

UCSCXenaShiny: an R package for exploring and analyzing UCSC Xena public datasets in web browser

Shixiang Wang
wangshx@shanghaitech.edu.cn

*

Yi Xiong
xiongyi123@csu.edu.cn

Kai Gu
gukai1212@163.com

Longfei Zhao
longfei8533@live.cn

Yin Li
yinli@openbio.org

Fei Zhao
zhaofei@openbio.org

Xuejun Li
lxjneuro@csu.edu.cn

Xue-Song Liu
liuxs@shanghaitech.edu.cn

†

2020-07-06

Abstract

Motivation: UCSC Xena platform provides huge amounts of processed cancer omics data from big public projects like TCGA or individual research groups for enabling unprecedented research opportunities. In 2019, we developed UCSCXenaTools, an R package for retrieval of UCSC Xena data. However, an easier dataset exploration and analysis tool is still lack, especially for researchers without programming experience.

Results: We develop UCSCXenaShiny, an R Shiny package to quickly explore, download all datasets from UCSC Xena data hubs. In addition, a module based analysis framework is constructed to analyze and visualize data.

Availability: <https://github.com/openbioX/UCSCXenaShiny> or <https://cran.r-project.org/package=UCSCXenaShiny>.

Contact: wangshx@shanghaitech.edu.cn or liuxs@shanghaitech.edu.cn.

Keywords: UCSC Xena; cancer genomics; TCGA

1 Introduction

Over the past decade, programs including TCGA [Weinstein et al., 2013], ICGC [Zhang et al., 2011], PCAWG [The et al., 2020], GTEx [Consortium et al., 2015], CCLE [Barretina et al., 2012] and etc. have generated large amounts of molecular data characterizing the landscape of more than ten thousands of tumors from genomic, epigenetic and proteomic aspects. The data have been preprocessed and stored at data hubs of UCSC Xena platform along with many public cancer datasets from individual research groups, providing unprecedented opportunities for either simple or systematic exploration of cancer behaviors and mechanisms at multiple molecular layers in individual cancer type or across cancer types [Goldman et al., 2019].

In 2019, we developed UCSCXenaTools, an open-source R package for retrieving and assembling public UCSC Xena data [Wang and Liu, 2019]. UCSCXenaTools was developed to communicate with UCSC Xena data hubs for downloading datasets or dataset subsets, querying metadata of data hub, cohort or dataset. Despite UCSC Xena platform itself allows users to explore and analyze data, it is hard for researchers to quickly explore all available datasets, locate what they need in their research and download useful datasets. Besides, the analysis features provided by UCSC Xena platform mainly focus on individual cohort data, thus lack of full-feature functionality.

To this end, we develop an open-source R Shiny package UCSCXenaShiny for cancer community to allow researchers to explore and analyze datasets from UCSC Xena data

*Corresponding author; Email: wangshx@shanghaitech.edu.cn

†Corresponding author; Email: liuxs@shanghaitech.edu.cn

hubs in web browser. In addition, an extensible module based analysis framework is constructed to analyze data. Currently, several modules providing single-gene expression analysis and visualization are implemented.

2 Materials and methods

2.1 Dataset exploration

UCSCXenaShiny opens a web page in user's browser to provide service. The page "Repository" is used to explore all available UCSC Xena datasets. Users can find desired datasets by either defined buttons or searching in dataset table. Once one or several datasets selected, users can query their metadata or download them (Fig.1 and Fig.3). To improve the performance of downloading large datasets, we provide a button to download a Shell script containing 'wget' commands which can run in Unix-like system.

2.2 Module and pipeline

For now, several modules targeting at single-gene expression analysis are available at page "Module" (Fig.2 and Fig.4), a pipeline based on them is available at page "Pipeline" (Fig.2). The usage is quite easy, users just need to type the gene symbol name and all procedures will be properly done by UCSCXenaShiny, including downloading data from UCSC Xena data hubs, cleaning data, analyzing data and visualizing the result. We are happy to accept new feature requests and they can be discussed at <https://github.com/openbioX/UCSCXenaShiny/issues>.

