TCGA-Assembler Version 2.0 User Manual

May 30, 2018

Introduction

TCGA-Assembler can acquire and process The Cancer Genome Atlas (TCGA) public data including level-3 data of RNA-seq, miRNA-seq, DNA methylation, DNA copy number, and protein expression, and somatic mutation data, as well as de-identified patient clinical information. The TCGA data are maintained by the Genomic Data Commons (GDC). TCGA-Assembler can also acquire and process mass spectrometry proteomics data of TCGA samples from the Clinical Proteomic Tumor Analysis Consortium (CPTAC). TCGA-Assembler is open-source and freely available at http://www.compgenome.org/TCGA-Assembler/. Click on Download Software and unzip the downloaded file to your desired location. The folder of the package is TCGA-Assembler.

The TCGA-Assembler package includes two modules, Module A and Module B.

Module_A.R includes all Module A functions. Module A downloads data from the GDC TCGA data portal and CPTAC. Each patient sample in TCGA has one or multiple data files produced by each assay platform. Module A assembles the data files of individual samples into data matrices, most of them has the format with genomic features as rows and samples as columns.

Module_B.R includes all Module B functions. Module B processes data matrix files to fulfill various data manipulation needs. Module B can process both data matrix files assembled by TCGA-Assembler Module A and data matrix files downloaded from Firehose website at the Broad Institute (https://confluence.broadinstitute.org/display/GDAC/Dashboard-Stddata).

QuickStartExample.R includes the example code running a complete process of using TCGA-Assembler to acquire, process, and manipulate TCGA data.

SupportingFiles folder includes supporting files and annotation files used by Module B functions.

ManualExampleData folder holds some data files either acquired/processed by TCGA-Assembler or downloaded from Firehose website, which will be used by the examples in this user manual. There are four subfolders in the ManualExampleData folder. The RawData.TCGA-Assembler subfolder is used to hold data downloaded from GDC TCGA data portal by TCGA-Assembler. The ProcessedData.TCGA-Assembler subfolder is used to contain results obtained by processing the raw data files in the RawData.TCGA-Assembler subfolder. The RawData.Firehose subfolder contains data files downloaded from the Firehose website. We keep only several samples in the Firehose data files to make their sizes small, so that they can be included in the package, and the data formats are not changed. The ProcessedData.Firehose subfolder is used to contain results obtained by processing the Firehose data files in the RawData.Firehose subfolder.

Because downloading and processing TCGA data may require significant memory space depending on the size of data, we recommend using TCGA-Assembler on computers with 8GB or larger RAM.

To use TCGA-Assembler, R and R packages including RCurl, rjson, httr, stringr, HGNChelper, and their dependent packages need to be installed. They are freely available from http://www.r-project.org/.

General Procedure of Using TCGA-Assembler Pipeline

Step 1: Start R, and set the TCGA-Assembler folder (i.e. the package folder) as the Present Working Directory (PWD) of R.

Step 2: Source Module_A.R to load Module A functions into your workspace. Use data downloading functions whose names start with "Download" to (1) acquire TCGA public data of specified cancer type(s) and assay platform and (2) assemble the data files of individual samples into data matrices in tab-delimited .txt data files.

Step 3: Source Module_B.R to load Module B functions into your workspace. The data matrix .txt files downloaded by Module A should be processed by corresponding Module B functions, whose names start with "Process". These functions extract useful measurements from the .txt files and import the data into R. Data quality check, removal of redundant information, and some data calculations, such as calculating gene-level copy number values, are fulfilled by these functions.

Step 4: Based on the processed data generated in Step 3, use other data processing functions in Module B to do various data manipulations, such as merging multi-platform data into a single mega data table.

Notices

1. Users who run TCGA-Assembler on ***Windows*** operating system need to make curl available as a system command. The easiest way to do so is to copy curl.exe in TCGA-Assembler package to C:\Windows\System32 directory. You can also download the curl command supporting SSL and SSH and applicable to your operating system from https://curl.haxx.se/download.html. Mac and Linux users do not need to do this, because curl is already included as a system command.

2. ***Windows*** operating system usually has a limitation on the length of file path, which is 260 characters. TCGA data files usually have a long file name and folder name. So the downloaded data files may have paths (including both the full directory and file name) longer than the limitation, causing failure when writing the data files to your local hard disk. If you see something similar to the following messages in your R console, it most likely indicates a failure caused by the limitation on file path length, and the program can not save the data files and keeps retrying.

   ```
   [1] "metadata file: preparing ..."
   [1] "metadata file: preparing done!"
   [1] "*.tar.gz file: downloading & unzipping ..."
   [1] "cannot open the connection"
   [1] "cannot open the connection"
   [1] "cannot open the connection"
   ...
   ```

   To solve this problem, either put TCGA-Assembler package in a directory with a short path (such as the root directory C:\) or set your operating system to allow long file path. You can Google on the internet for solutions about how to configure your Windows operating system to allow long file path. Mac and Linux users usually do not encounter this problem.

3. Data downloading functions in Module A, such as DownloadCNAData, use a temporary folder to organize the temporary data files downloaded. The temporary folder name will be "tmp_YYYYMMDDhhmmss", where YYYY, MM, DD, hh, mm, and ss are year, month, date, hour, minute, and second, respectively. The temporary folder will be removed after the data downloading function is successfully executed. If the function execution is interrupted (for example due to internet connection problem), the temporary folder (including files in it) will not be automatically removed and users need to remove it manually.

Contents

Module A includes the following functions:

Module B includes the following functions:

DownloadBiospecimenClinicalData

Usage

```
DownloadBiospecimenClinicalData(cancerType, saveFolderName = "./BiospecimenClinicalData",
    outputFileName = "")
```

Description

This function downloads all biospecimen and clinical data files of user-specified cancer type and save them in the specified folder.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is "./BiospecimenClinicalData".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

Details

This function retrieves the biospecimen and clinical data files, which are tab-delimited text files whose file type is called biotab. In the data files, patients that fit within the clinical data types of interest are indicated by TCGA barcodes. For information about the biospecimen and clinical data formats, please refer to https://wiki.nci.nih.gov/display/TCGA/Biotab. Please refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode, for information about TCGA patient barcode. The downloaded data files have names composed of outputFileName and their original filenames, with "__" separating the two. If outputFileName is an empty string, the names of the downloaded data files are the same as their original TCGA filenames.

