The Journal of Open Source Software

The UCSCXenaTools R package: a toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq

Shixiang Wang[1, 2, 3] **and Xuesong Liu**[1]

1 School of Life Science and Technology, ShanghaiTech University **2** Shanghai Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences **3** University of Chinese Academy of Sciences

DOI: 10.21105/joss.01627

Software

- Review
- Repository
- Archive

Submitted: 02 August 2019
Published: 05 August 2019

Summary

UCSC Xena platform (https://xenabrowser.net/) provides unprecedented resource for public omics data (Goldman et al., 2019) from big projects like The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013), International Cancer Genome Consortium Data Portal (ICGC) (Zhang et al., 2011), The Cancer Cell Line Encyclopedia (CCLE) (Barretina et al., 2012), or reserach groups like Mullighan et al. (2008), Puram et al. (2017). All available data types include single-nucleotide variants (SNVs), small insertions and deletions (INDELs), large structural variants, copy number variation (CNV), expression, DNA methylation, ATAC-seq signals, and phenotypic annotations.

Despite UCSC Xena platform itself allows users to explore and analyze data, it is hard for users to incorporate multiple datasets or data types, integrate the selected data with popular analysis tools or homebrewed code, and reproduce analysis procedures. R language is well established and extensively used standard in statistical and bioinformatics research. Here, we introduce an R package UCSCXenaTools for enabling data retrieval, analysis integration and reproducible research for omics data from UCSC Xena platform.

Currently, UCSCXenaTools supports downloading over 1600 datasets from 10 data hubs of UCSC Xena platform as shown in the following table. Typically, downloading UCSC Xena datasets and loading them into R by UCSCXenaTools is a workflow with generate, filter, query, download and prepare 5 steps, which are implemented as functions. They are very clear and easy to use and combine with other packages like dplyr (Wickham, Francois, Henry, Müller, & others, 2015). Besides, UCSCXenaTools can also query and download subset of a target dataset, this is particularly useful when user focus on studying one object like gene or protein. The key features are summarized in Figure 1.

Data hub	Dataset count	URL
tcgaHub	879	https://tcga.xenahubs.net
gdcHub	449	https://gdc.xenahubs.net
publicHub	104	https://ucscpublic.xenahubs.net
pcawgHub	53	https://pcawg.xenahubs.net
toilHub	50	https://toil.xenahubs.net
singlecellHub	45	https://singlecell.xenahubs.net
icgcHub	23	https://icgc.xenahubs.net
pancanAtlasHub	19	https://pancanatlas.xenahubs.net
treehouseHub	15	https://xena.treehouse.gi.ucsc.edu
atacseqHub	9	https://atacseq.xenahubs.net

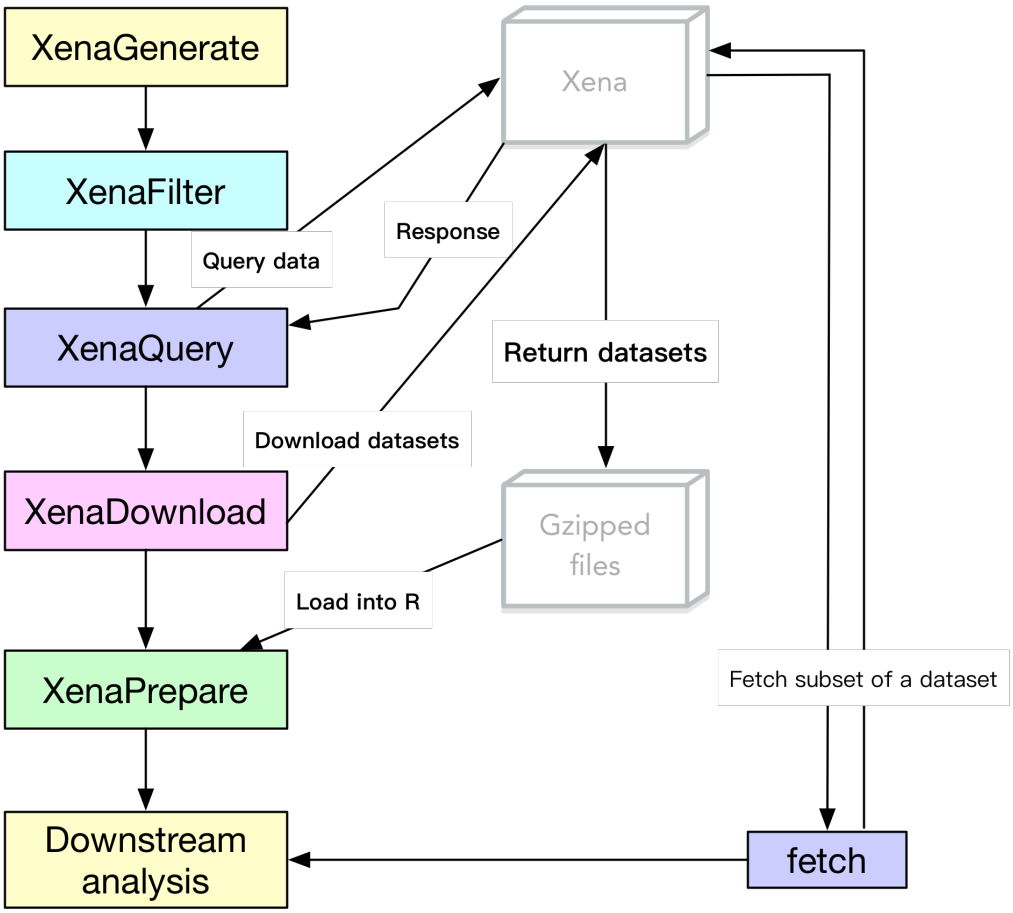

Figure 1: Overview of UCSCXenaTools

Acknowledgements

We thank Christine Stawitz and Carl Ganz for their constructive comments. This package is based on R package XenaR, thanks Martin Morgan for his work.

References

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., et al. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), 603.

Goldman, M., Craft, B., Hastie, M., Repečka, K., Kamath, A., McDade, F., Rogers, D., et al. (2019). The ucsc xena platform for cancer genomics data visualization and interpretation. *BioRxiv*, 326470.

Mullighan, C. G., Phillips, L. A., Su, X., Ma, J., Miller, C. B., Shurtleff, S. A., & Downing, J. R. (2008). Genomic analysis of the clonal origins of relapsed acute lymphoblastic leukemia. *Science*, *322*(5906), 1377–1380.

Puram, S. V., Tirosh, I., Parikh, A. S., Patel, A. P., Yizhak, K., Gillespie, S., Rodman, C., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, *171*(7), 1611–1624.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, *45*(10), 1113.

Wickham, H., Francois, R., Henry, L., Müller, K., & others. (2015). Dplyr: A grammar of data manipulation. *R package version 0.4*, *3*.

Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., Liang, Y., et al. (2011). International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database*, *2011*.