3 Results

3.1 Package structure

The structure and workflow of UCSCXenaShiny is described in Fig.1. Currently, the core components of this package are page "Repository" and page "Module". Page "Repository" allows researchers to explore and download datasets. Table 1 summaries the cohort and dataset number available at different UCSC Xena data hubs. There are total 1639 datasets and TCGA project is the major contributor. The development of UCSCXenaShiny is based on R Shiny platform (<https://shiny.rstudio.com/>), the overview of its graphic interface is shown in Fig.2.

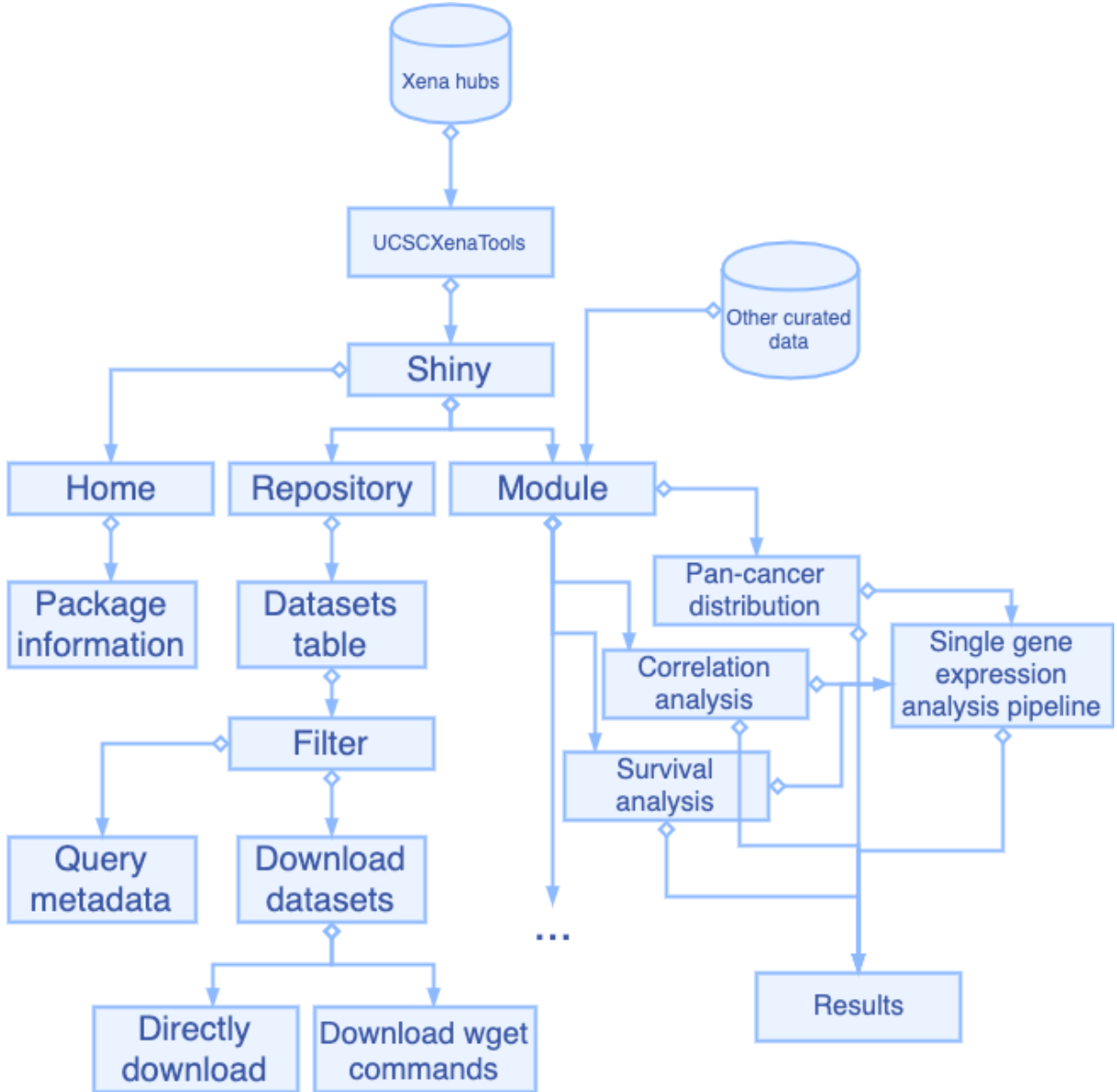


Figure 1: Package architecture and functional flowchart of UCSCXenaShiny.

3.2 Feature 1: dataset exploration and download