Table 1. Cancer Types Abbreviation

cancerType	Cancer Type Full Name
ACC	Adrenocortical carcinoma
BLCA	Bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma

CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
DLBC	Lymphoid neoplasm diffuse large B-cell lymphoma
ESCA	Esophageal carcinoma
GBM	Glioblastoma multiforme
HNSC	Head and neck squamous cell carcinoma
KICH	Kidney chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute myeloid Leukemia
LGG	Brain lower grade glioma
LIHC	Liver hepatocellular carcinoma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MESO	Mesothelioma
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCPG	Pheochromocytoma and paraganglioma
PRAD	Prostate adenocarcinoma
READ	Rectum adenocarcinoma
SARC	Sarcoma
SKCM	Skin cutaneous melanoma
STAD	Stomach adenocarcinoma
TGCT	Testicular germ cell tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine corpus endometrial carcinoma
UCS	Uterine carcinosarcoma
UVM	Uveal melanoma

Value

A character vector of the paths of downloaded biospecimen and clinical files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

source("Module_A.R")   # Load Module A functions

# Acquire biospecimen and clinical information of breast invasive carcinoma (BRCA) patients, save data files in the specified directory and add the prefix word "test" to the filenames.

filename_biosClin <- DownloadBiospecimenClinicalData(cancerType = "BRCA", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler/BiospecimenClinicalData", outputFileName = "test")
```

DownloadCNAData

Usage

```
DownloadCNAData(cancerType, assayPlatform = NULL, tissueType = NULL, saveFolderName = ".",
    outputFileName = "", inputPatientIDs = NULL)
```

Description

This function downloads copy number data of samples belonging to the user-specified cancer type and tissue type(s), measured by the specified assay platform, and then combines them into tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can be one or a combination of cna_cnv.hg18, cna_cnv.hg19, cna_nocnv.hg18, and cna_nocnv.hg19. Its default value is NULL, which indicates all above assay platforms (if available). Table 3 shows the description of assay platforms.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all above tissue types if available. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data from all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

Details

For each different assayPlatform, this function generates a tab-delimited .txt data file. The filename consists of five components: (1) outputFileName; (2) cancer type; (3) assay platform used to generate the data; (4) specified tissue types connected by "_" or "tissueTypeAll" that indicates all available tissue types; (5) the date and time when the data were downloaded. Double-underscore "__" is used to separate the five components in the filename. If outputFileName is an empty string, the filenames consist of only the other four components.

All data files have the same format. The 1st row includes column names, while each other row corresponds to a DNA segment. The 1st column is TCGA barcode. The 2nd column is chromosome ID. The 3rd column is the start position of the segment. The 4th column is the end position of the segment. The 5th column is the number of probes in the segment. The 6th column is the copy number value transferred by base2 log(copyNumber/2), centered on 0. For more details of the data format, data type, and data generation pipeline, please refer to https://wiki.nci.nih.gov/display/TCGA/SNP+array-based+data.

Table 2. Tissue Types

tissueType	Category	Tissue Type	TCGA_SampleTypeId
TP	Tumor	Primary Solid Tumor	01
TR	Tumor	Recurrent Solid Tumor	02
TB	Tumor	Primary Blood Derived Cancer - Peripheral Blood	03
TRBM	Tumor	Recurrent Blood Derived Cancer - Bone Marrow	04
TAP	Tumor	Additional - New Primary	05
TM	Tumor	Metastatic	06
TAM	Tumor	Additional Metastatic	07
THOC	Tumor	Human Tumor Original Cells	08
TBM	Tumor	Primary Blood Derived Cancer - Bone Marrow	09
NB	Normal	Blood Derived Normal	10
NT	Normal	Solid Tissue Normal	11
NBC	Normal	Buccal Cell Normal	12
NEBV	Normal	EBV Immortalized Normal	13
NBM	Normal	Bone Marrow Normal	14

Table 3. Assay platforms for function DownloadCNAData

assayPlatform	Assay	Reference Genome	Additional Description	Filename Pattern
cna_cnv.hg18	Affymetrix SNP Array 6.0	Hg18	All probes are included in measurements	.hg18.seg.txt
cna_cnv.hg19	Affymetrix SNP Array 6.0	Hg19	All probes are included in measurements	.hg19.seg.txt
cna_nocnv.hg18	Affymetrix SNP Array 6.0	Hg18	Probes frequently containing germline CNVs are excluded	.nocnv_hg18.seg.txt
cna_nocnv.hg19	Affymetrix SNP Array 6.0	Hg19	Probes frequently containing germline CNVs are excluded	.nocnv_hg19.seg.txt

Value

A character vector of filenames (including the paths) of the obtained tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

source("Module_A.R")   # Load Module A functions

# Acquire copy number data of six rectum adenocarcinoma (READ) patient samples.

filename_READ_CNA <- DownloadCNAData(cancerType = "READ", assayPlatform = "cna_cnv.hg19",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-
6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AF-2692", "TCGA-AG-4021"))

# Acquire copy number data of all bladder urothelial carcinoma (BLCA) patient samples.

filename_BLCA_CNA <- DownloadCNAData(cancerType = "BLCA", assayPlatform = NULL,
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire copy number data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_CNA <- DownloadCNAData(cancerType = "BRCA", assayPlatform = "cna_cnv.hg19",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-3C-
AAAU", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA-BH-A18Q", "TCGA-BH-A18R" ) )
```

DownloadMethylationData

Usage

```
DownloadMethylationData(cancerType, assayPlatform = NULL, tissueType = NULL, saveFolderName = ".",
    outputFileName = "", inputPatientIDs = NULL)
```

Description

This function downloads methylation data of samples belonging to the user-specified cancer type and tissue types, measured by the specified assay platforms, and combines them into tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can be one or a combination of methylation_27 and methylation_450. Its default value is NULL, which indicates both assay platforms (if available). Table 4 shows the description of assay platforms.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all above tissue types if available. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is the current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data from all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

Details

For each assayPlatform, this function generates a tab-delimited .txt data files. The filenames consist of five components: (1) outputFileName; (2) cancer type; (3) assay platforms used to generate the data; (4) specified tissue types connected by "_" or "tissueTypeAll" that indicates all available tissue types; (5) the date and time when the data were acquired. Double-underscore "__" is used to separate the five components in the filename. If outputFileName is an empty string, the filenames consist of only the other four components.

