression
Copy Number Variation
Input a gene or formula (as signature)
TP53
Select the immune signature source
Cibersort
Select Correlation method
spearman
Height
8
Width
12
Choose plot format
pdf
png
Download
The correlation between TP53 mRNA with immune signatures
NOTEs:
Genomic profile data source
Immune signature data source
Search:
gene
immune_cells
cor
p.value
cancer
pstar
ACC.B.cells.naive
TP53
B.cells.naive
0.0670416674071397
0.565009061987999
ACC

F


Choose a dataset:
LUAD
Select a genomic profile:
mRNA Expression
Transcript Expression
miRNA Expression
Mutations
Copy Number Variation
DNA Methylation
Protein Expression
Item:
KRAS
Submit
Next step.
NOTEs:
1. Not all dataset have clinical/pathological stages, so, in this case, the stage option is disabled.
2. The default option <Auto> will return the best p value, if you do not want to do so please choose <Custom>.
Data source from Pan-Cancer Atlas Hub
Age
Sex
Female
Male
Unknown
Clinical/Pathological stage
I
II
III
IV
Unknown
Primary endpoint
OS
DSS
DFI
PFI
Cutoff mode
Auto
Custom
GO!
Height
25
Width
20
Dataset : LUAD
Profile : mRNA
Item : KRAS
Number of cases : 496
Strata group=High group=Low
Survival probability
Log-rank
p = 0.0058
Duration overall survival (days)
Number at risk
Number of censoring

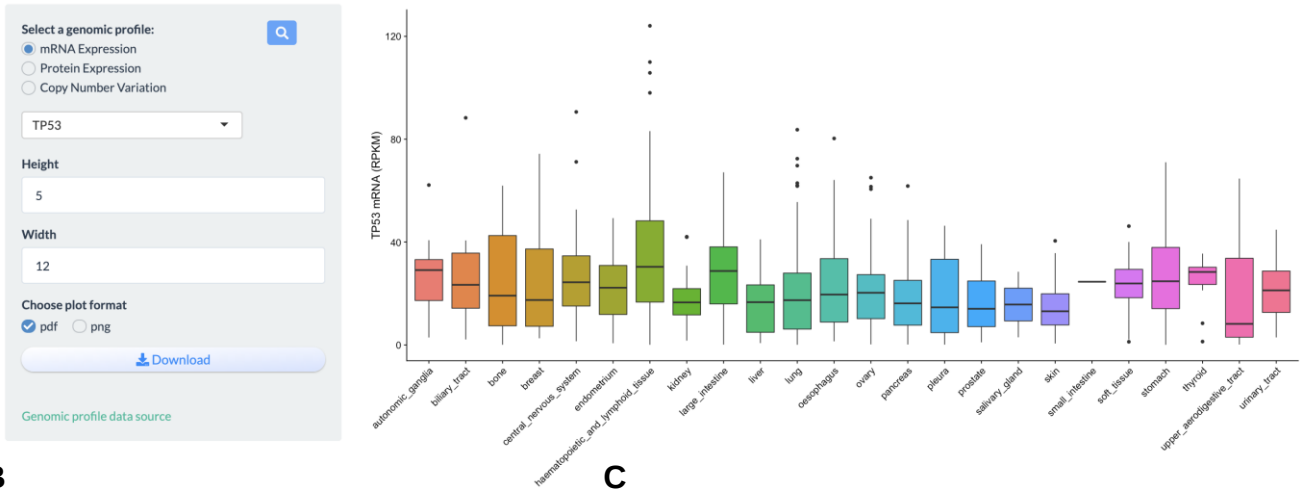
G


Select a genomic profile:
mRNA Expression
Transcript Expression
DNA Methylation
Protein Expression
miRNA Expression
Copy Number Variation
Input a gene or formula (as signature)
TP53
Select Measure for plot
OS
Select Threshold for plot
0.5
First color
#6A6F68
Second color
#E31A1C
Third color
#377DB8
TP53 Expression
Type
NS
Risky
log (Hazard Ratio)
NOTEs:
1. We define gene in certain cancer type as risky (log(Hazard Ratio) > 0) or protective (log(Hazard Ratio) < 0) or NS (No statistical significance, P value > 0.05)
2. We divide patients into different groups for comparison according to gene expression, you could choose the threshold for grouping (0.5 by default)
3. Genomic profile data source

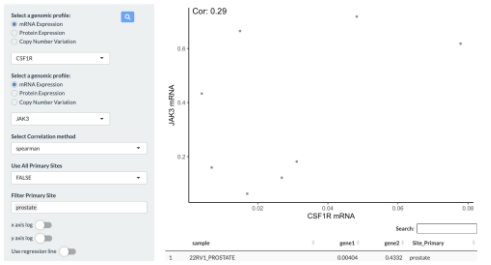
Figure S3. Analysis features with TCGA/GTEx data provided by “Quick PanCan Analysis” page of Shiny application of *UCSCXenaShiny*. The analysis modules contains panels for parameters and outputting setting, visualization and data presentation. (A) The module showing molecular profile distribution, value difference between tumor and normal tissue across TCGA cancer types. (B) The module showing molecular profile distribution, value difference (between tumor and normal tissue) across tumor tissue origins. (C) The module showing correlation between two molecular profiles. (D) The module showing association between molecular profile and tumor features including stemness, TMB (tumor mutational burden) and MSI (microsatellite instability). (E) The module showing association between molecular profile and tumor immune signatures or immune cell subsets. (F) The module showing survival difference between sample groups corresponding different molecular profile levels in TCGA pan-cancer or specified tumor type. In this sub-plot, TCGA LUAD cohort is selected. (G) The module showing association between survival risk and molecular profile level across TCGA cancer types with univariable Cox models. NOTE: Some analysis pages are too large to show all the details.