UCSCXenaShiny allows users to explore UCSC Xena datasets quickly and easily (Fig.3). A table storing all datasets is shown in Fig.3A, users can filter datasets by either typing some key words in search bar or selecting data hubs or data types. Once desired datasets are selected in the table, users can click the button on the bottom to check metadata of datasets or download datasets (Fig.3B-D).

3.3 Feature 2: Single-gene expression analysis

UCSCXenaShiny provides modules implementing basic analysis functionality and modules can be go further assembled as analysis pipeline (Fig.1). For example, we constructed

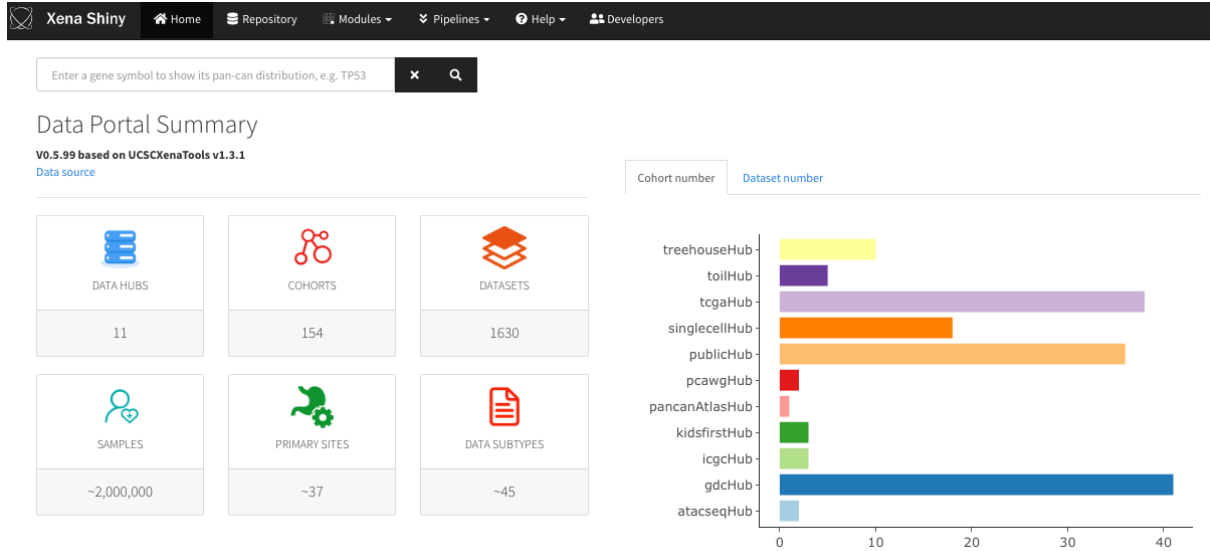


Figure 2: UCSCXenaShiny graphical interface overview.