All output data files have the same format. The first row gives the TCGA barcodes of samples, while each other row corresponds to a CpG site. The 1st column is the index of CpG site. The 2nd column is gene symbol. The 3rd column is chromosome ID. The 4th column is the genomic coordinate of CpG site. Starting from the 5th column, each column is the "Beta_value" of a sample. For more details of the data format, data type, and data generation pipeline, please refer to https://wiki.nci.nih.gov/display/TCGA/DNA+methylation.

Table 4. Assay platforms for function DownloadMethylationData

assayPlatform	Assay	Filename Pattern
methylation_27	Infinium HumanMethylation27 BeadChip	jhu-usc.edu_.*.HumanMethylation27.
methylation_450	Infinium HumanMethylation450 BeadChip	jhu-usc.edu_.*.HumanMethylation450.

* indicates any character string

Value

A character vector of filenames (including paths) of the obtained tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

# Acquire humanmethylation450 data of six rectum adenocarcinoma (READ) patient samples.

filename_READ_Methylation450 <- DownloadMethylationData(cancerType = "READ", assayPlatform =
"methylation_450", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-
A01W", "TCGA-AG-3731"))

# Acquire humanmethylation27 data of all rectum adenocarcinoma (READ) patient samples.

filename_READ_Methylation27 <- DownloadMethylationData(cancerType = "READ", assayPlatform =
"methylation_27", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire humanmethylation450 data of three breast invasive carcinoma (BRCA) patient samples.
filename_BRCA_Methylation450 <- DownloadMethylationData(cancerType = "BRCA", assayPlatform =
```

```
"methylation_450", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-3C-AAAU", "TCGA-A7-A13F", "TCGA-BH-A0BZ", ) )

# Acquire humanmethylation27 data of four breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_Methylation27 <- DownloadMethylationData(cancerType = "BRCA", assayPlatform =
"methylation_27", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-A1-A0SD", "TCGA-BH-A18N", "TCGA-BH-A18Q", "TCGA-BH-A18R" ) )
```

DownloadmiRNASeqData

Usage

```
DownloadmiRNASeqData(cancerType, assayPlatform = NULL, tissueType = NULL, saveFolderName = ".",
    outputFileName = "", inputPatientIDs = NULL)
```

Description

This function downloads miRNASeq expression data of the user-specified cancer type and tissue types, measured by the specified assay platform, and then combines them into tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can be one or a combination of mir_GA.hg18, mir_GA.hg19, mir_GA.hg19.mirbase20, mir_HiSeq.hg18, mir_HiSeq.hg19, and mir_HiSeq.hg19.mirbase20. Its default value is NULL, which indicates all above assay platforms (if available). Table 5 shows the description of assay platforms.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all above tissue types if available. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is the current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data from all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

Details

For each assayPlatform, this function generates a tab-delimited .txt data files. The filenames consist of five components: (1) outputFileName; (2) cancer type; (3) assay platform used to generate the data; (4) specified tissue types connected by "_" or "tissueTypeAll" that indicates all available tissue types; (5) the date and time when the data were acquired. Double-underscore "__" is used to separate the five components in the filename. If outputFileName is an empty string, the filenames consist of only the other four components.

All data files have the same format. The 1st row includes the TCGA barcodes of samples, and the 2nd row shows whether a column of data is read_count or reads_per_million_miRNA_mapped, while all other rows are miRNA expression values. The 1st column is miRNA name. Starting from the 2nd column, every two columns, i.e. a read_count column and a reads_per_million_miRNA_mapped column, correspond to one sample. For more details of the data format, data type, and data generation pipeline, please refer to https://wiki.nci.nih.gov/display/TCGA/miRNASeq.

Table 5. Assay platforms for function DownloadmiRNASeqData

assayPlatform	Assay	Reference Genome	Additional Description	Filename Pattern
mir_GA.hg18	Illumina GA	Hg18		.mirna.quantification.txt
mir_GA.hg19	Illumina GA	Hg19		.hg19.mirna.quantification.txt
mir_GA.hg19.mirbase20	Illumina GA	Hg19	Use miRNA information from mirBase 20	.hg19.mirbase20.mirna.quantification.txt
mir_HiSeq.hg18	Illumina HiSeq	Hg18		.mirna.quantification.txt
mir_HiSeq.hg19	Illumina HiSeq	Hg19		.hg19.mirna.quantification.txt
mir_HiSeq.hg19.mirbase20	Illumina HiSeq	Hg19	Use miRNA information from mirBase 20	.hg19.mirbase20.mirna.quantification.txt

Value

A character vector of filenames (including path) of the downloaded tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

source("Module_A.R")   # Load Module A functions

# Acquire miRNASeq data of six rectum adenocarcinoma (READ) patient samples.
```

```
filename_READ_miRNASeq <- DownloadmiRNASeqData(cancerType = "READ", assayPlatform =
"mir_HiSeq.hg18", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AF-
2689", "TCGA-AF-2691"))

# Acquire miRNASeq data of all bladder urothelial carcinoma (BLCA) patient samples.

filename_BLCA_miRNASeq <- DownloadmiRNASeqData(cancerType = "BLCA", assayPlatform = NULL,
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire miRNASeq data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_miRNASeq <- DownloadmiRNASeqData(cancerType = "BRCA", assayPlatform =
"mir_HiSeq.hg19.mirbase20", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-3C-AAAU", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA-BH-
A18Q", "TCGA-BH-A18R" ) )
```

