TCGA Methods - Final Project Summary

Margaret Hannum (rh2623)
P8119 - Advanced Statistical and Computational Methods in Genetics and Genomics

April 7, 2017

Background: The Cancer Genome Project (TCGA)

TCGA is a large collaborative effort led by National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI).

The Project Mission involved three main steps: 1. Large scale high-quality tissue sampling 2. DNA sequencing and analysis (Genome Data Analysis Centers) 3. Public data sharing (GDC Data Portal). There are 3 levels of data to protect subject identities. Level 3 is metadata available to anyone.

The project itself ran from 2007-2017. The data is publicly available online and has contributed to more than a thousand cancer studies. New studies using the data continue to come out.

Using TCGA Data: Methods and Complications

NCI's Genomic Data Commons is a collection of validated data anyone can query and download. All TCGA data is housed there, in addition to newer and ongoing genetic repository projects. cBioPortal for Cancer Genomics, which is housed by Memorial Sloan Kettering, provides visualization, analysis and download of large-scale cancer genomics data sets.

Still, there is such a huge amount of data and a lack of tools to easily acquire and assemble it. TCGA includes 2.5 petabytes of comprehensive whole genome and whole exome maps for 33 disease types, collected from 11,000 patients. Navigating the shear amount of data can be overwhelming for researchers.

TCGA-Assembler

This is where TCGA-Assembler comes in. TCGA-Assembler is an R package that "streamlines retrieval, assembly, and processing" of public TCGA data. It was originally published in *Nature* in 2014 by Yitan Zhu, Peng Qiu, and Yuan Ji (University of Chicago). Version 2.01 was released in January 2017. The package involves two modules, the first of which extracts URLs of all data files, and then downloads user-specified data files and assembles

them into usable matrices. The second module is for processing the data (correcting gene symbols, removing redundancies, combining multimodal and multi-platform data).

Using TCGA-Assembler

TCGA-Assembler v 2.01 can be downloaded online. (You must register and an email will be sent to you before you can download. Recommend using personal email like gmail.) R and R packages including RCurl, rjson, httr, stringr, HGNChelper, and bitops also need to be installed. When you open R, set working directory as TCGA folder. Finally, source to Module A for data downloading functions, and Module B for for data processing functions. You can do further data processing such as merging multi-platform data down the line.

Comments

While TCGA-Assembler may be a tool that streamlines the download and assembly of TCGA data, it is by no means easy or intuitive. It still requires a level of comfort with R as well as with TCGA data itself. However, for researchers and statisticians who do not have or are not able to write their own scripts to process these large, multi-modal datasets, it may be a good investment of time to learn how to use this package to get over at least one hurdle of working with TCGA data.

There are also many other methods to explore when using TCGA data, such as Web-TCGA, Firehose Pipeline (Broad Institute), FirebrowseR (an R client to Firehose), and tools to visualize genomic data such as Gviz, Bioconductor, and cBioPortal.

References

1. Y. Zhu, P. Qiu, Y. Ji. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data.Nature Methods. Vol. 11, No. 6, pp:599-600, 2014. — doi:10.1038/nmeth.2956

2. Wang, Z., Jensen, M. A., & Zenklusen, J. C. (2016). A Practical Guide to The Cancer Genome Atlas (TCGA) . In Statistical Genomics: Methods and Protocols (Vol. 1418, Methods in Molecular Biology, pp. 111-141). New York, NY: Springer.

3. The future of cancer genomics. (2015). Nature Medicine, 21(2), 99-99. doi:10.1038/nm.3801

4. Tomczak, K., Czerwinska, P., and Wiznerowicz, M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology. 19(1A): A68-A77.