R package development to analyze the cancer genome atlas data: a study case based on hypoxia induced factor-a3 isoforms

Fábio Malta de Sá Patroni[1], Douglas Adamoski[2], Marcelo Falsarella Carazzolle[3], Sandra Martha Gomes Dias[4],

1 Graduate Program in Genetics and Molecular Biology, Institute of Biology University of Campinas, Campinas
2 Graduate Program in Genetics and Molecular Biology, Institute of Biology University of Campinas
3 Biology Institute - UNICAMP, National Center for High Performance Computing/Unicamp
4 Brazilian National Center for Research in Energy and Materials, Brazilian Biosciences National Laboratory

Abstract

A great magnitude of information on multi-resource omics data is being created and made freely available through the project The Cancer Genome Atlas (TCGA). Although the amount of data rises every year, data mining tools are not following at the same pace. The efficient use of this information, also called the "big data", has the potential to unveil new observations and mechanisms that can impact on cancer treatment. TCGA data are stored at the GDC Data Portal and GDC Legacy Archive, both of which hosted by the US Nacional Cancer Institute (NCI). The totality of the data comprehends genomic, transcriptomic, proteomic and metilome information from 33 types of cancer. Given the complexity and extension of the available data, new analytical tools are necessary to automate and facilitate the data mining process. R programming language is being widely used for dealing with "big data". This work has two main goals: First, to create an R package aiming at to download, organize, analyze and report TCGA data; second, apply the package to identify potential downstream targets of the transcriptional factor HIF-3a isoforms. HIF regulates the expression of genes as a response to hypoxia and is an important player on the tumor metabolism adaptation process. Our R package, GDCRtools (version: 0.0.9) was developed and used to analyze the HIF3a2 isoform in four tumor types: Ovarian serous cystadenocarcinoma [OV], Testicular Germ Cell Tumors [TGCT], Uterine Carcinosarcoma [UCS] and Stomach adenocarcinoma [STAD].Tumors were divided among higher and lower HIF3a2 expression, and differential gene expression determined. Gene Ontology and Reactome was employed for pathway enrichment analysis and revealed enriched terms related with extracellular matrix organization, blood vessel development and GPCR downstream signaling, potentially linking this isoform with these processes.

Funding: This work was supported by grants from São Paulo Research Foundation (2015/26059-7)