DownloadRNASeqData

Usage

```
DownloadRNASeqData(cancerType, assayPlatform = NULL, tissueType = NULL, saveFolderName = ".",
    outputFileName = "", inputPatientIDs = NULL)
```

Description

This function downloads RNA expression data of the user-specified cancer type and tissue types, measured by the specified assay platform, and then combines them into tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can be one or a combination of gene_Array, gene.normalized_RNAseq, gene_RNAseq, isoform.normalized_RNAseq, isoform_RNAseq, exon_RNAseq, and exonJunction_RNAseq. Its default value is NULL, which indicates all above assay platforms (if available). Table 6 shows the description of assay platforms.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all above tissue types if available. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data from all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

Details

For each assayPlatform, this function generates a tab-delimited .txt data files. The filenames consist of five components: (1) outputFileName; (2) cancer type; (3) assay platform used to generate the data; (4) specified tissue types separated by "_" or "tissueTypeAll" indicating all available tissue types; (5) the date and time when the data were acquired. Double-underscore "__" is used to separate the five components in the filename. If outputFileName is an empty string, the filenames consist of only the other four components.

When assayPlatform is gene_Array, the downloaded data are log2 lowess normalized (cy5/cy3) expression values collapsed by gene symbol. The 1st row is the TCGA barcodes of samples, while each of the other rows corresponds to a gene. The 1st column is the gene symbol of gene. And starting from the 2nd column, every column is the expression data of one sample.

When assayPlatform is gene.normalized_RNAseq, the downloaded data are normalized counts of genes. The 1st row is the TCGA barcode of sample, while each other row corresponds to a gene. The 1st column is the gene symbol (before "|") and Entrez ID (after "|") of each gene. And starting from the 2nd column, every column is the data of a sample.

When assayPlatform is gene_RNAseq, the data file includes raw_count and scaled_estimate data. The 1st row is the TCGA barcode of sample, and the 2nd row indicates whether a column is raw_count or scaled_estimate, while each of the other rows corresponds to a gene. The 1st column is gene symbol (before "|") and Entrez ID (after "|") of each gene. And starting from the 2nd column, every two columns correspond to one sample including a raw_count column and a scaled_estimate column.

When assayPlatform is isoform.normalized_RNAseq, the data file includes normalized counts of isoforms. The 1st row is the TCGA barcode of sample, while each other row corresponds to an isoform. The 1st column is isoform ID. And starting from the 2nd column, every column corresponds to a sample.

When assayPlatform is isoform_RNAseq, the data file includes raw_count and scaled_estimate data of isoforms. The 1st row is the TCGA barcode of sample, and the 2nd row indicates whether a column is raw_count or scaled_estimate, while each other row corresponds to a isoform. The 1st column is isoform ID. And starting from the 2nd column, every two columns correspond to one sample.

When assayPlatform is exon_RNAseq, the data file includes RPKM values of exon. The 1st row is the TCGA barcode of sample, while each of the other rows corresponds to an exon. The 1st column is the genomic coordinate of exon. Starting from the 2nd column, every column includes RPKM values of a sample.

When assayPlatform is exonJunction_RNAseq, the data file includes raw counts of exon junctions. The 1st row is the TCGA barcode of sample, while each of the other rows corresponds to an exon. The 1st column is the genomic coordinate of exon. The data of samples starting from the 2nd column.

For more details of the data format, data type, and data generation pipeline, please refer to https://wiki.nci.nih.gov/display/TCGA/RNASeq+Version+2.

Table 6. Assay platforms for function DownloadRNASeqData

assayPlatform	Assay	Additional Description	Filename Pattern
gene_Array	Agilent array	Gene expression quantification	.txt_lmean.out.logratio.gene.tcga_level3.data.txt

	G4502A		
gene.normalized_RNAseq	Illumina HiSeq	Normalized gene expression quantification	.rsem.genes.normalized_results
gene_RNAseq	Illumina HiSeq	Gene expression quantification	.rsem.genes.results
isoform.normalized_RNAseq	Illumina HiSeq	Normalized isoform expression quantification	.rsem.isoforms.normalized_results
isoform_RNAseq	Illumina HiSeq	Isoform expression quantification	.rsem.isoforms.results
exon_RNAseq	Illumina HiSeq	Exon quantification	.bt.exon_quantification.txt
exonJunction_RNAseq	Illumina HiSeq	Exon junction quantification	.junction_quantification.txt

Value

A character vector of filenames (including path) of the downloaded tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

# Acquire normalized gene expression data of six rectum adenocarcinoma (READ) patient samples.

filename_READ_RNASeq <- DownloadRNASeqData(cancerType = "READ", assayPlatform =
"gene.normalized_RNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-
3732", "TCGA-AG-3742"))

# Acquire exon junction expression data of all READ patient samples,

filename_READ_RNASeq <- DownloadRNASeqData(cancerType = "READ", assayPlatform =
"exonJunction_RNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire microarray gene expression data of all READ patient samples.

filename_READ_Microarray <- DownloadRNASeqData(cancerType = "READ", assayPlatform =
"gene_Array", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire normalized gene expression data of six breast invasive carcinoma (BRCA) patient samples
```

```
filename_BRCA_RNASeq_gene <- DownloadRNASeqData(cancerType = "BRCA", assayPlatform =
"gene.normalized_RNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-3C-AAAU", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA-BH-
A18Q", "TCGA-BH-A18R" ) )

