DriverGenePathway: an R package for detecting cancer driver gene sets based on multiple statistical hypothesis tests and adaptively weighted mathematical programming

Zeyuan Wang

2023-04

Getting started

Installing the package.

To install the *DriverGenePathway* package, first download the installation package from site https://github.com/FrancisWang96/DriverGenePathway. For Windows, start R and select the `Packages` menu, then `Install package from local zip file`. Find and highlight the location of the zip file and click on `open`. For Linux/Unix, use the usual command `R CMD INSTALL` or install from bioconductor.

Loading the package.

To load the `DriverGenePathway` package in your R session, type `library(DriverGenePathway)`.

Help files.

Detailed information on `DriverGenePathway` package functions can be obtained in the help files. For example, to view the help file for the function `DriverGenePathway` in a R session, use `?DriverGenePathway`.

Introduction

The main purpose of DriverGenePathway is to identify significantly important mutated genes and gene sets (pathways) that are responsible for cancer, called driver genes and driver pathways, thus this package contains two main functions, `DriverGenes` and `DriverPathway`, aimming respectively at searching driver genes and driver pathways. For `DriverGenes`, we provided 5 methods of hypothesis test. The process mainly consists of three sections. First the mutation and coverage data are preprocessed after guaranteed available. Then with covariate data, the background mutation rate for each gene is calculated. Finally the significant genes are discovered via hypothesis test methods. Regarding `DriverPathway`, a de novo method is employed to search the driver pathway. It models an optimal submatrix function based on coverage and mutual exclusivity, which are basic characteristics of driver pathways. In addition, the submatrix optimization problem (a quadratic programming problem) is solved by genetic algorithm.

Functions

DriverGenes()

To run `DriverGenes`, 3 input data sets and a directory of chromosome files are required. The first data set is mutation MAF(Mutation Annotation Format) of a paticular cancer type, containing information of mutations. The second is coverage data, containing information of coverages. The third data set is covariate data, which contains values of covariates, and is used for background mutation rate discovery. In addition, the chromosome files directory can be either hg19 or hg38.

```
library(DriverGenePathway)
laml_maf <- system.file("extdata", "tcga_laml.maf", package = "DriverGenePathway")
coverage <- system.file("extdata", "coverage.rda", package = "DriverGenePathway")
covariate <- system.file("extdata", "gene.covariates.txt", package = "DriverGenePathway")
load(coverage)
bmr_result <- backgroundMutationRate(laml_maf, coverage, covariate, original_mutation_rate = 1.2e-6,
                                     max_neighbors = 50, ref_genome = NULL, category_num = 1,
                                     output_file = TRUE, quiet = FALSE)
driverGenes(bmr_result, p_class = "betaBinomial", output_file = TRUE, filter = TRUE,
            sigThreshold = 0.1, quiet = FALSE)
```

DriverPathway()

The input of `DriverPathway` can either be MAF or mutation matrix.

```
library(DriverGenePathway)
mutation_matrix <- system.file("extdata", "sample_mutation_matrix.rda",
package = "DriverGenePathway")
load(mutation_matrix)
driverPathway(mutation_matrix, driver_size = 3, pop_size = 30, iters = 100, permut_time = 50)
```