	Data hub	Cohorts	Datasets	URL
1	tcgaHub	38	715	https://tcga.xenahubs.net
2	gdcHub	41	528	https://gdc.xenahubs.net
3	publicHub	36	109	https://ucscpublic.xenahubs.net
4	pcawgHub	2	53	https://pcawg.xenahubs.net
5	toilHub	5	50	https://toil.xenahubs.net
6	singlecellHub	18	54	https://singlecellnew.xenahubs.net
7	icgcHub	3	23	https://icgc.xenahubs.net
8	pancanAtlasHub	1	22	https://pancanatlas.xenahubs.net
9	treehouseHub	10	26	https://xena.treehouse.gi.ucsc.edu
10	atacseqHub	2	9	https://atacseq.xenahubs.net
11	kidsfirstHub	3	50	https://kidsfirst.xenahubs.net

Table 1: Summary of UCSC Xena data hubs.

a few modules to analyze and visualize the single gene expression, including its pan-cancer distribution with violin plot or anatomy heatmap [Maag, 2018], and survival effects [Terry M. Therneau and Patricia M. Grambsch, 2000] under different expression cutoff. We combined some of them and built single-gene expression analysis pipeline so researchers can get as much information as possible in one click for a same task view. An example for gene *TP53* is given in Fig.4.

Acknowledgements

We thank projects TCGA, GTEx, ICGC, CCLE, PCAWG, etc., and individual research groups for making cancer genomics data public. We thank UCSC Xena platform for providing data processing, integration and download.

References

- Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603, 2012.
- GTEx Consortium et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.
- Mary Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Akhil Kamath, Fran McDade, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. The ucsc xena platform for cancer genomics data visualization and interpretation. *BioRxiv*, page 326470, 2019.
- Jesper Maag. gganatogram: An r package for modular visualisation of anatograms and tissues based on ggplot2. *f1000research*, 2018. URL <https://f1000research.com/articles/7-1576/v1>. Version 1: Awaiting peer review.
- Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.
- ICGC The, TCGA Pan-Cancer Analysis of Whole, Genomes Consortium, et al. Pan-cancer analysis of whole genomes. *Nature*, 578(7793):82, 2020.
- Shixiang Wang and Xuesong Liu. The ucscxenatools r package: a toolkit for accessing genomics data from ucsc xena platform, from cancer multi-omics to single-cell rna-seq. *Journal of Open Source Software*, 4(40):1627, 2019.
- John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.
- Junjun Zhang, Joachim Baran, Anthony Cros, Jonathan M Guberman, Syed Haider, Jack Hsu, Yong Liang, Elena Rivkin, Jianxin Wang, Brett Whitty, et al. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, 2011, 2011.

A

Dataset Filters

Active Data Hub :

- ☐ UCSC Public
- ☐ TCGA
- ☒ GDC
- ☐ ICGC
- ☐ Pan-Cancer Atlas
- ☐ TOIL
- ☐ Treehouse
- ☐ PCAWG
- ☐ ATAC-seq
- ☐ Single Cell

Cohort Name :

e.g. Breast (separator is :)

Data Type :

- ☐ Phenotype
- ☐ Feature by sample matrix
- ☐ Genomic segments
- ☐ Mutations

Data Subtype :

e.g. gene expression (separator is :)

How to use repository

Show 10 entries

Dataset	Hub	Cohort	Samples	Subtype	Label	Unit
1 TCGA-BLCA.cnv.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	415	copy number	Copy Number Segment	log2(copy-number/2)
2 TCGA-BLCA.GDC_phenotype.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	454	phenotype	Phenotype	
3 TCGA-BLCA.gistic.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	413	copy number (gene-level)	GISTIC - focal score by gene	Gistic2 copy number
4 TCGA-BLCA.htseq_counts.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	430	gene expression RNAseq	HTSeq - Counts	log2(count+1)
5 TCGA-BLCA.htseq_fpkm.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	430	gene expression RNAseq	HTSeq - FPKM	log2(fpkm+1)
6 TCGA-BLCA.htseq_fpkm-uvq.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	430	gene expression RNAseq	HTSeq - FPKM-UQ	log2(fpkm-uvq+1)
7 TCGA-BLCA.masked_cnv.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	415	copy number	Masked Copy Number Segment	log2(copy-number/2)
8 TCGA-BLCA.methylation450.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	437	DNA methylation	Illumina Human Methylation 450	beta value
9 TCGA-BLCA.mirna.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	432	stem loop expression	miRNA Expression Quantification	log2(RPM+1)
10 TCGA-BLCA.muse_smv.tsv	gdcHub	GDC TCGA Bladder Cancer (BLCA)	411	somatic mutation (SNPs and small INDELs)	MuSE Variant Aggregation and Masking	