# Acquire exon expression data of six breast invasive carcinoma (BRCA) patient samples

filename_BRCA_RNASeq_exon <- DownloadRNASeqData(cancerType = "BRCA", assayPlatform =
"exon_RNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs
= c("TCGA-3C-AAAU", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA-BH-A18Q", "TCGA-
BH-A18R" ) )
```

DownloadRPPAData

Usage

```
DownloadRPPAData(cancerType, assayPlatform = NULL, tissueType = NULL, saveFolderName = ".",
    outputFileName = "", inputPatientIDs = NULL)
```

Description

This function downloads Reverse Phase Protein Array (RPPA) protein expression data of samples belonging to the user-specified cancer type and tissue types, measured by the specified assay platform, and combines them into tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name. Please note, there is no RPPA data for LAML (Acute Myeloid Leukemia) patients.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can only be protein_RPPA. Its default value is NULL, which also indicates protein_RPPA. Table 7 shows the description of assay platform.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all above tissue types if available. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data from all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

Details

For each assayPlatform, this function generates a tab-delimited .txt data files. The filenames consist of five components: (1) outputFileName; (2) cancer type; (3) assay platform used to generate the data; (4) specified tissue types separated by "_" or "tissueTypeAll" indicating all available tissue types; (5) the date and time when the data were downloaded. Double-underscore "__" is used to separate the five components in the filename. If outputFileName is an empty string, the filenames consist of only the other four components.

In the data file, the 1st row is the TCGA barcode of sample, while each of the other row corresponds to a protein antibody. The 1st column shows the protein antibody name (after "|") and corresponding gene symbol (before "|") that encodes the protein. And starting from the 2nd column, each column corresponds to a sample. For details of the data format, data type, and data generation pipeline, please refer to https://wiki.nci.nih.gov/display/TCGA/Protein+Array+Data+Format+Specification.

Table 7. Assay platforms for function DownloadRPPAData

assayPlatform	Assay	Filename Pattern
protein_RPPA	Reserve Phase Protein Array (RPPA)	mdanderson.org_.*.MDA_RPPA_Core.protein_expression.Level_3.

* indicates any character string

Value

A character vector of filenames (including path) of the obtained tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

source("Module_A.R")   # Load Module A functions

# Acquire RPPA protein expression data of six rectum adenocarcinoma (READ) patient samples.

filename_READ_RPPA <- DownloadRPPAData(cancerType = "READ", assayPlatform = "protein_RPPA",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-
6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-3582", "TCGA-AG-4001"))

# Acquire RPPA protein expression data of all bladder urothelial carcinoma (BLCA) patient samples.

filename_BLCA_RPPA <- DownloadRPPAData(cancerType = "BLCA", assayPlatform = "protein_RPPA",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire RPPA protein expression data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_RPPA <- DownloadRPPAData(cancerType = "BRCA", assayPlatform = "protein_RPPA",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-3C-
AALI", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA-BH-A18Q", "TCGA-BH-A18R" ) )
```

DownloadSomaticMutationData

Usage

```
DownloadSomaticMutationData(cancerType, assayPlatform = NULL, tissueType = NULL,
    saveFolderName = ".", outputFileName = "", inputPatientIDs = NULL, supportFolderName =
    "./SupportingFiles")
```

Description

This function downloads somatic mutation data of samples belonging to the user-specified cancer type measured by the specified assay platform, and generates tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Its value can be any cancer type abbreviation, i.e. ACC, BLCA, BRCA, CESC, CHOL, COAD, DLBC, ESCA, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, MESO, OV, PAAD, PCPG, PRAD, READ, SARC, SKCM, STAD, TGCT, THCA, THYM, UCEC, UCS, UVM. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type. The cancer type abbreviation table (Table 1) shows the full cancer type name.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can only be somaticMutation_DNAseq. Its default value NULL is equivalent to somaticMutation_DNAseq. Table 8 shows the description of assay platform.

tissueType: a character vector indicating the specified tumor types, for which data should be downloaded. Its value can be one or multiple of the tumor type abbreviations, including TP, TR and TM. Its default value is NULL, which indicates all available tumor types. The tissue type abbreviation table (Table 2) shows the detail tumor type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data from all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

supportFolderName: a character string used as the path of the directory of Supporting files, whose default value is "./SupportingFiles".

Details

For each original somatic mutation file provided by GDC, this function generates a tab-delimited .txt data file. The filename consists of six components: (1) outputFileName; (2) cancer type; (3) assay platform used to generate the data; (4) specified tissue types separated by "_" or "tissueTypeAll" indicating all available tissue types; (5) the date and time when the data were downloaded; (6) original filename provided by GDC. Double-underscore "__" is used to separate the five components in the filename. If outputFileName is an empty string, the filename consists of only the other five components.

In the data file, the 1st row includes column names, while each of the other rows corresponds to a mutation. For details of the data format, data type, and data generation pipeline, please refer to https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format+(MAF)+Specification.

We found the Ctrl-Z character exists in the context of some original somatic mutation file from GDC. On Windows operating systems, the Ctrl-Z character will be recognized as a signal of EOF (End of File), which means those files can not be read completely, and the output file of this function will be truncated. So, we keep the original somatic mutation files in the sub directory "originalSomaticMutationFiles" for user's reference. Please note, the Ctrl-Z problem only effect the results on Windows operating system, but no effect on Mac OS or Linux.

