In silico molecular subtype-specific drug targets prospection by integrating colorectal cancer and tumor-derived cell lines data

Cristóvão Antunes de Lanna[1], Nicole Scherer[2], Luís Felipe Ribeiro Pinto[3], Mariana Boroni[2]

1 LABORATÓRIO DE BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL [LBBC/INCA]
2 LABORATÓRIO DE BIOINFORMÁTICA E BIOLOGIA COMPUTACIONAL
3 PROGRAMA DE CARCINOGÊNESE MOLECULAR

Abstract

Colorectal cancer (CRC) is the third most prevalent carcinoma in the world. The molecular basis of CRC progression is well known and the vast amount of available data has allowed the classification of CRC tumors into molecular subtypes. However, many classification systems have been developed independently and they are generally inconsistent. Recently, the Colorectal Cancer Subtyping Consortium (CRCSC), an initiative involving several independent groups, has identified four consensus molecular subtypes (CMS) based on gene expression data from more than 4,000 primary tumor samples. Since cell lines are frequently used as in vitro tumor models, our study aims to classify CRC cell lines into their respective CMS and use them as platform to discovery/validate potentially new drug targets. In the present study we analysed a total of 155 CRC-derived cell lines using both Loess-normalized microarray data obtained from the GEO database and RNA-seq data from the SRA database. For RNA-seq data, after quality control and low-quality bases removal using FastQC and Trimmomatic, respectively, reads were aligned to the human genome and sorted by coordinate using STAR. After that, mapped reads for each gene were quantified and gene expression levels were normalized using RSEM. Additionally, raw RNA-seq aligned read count data for The Cancer Genome Atlas (TCGA) CRC primary tumor samples was downloaded using the TCGABiolinks tool and normalized using the DESeq2 tool, both written in R. Normalized counts from all sample sets were used to classify each sample using the CMSClassifier tool, written in R and developed by CRCSC. This tool allows us to make a molecular classification of CRC samples using high-throughput expression data (RNA-seq or microarray) based on the random forest method. Mutation Allele Frequency (MAF) files were also obtained from TCGA and separated by the samples' CMS classification. A list containing overlapping frequently mutated and differentially expressed genes will be used for metabolic pathway enrichment of each CMS using tools such as MetaCore, David and Reactome. Pathway components will be used for building interaction networks using String. Features containing the most connections will be used as candidates for novel drug target discovery using the Integrity database. Identified targets will then be compared with CMS representative cell lines' targets identified in the CancerRXGene database in order to compare cell lines' and primary tumors' potential treatment responses. First results include sample classification, which for the cell lines is as follows: CMS1, 42 cell lines (27%); CMS2, 62 (40%); CMS3, 10 (6%); CMS4, 2 (1%); undetermined, 39 (25%). For TCGA samples, classification distribution is: CMS1, 83 samples (13%); CMS2, 349 (54%); CMS3, 51 (8%); CMS4, 60 (9%); undetermined, 104 (16%).

Differentially expressed and mutated genes analysis is in progress. The identification of CRC cell lines representing each CMS is important to assess the functional consequences of genome-aberration-mediated gene deregulation of each CRC subtype. Representative cell lines of each subtype will be used to validate the molecular features that predict resistance/sensitivity to agents that target such aberrations. This will improve our comprehension of the mechanisms of tumorigenesis and foster the development of new therapies aiming to selectively interfere with one or more of these processes.

Funding: CAPES, INCA, Ministério da Saúde