Showing 1 to 10 of 528 entries

Previous 1 2 3 4 5 ... 53 Next

Show Metadata **Request Data**

B

Submitted datasets:

	hosts	datasets	url
1	https://gdc.xenahubs.net	TCGA-BLCA.GDC_phenotype.tsv	https://gdc.xenahubs.net/download/TCGA-BLCA.GDC_phenotype.tsv.gz
2	https://gdc.xenahubs.net	TCGA-BLCA.htseq_counts.tsv	https://gdc.xenahubs.net/download/TCGA-BLCA.htseq_counts.tsv.gz

Download data directly **Batch download in terminal**

D

```
2020-07-06-commands.sh — Downloads
1 #!/usr/bin/env bash
2 #Usage: run bash 2020-07-06-commands.sh in your terminal under a desired
3 directory
4 wget -c https://gdc.xenahubs.net/download/TCGA-BLCA.GDC_phenotype.tsv.gz
5 wget -c https://gdc.xenahubs.net/download/TCGA-BLCA.htseq_counts.tsv.gz
```

Line: 1 Shell Script (Bash) Tab Size: 4

C

Please select a folder

Create new folder **Sort content** **Refresh**

home

Directories

- home
 - anaconda3
 - Applications
 - biodata
 - bioinfo
 - Desktop
 - Documents
 - Downloads
 - fsdownload
 - go
 - icmp4j
 - igv
 - Library
 - MessageBoard
 - Movies
 - Music
 - Nutstore Files
 - OneDrive - shanghaiitech.edu.cn
 - Pictures
 - Public
 - quicklisp
 - R-dev

Content

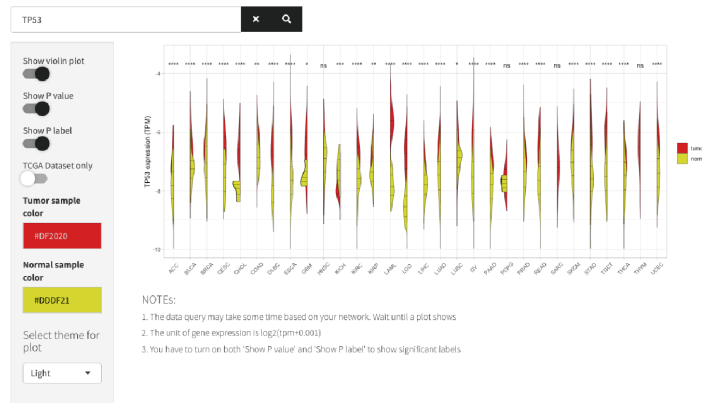
No folder selected

Cancel **Select**

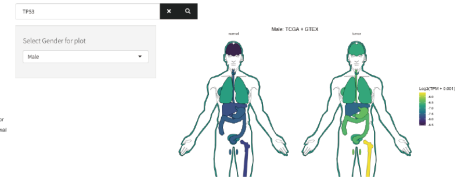
Figure 3: Dataset search and download.