Table 8. Assay platforms for function DownloadSomaticMutationData

assayPlatform	Assay	Filename Pattern
somaticMutation_DNAseq	DNA-Seq	.somatic.maf

Value

A character vector of filenames (including path) of the obtained tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

source("Module_A.R")   # Load Module A functions

# Acquire somatic mutation data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_somatic <- DownloadSomaticMutationData(cancerType = "BRCA", assayPlatform =
"somaticMutation_DNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-E2-A1IU", "TCGA-A7-A13F", "TCGA-BH- A0BZ", "TCGA-BH-A18N", "TCGA-BH-
A18Q", "TCGA-BH-A18R" ))

# Acquire somatic mutation data of glioblastoma multiforme (GBM) patient samples.

filename_GBM_somatic <- DownloadSomaticMutationData(cancerType = "GBM", assayPlatform =
"somaticMutation_DNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")
```

DownloadCPTACData

Usage

```
DownloadCPTACData(cancerType, assayPlatform = NULL, tissueType = NULL, saveFolderName = ".",
    outputFileName = "", inputPatientIDs = NULL)
```

Description

This function downloads Clinical Proteomic Tumor Analysis Consortium (CPTAC) proteomics data of TCGA patient samples belonging to the user-specified cancer type, measured by the specified assay platform, and generates tab-delimited .txt data files.

Arguments

cancerType: a character string indicating the specified cancer type for which data should be downloaded. Currently, CPTAC provides data of breast invasive carcinoma (BRCA), colorectal cancer including colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ), and ovarian serous cystadenocarcinoma (OV). So there are three possible values for cancerType, which are BRCA, COAD, READ, OV. Please refer to TCGA (https://tcga-data.nci.nih.gov/docs/publications/tcga/) for information about cancer type.

assayPlatform: a character vector indicating the assay platforms, for which data should be downloaded. Its value can be any or a combination of proteome_iTRAQ, phosphoproteome_iTRAQ, and glycoproteome_iTRAQ. The default value NULL indicates all available assay platforms. proteome_iTRAQ means total expressions of proteins; phosphoproteome_iTRAQ means expressions of phosphorylated proteins; glycoproteome_iTRAQ means expressions of proteins containing carbohydrates as posttranslational modifications. Not all cancer types have data of all three different assay platforms. Table 9 shows more information of assay platform.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all available tissue types. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

saveFolderName: a character string used as the path of the directory to save downloaded data files, whose default value is current working directory, i.e. ".".

outputFileName: a character string used to form the names of output data files. Its default value is an empty string.

inputPatientIDs: NULL or a character vector of TCGA barcode identify the patients whose data should be acquired. If it is NULL (by default), data of all patients in the specified cancer type and tissue type(s) will be acquired. Barcodes in inputPatientIDs must start with "TCGA-" and have 12 characters in length (eg. "TCGA-XX-XXXX"), but do not need to be full-length and complete, because the leading 12 characters of barcodes provide enough information for identifying patients. For details

of TCGA barcodes, please refer to TCGA Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

Details

For each assayPlatform, this function generates a tab-delimited .txt data files. The filenames consist of five components: (1) outputFileName; (2) cancer type; (3) assay platform used to generate the data; (4) specified tissue types separated by "_" or "tissueTypeAll" indicating all available tissue types; (5) the name abbreviation of the center that generated the data, which can be one of "BI" (Broad Institute), "JHU" (Johns Hopkins University), "PNNL" (Pacific Northwest National Laboratory) and "VU" (Vanderbilt University); (6) the date and time when the data were downloaded. Double-underscore "__" is used to separate the five or six components in the filename.

In the proteome data files of BRCA and OV, the 1st row is the column names and TCGA barcodes of samples, while each of the other rows corresponds to a protein. The 1st column shows the gene symbol that encodes the protein. The 2nd column is the gene description. The 3rd column is the organism. The 4th column is the chromosome ID. The 5th column is the genomic location of the gene. Starting from the 6th column, each two columns correspond to a sample, in which the first column is Log Ratio and the second column is Unshared Log Ratio. Log Ratio is the log of the ratio between the spectral count of a protein in a sample verses the spectral count of this protein in the reference sample, while all peptides mapping to this protein are counted. Unshared Log Ratio is the log of the ratio between the spectral count of a protein in a sample verses the spectral count of this protein in the reference sample without counting the peptides that can map to more than one protein.

In the proteome data file of colorectal cancer (COAD, READ), the 1st row is the column names and TCGA barcodes of samples, while each of the other rows corresponds to a protein. The 1st column shows the gene symbol of the protein. The 2nd column is gene description. The 3rd column is the organism. The 4th column is the chromosome ID. The 5th column is the genomic location of the gene. Starting from the 6th column, each two columns correspond to a sample, in which the first column is Spectral Counts and the second column is Unshared Spectral Counts. Spectral Counts refer to the count of all peptides mapping to a protein, while Unshared Spectral Counts refer to the count of peptides uniquely mapping to this protein.

In the phosphoproteome data files of BRCA and OV, only Log Ratio data are included. The 1st row is the column names and TCGA sample barcodes, while each of the other rows corresponds to a phosphosite. The 1st column shows the position of the phosphosite. The 2nd column is the sequence of the peptide. The 3rd column is the gene symbol. The 4th column is the organism. Data start from the 5th column.

In the glycoproteome data file of OV, only Log Ratio data are included. The 1st row is the column names and TCGA sample barcodes, while each of the other rows corresponds to a glycosite. The 1st column shows the position of the glycosite. The 2nd column is the sequence of the peptide. The 3rd column is the gene symbol. The 4th column is the organism. Data start from the 5th column.

Table 9. Assay platforms for function DownloadCPTACData

cancerType	Introduction	assayPlatform
BRCA	https://cptac-data-portal.georgetown.edu/cptac/s/S015	proteome_iTRAQ phosphoproteome_iTRAQ
COAD, READ	https://cptac-data-	proteome_iTRAQ

	portal.georgetown.edu/cptac/s/S016	
OV	https://cptac-data-portal.georgetown.edu/cptac/s/S020	proteome_iTRAQ phosphoproteome_iTRAQ glycoproteome_iTRAQ

Value

A character vector of filenames (including path) of the obtained tab-delimited .txt data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

source("Module_A.R")   # Load Module A functions

# Acquire CPTAC protein expression data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_iTRAQ <- DownloadCPTACData(cancerType = c("BRCA"), assayPlatform =
"proteome_iTRAQ", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-A2-A0CM", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA- BH-
A18Q", "TCGA-BH-A18R" ))

# Acquire all proteomics data of ovarian serous cystadenocarcinoma (OV) patient samples.

filename_OV_iTRAQ <- DownloadCPTACData(cancerType = c("OV"), saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler")

# Acquire CPTAC phosphoproteome data of colorectal cancer (COAD) patient samples.

filename_COAD_iTRAQ <- DownloadCPTACData(cancerType = c("COAD"), assayPlatform =
"proteome_iTRAQ", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Acquire CPTAC phosphoproteome data of colorectal cancer (READ) patient samples.

filename_READ_iTRAQ <- DownloadCPTACData(cancerType = c("READ"), assayPlatform =
"proteome_iTRAQ", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")
```

