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@openbiox.org

Fei Zhao
zhaofei@openbiox.org

Xuejun Li
lxjneuro@csu.edu.cn

Xue-Song Liu
liuxs@shanghaitech.edu.cn
†

2020-07-05

Abstract

Motivation: UCSC Xena platform provides huge amounts of processed cancer omics data from big public projects like TCGA or individual reserach 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 addiction, 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 caner 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 addiction, 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). 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", a pipeline based on them is available at page "Pipeline" (Fig.1). 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

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. Page "Module" provides modules implementing basic analysis functionality and modules can be go further assembled as analysis pipeline. For example, we constructed a few modules to analyze the single gene expression, including its pan-cancer distribution and survival effects 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.

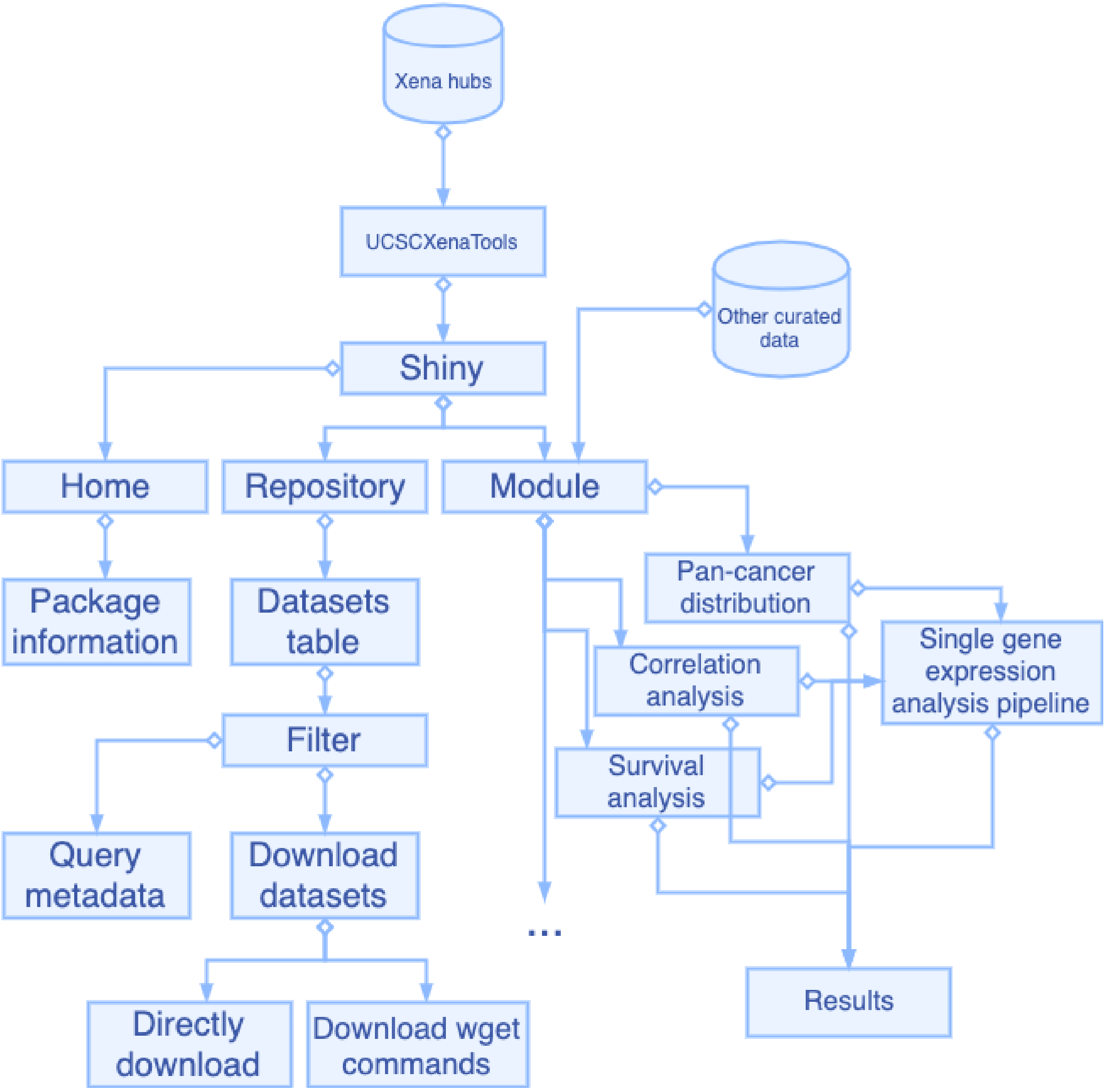

Figure 1: Package architecture and functional flowchart of UCSCXenaShiny

	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

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.

ICGC The, TCGA Pan-Cancer Analysis of Whole, Genomes Consortium, et al. Pancancer 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.