A

Module: Gene Pancan Expression Distribution

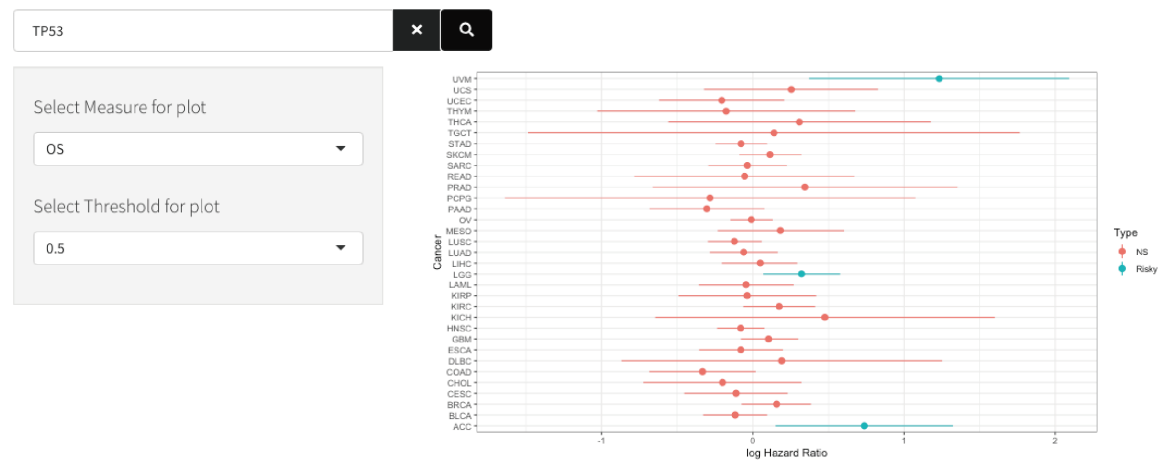


Module: Gene Pancan Expression Anatomy Visualization



B

Module: Gene Pancan Uni-cox analysis



C

Module: Surviva Analysis

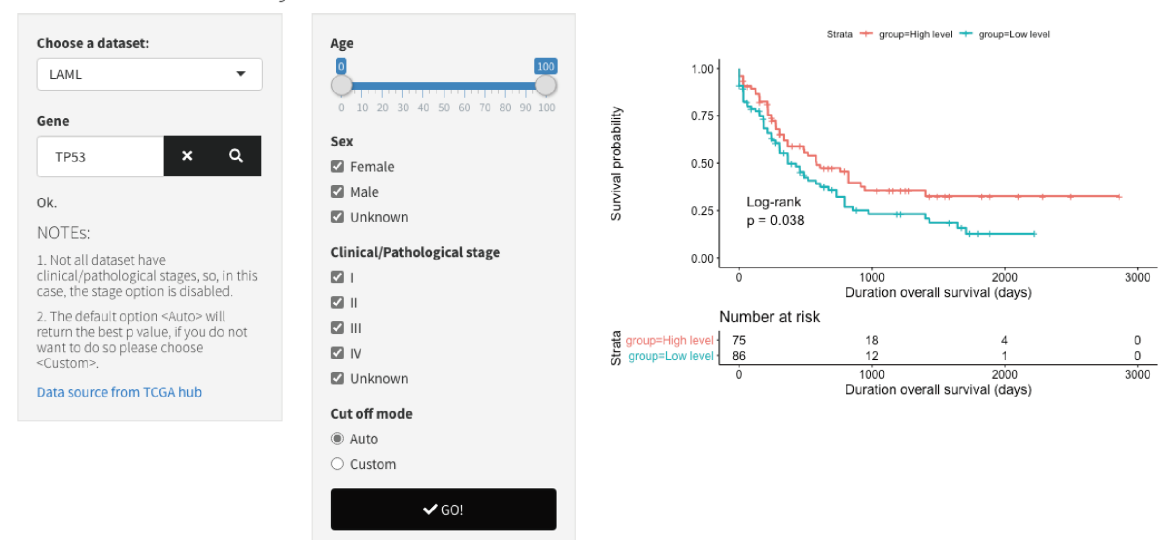


Figure 4: Current available analysis modules provided UCSCXenaShiny.