CalculateSingleValueMethylationData

Usage

```
CalculateSingleValueMethylationData(input, regionOption, DHSOption, outputFileName,
    outputFileFolder, chipAnnotationFile = "./SupportingFiles/MethylationChipAnnotation.rda")
```

Description

This function does the following

(1) Calculate an average methylation value for each gene based on certain CpG sites according to the specified option (see details of regionOption and DHSOption).

(2) Draw and save a box plot of the obtained single-value methylation data. The filename of the box plot picture is composed of outputFileName, regionOption, DHSOption, and "boxplot.png", with double-underscore "__" separating them.

(3) Save the single-value methylation data as a tab-delimited .txt file. The first two columns are gene symbol and the single-value type that shows the regionOption and DHSOption used to calculate the data with "|" in between to separate them. The other columns are data of individual samples. The first row gives TCGA barcodes of samples (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode). The name of the file is composed of outputFileName, regionOption, and DHSOption with double-underscore "__" separating them.

(4) Save the single-value methylation data as an R data file (.rda) that includes two variables. The first variable is Des, which is a two-column character matrix including gene symbol and the single-value type that shows the regionOption and DHSOption used to calculate the data, with "|" in between to separate them. Des serves as the description of the second variable, Data, which is a numeric matrix. Each column in Data is a sample and each row corresponds to a gene. The column names of Data are TCGA barcodes of samples. The name of the file is composed of outputFileName, regionOption, and DHSOption with double-underscore "__" separating them.

The output .txt file and .rda file contain the same data. Having the same data in two different file formats is for the convenience of using data under different software environments.

Arguments

input: a list object containing the methylation data based on which a single methylation value needs to be calculated for each gene. This list object can be generated by the ProcessMethylation27Data, ProcessMethylation450Data, and MergeMethylationData functions. It is a list object formed by two variables Des and Data. Des is a four-column character matrix including Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. Data is a numeric matrix containing methylation data of samples, where each column corresponds to a sample and the column names are the TCGA barcodes of samples. Des serves as the description of Data. Another way to generate this list object is to load the .rda file produced by the ProcessMethylation27Data, ProcessMethylation450Data, and MergeMethylationData functions, and form a list object using the loaded Des and Data variables.

regionOption: a character string indicating for which genomic region of a gene the average methylation value should be calculated, based on HumanMethylation450 BeadChip annotations provided by Illumina (http://support.illumina.com/downloads/humanmethylation450_15017482_v1-2_product_files.ilmn). Available options are TSS1500, TSS200, 5'UTR, 1stExon, Body, 3'UTR, and All. TSS1500: within 1500 bps ahead of a Transcription Start Site (TSS), but not including the 200 bps ahead of the TSS. TSS200: within 200 bps ahead of a TSS. 5'UTR: 5' untranslated region. 1stExon: first exon. Body: gene body. 3'UTR: 3' untranslated region. All: all CpG sites associated with a gene no matter which genomic region the CpG sites are in. For details about the definitions of the region options, please refer to the Illumina annotations. Including more than one options to calculate an average methylation value over more than one regions.

DHSOption: a character string that can be DHS, notDHS, or Both, indicating whether only CpG sites that are DNAse hypersensitive will be included in the calculation. DHS selects CpG sites that are DNAse hypersensitive. notDHS selects CpG sites that are not labeled as DNAse hypersensitive. Both selects all CpG sites no matter whether they are DNAse hypersensitive or not. The DNAse hypersensitivity of CpG sites are experimentally determined and provided by Illumina chip annotations.

outputFileName: a character string used to form the names of output data files and box plot picture file.

outputFileFolder: a character string indicating the directory in which the output files will be saved.

chipAnnotationFile: a character string indicating the path of the chip annotation file to be used by the function, which is the MethylationChipAnnotation.rda file in the SupportingFiles folder in the package.

Value

A list object of two variables. The first variable is Des, a two-column character matrix including gene symbol and the single-value type that shows the regionOption and DHSOption used to calculate the data with "|" in between to separate them. Des serves as the description of the second variable, Data, which is a numeric data matrix. Each column in the data matrix is a sample and each row corresponds to a gene. The column names of Data are the TCGA barcodes of samples.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download humanmethylation450 data of six rectum adenocarcinoma (READ) patient samples

filename_READ_Methylation450 <- DownloadMethylationData(cancerType = "READ", assayPlatform =
"methylation_450", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-
A01W", "TCGA-AG-3731"))

# Process the downloaded humanmethylation450 data and import the data into R.
```

```
Methylation450Data <- ProcessMethylation450Data(inputFilePath = filename_READ_Methylation450[1],
outputFileName = "READ__humanmethylation450", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Calculate average methylation values of all CpG sites associated with the genes.

Methylation450_All_Both <- CalculateSingleValueMethylationData(input = Methylation450Data,
regionOption = "All", DHSOption = "Both", outputFileName =
"READ__humanmethylation450__SingleValue", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler", chipAnnotationFile =
"./SupportingFiles/MethylationChipAnnotation.rda")

# Calculate average methylation values of CpG sites within TSS200 region of the genes.

Methylation450_TSS200_Both <- CalculateSingleValueMethylationData(input = Methylation450Data,
regionOption = "TSS200", DHSOption = "Both", outputFileName =
"READ__humanmethylation450__SingleValue", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Calculate average methylation values of CpG sites that are within TSS1500 region of the genes
# and are DNAse hypersensitive.