Figure S4

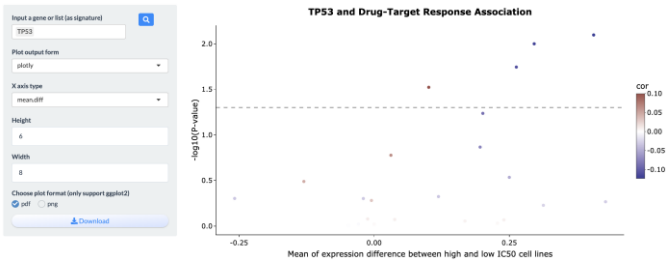
A



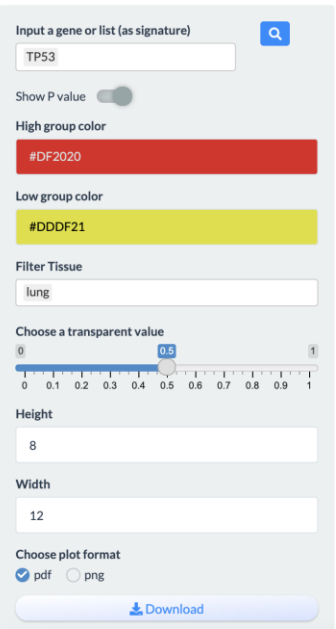
B

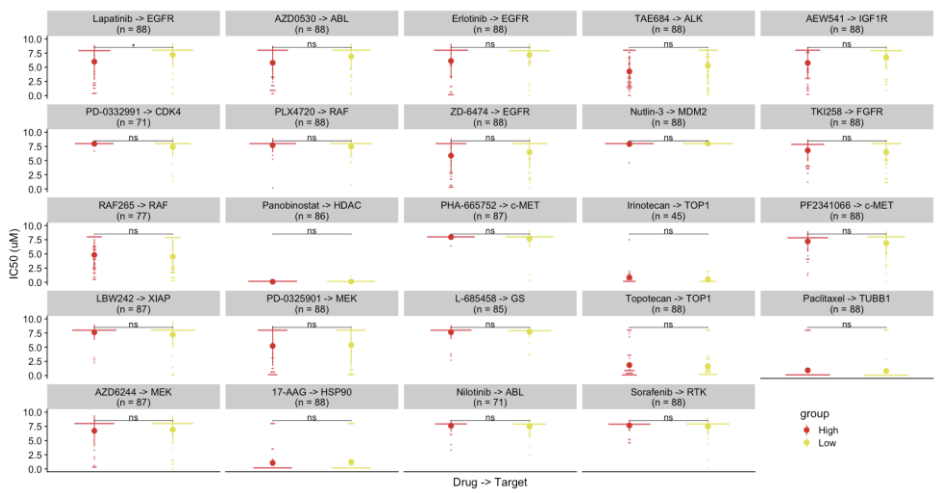


C



D





Method:

Analyze difference of drug response (IC50 value (uM)) between gene (or signature) high and low expression.

When there are multiple genes, geometrical mean of expression of these genes are used as a signature.

NOTEs: You can select multiple tissues, even ALL for all tissues. In this case 'number_of_cell_lines' indicates sample size for each gene-drug group instead of gene-drug-tissue group.

Figure S4. Analysis features with CCLE cancer cell line data provided by “Quick PanCan Analysis” page of Shiny application of *UCSCXenaShiny*. The analysis modules contains panels for parameters and outputting setting, visualization and data presentation. (A) The module showing molecular profile value distribution across CCLE cancer cell line tissue origins. (B) The module showing correlation between two molecular profiles. (C) The module showing association between gene mRNA expression and tumor drug response. (D) The module showing drug response IC50 value difference between low and high expression of a gene (use expression median as cutoff). NOTE: Some analysis pages are too large to show all the details.

Table S3

	Purpose/key feature	Downside
UCSCXenaShiny	An R package with both command caller and web GUI for interactively exploring UCSC Xena data.	It depends on R environment and has many dependent packages to support its analysis and visualization features.
UCSC Xena Browser	An online tool to explore functional genomic data sets for correlations between genomic and/or phenotypic variables.	1) It can not be used to download datasets in batch or download a dataset subset. 2) It has no specific analysis and visualization modules for the most commonly used pan-cancer (TCGA, PCAWG, etc.) cohorts.
UCSCXenaTools	An R API to access both a dataset and a dataset subset from UCSC Xena data hubs.	It has no analysis and visualization features for UCSC Xena datasets.
xenaPython	A Python API to access a dataset subset from UCSC Xena data hubs.	1) It can not be used to download a whole UCSC Xena dataset. 2) It has no analysis and visualization features for UCSC Xena datasets.
cBioPortal/GDC portal/ICGC portal	Web resources for exploring multidimensional cancer genomics data.	They mainly act as cancer genomics data providers. Their analysis and visualization features are limited.
CVCDAP	Seamless molecular and clinical analysis of a single TCGA cohort or comparison of two TCGA cohorts with relevance.	1) It cannot be used to download UCSC data. 2) Its analysis features are limited to TCGA datasets. 3) For user custom data, it only supports mutation analysis.

Table S3. Summary and comparison of the purpose/key feature and downside of UCSC Xena related tools (UCSCXenaShiny, UCSC Xena Browser, UCSCXenaTools, xenaPython) and other similar tools (cBioPortal, GDC portal, ICGC portal, CVCDAP).