Methylation450_TSS1500_DHS <- CalculateSingleValueMethylationData(input = Methylation450Data,
regionOption = "TSS1500", DHSOption = "DHS", outputFileName =
"READ__humanmethylation450__SingleValue", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler", chipAnnotationFile =
"./SupportingFiles/MethylationChipAnnotation.rda")
```

CheckGeneSymbol

Usage

```
CheckGeneSymbol(Des)
```

Description

This function identifies obsolete gene symbols and gene symbol errors caused by auto-conversion in spreadsheet programs, like Excel, and corrects them to official HGNC gene symbols. Symbols that are not known errors, obsolete gene symbols, or official gene symbols will be left unchanged.

Input argument

Des: a character matrix containing the description of genomic features. One column of the matrix MUST be gene symbol, and its column name MUST be "GeneSymbol".

Value

A character matrix with the same structure of Des, but with the gene symbols checked and corrected.

Details

If a symbol can not be uniquely mapped to one official gene symbol, for example a symbol is the alias for more than one genes, all official gene symbols that are associated with this alias will be included in the corrected gene symbol, with triple-underscore "___" separating each involved official gene symbol.

Example

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download humanmethylation27 data of six rectum adenocarcinoma (READ) patient samples

filename_READ_Methylation27 <- DownloadMethylationData(cancerType = "READ", assayPlatform =
"methylation_27", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-AG-3583", "TCGA-AG-A032", "TCGA-AF-2692", "TCGA-AG-4001", "TCGA-AG-
3608", "TCGA-AG-3574"))

# Process the downloaded humanmethylation27 data and import the data into R
```

```
Methylation27Data <- ProcessMethylation27Data(inputFilePath = filename_READ_Methylation27[1],
outputFileName = "READ__humanmethylation27", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

Des <- Methylation27Data$Des

# We purposely make a symbol error to simulate mistakes that Excel may
# make by transferring gene symbol to date format.

ID <- which(Des[, "GeneSymbol"] == "SEPT7")  # SEPT7 is an official gene symbol

# Change SEPT7 to a wrong gene symbol in date format

Des[ID, "GeneSymbol"] <- "7-Sep"

# Check and Correct gene symbol using the CheckGeneSymbol function

Des <- CheckGeneSymbol(Des)

print(Des[ID, "GeneSymbol"])
```

CombineMultiPlatformData

Usage

```
CombineMultiPlatformData(inputDataList, combineStyle = "Intersect")
```

Description

This function combines multi-platform datasets into a single mega data table, through matching of patient samples and genomic features. There are two possible ways to match patient samples across platforms. One is to identify samples measured by all assay platforms and merge the multi-platform data of these common samples, which is called Intersect and also the default setting of this function. The other is to include a sample as long as it is measured by at least one assay platform, which is called Union. Matching genomic features refers to putting together the multi-platform data of the same gene (i.e. making them adjacent rows in the data table). Currently, this function works on combining seven different types of data, including gene expression, protein expression produced by reverse phase protein array (RPPA) used in TCGA, DNA Methylation, DNA copy number, miRNA expression, gene-level somatic mutation data generated by the ProcessSomaticMutationData function, and protein expression produced by mass spectrometry used in CPTAC. To match genomic features, the data must contain the information of genes associated with the genomic features (such as through the gene symbols in Des variable produced by many data processing functions in this package). DNA copy number data must be preprocessed by the ProcessCNAData function to get gene-level copy number values for combining with data from other platforms.

Input argument

inputDataList: a vector of list objects. Each element in the vector is a list object of three variables that represent one dataset to be combined. The names of the three variables are Des, Data, and dataType. Des is a character matrix including descriptions of genomic features. One column in Des must be gene symbols and with a column name of "GeneSymbol", because matching of genomic features is based on this column. Data is a numeric matrix containing the data. Each row is a genomic feature and each column is a sample. The column names must be the TCGA sample barcodes (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode). dataType is a character string indicating the type of data. Options of dataType include methylation, copyNumber, somaticMutation, geneExp, miRNAExp, protein_RPPA, and protein_iTRAQ, standing for DNA methylation, DNA copy number, somatic mutation, gene expression, miRNA expression, protein expression produced by RPPA (from GDC) and protein expression generated by iTRAQ (from CPTAC), respectively.

combineStyle: a character string indicating how the samples should be combined. Intersect, which is the default setting, selects samples with all data types. Union includes a sample as long as it has at least one type of data.

Value

A list object of two variables Des and Data.

Des is a character matrix including descriptions of genomic features. Des has three columns. The first column is gene symbol. The second column is platform, which can be methylation, copyNumber, somaticMatation, geneExp, miRNAExp, protein_RPPA and protein_iTRAQ representing DNA methylation, DNA copy number, somatic mutation, gene expression, miRNA expression, protein expression produced by RPPA (from GDC) and protein expression generated by iTRAQ (from CPTAC), respectively. The third column is a description of genomic features. If the platform is geneExp, the description is Entrez ID of the gene. If the platform is protein_RPPA, the description is the name of the protein antibody used in the RPPA assay. For copyNumber platform, the description shows the chromosome ID and strand of gene. For miRNAExp platform, the description column is empty. For methylation platform, if the data are single-value methylation data calculated by the CalculateSingleValueMethylationData function, the description column gives the single-value type indicating how the data were calculated (refer to CalculateSingleValueMethylationData function for the definition of single-value type); if the data are methylation data of CpG sites, the description column gives the Illumina ID, chromosome ID, and genomic coordinate of the CpG sites with "|" separating them. For somaticMutation platform, the description is Entrez ID of the gene. For protein_iTRAQ platform, the description is the full protein name.

Data is a numeric matrix including the merged data. Each row is a genomic feature and each column is a sample. The column names of Data are the TCGA barcodes of samples with their first 15 characters, which indicate the tissue source site, patient index, and tissue type. Rows of Data are ordered so that genomic features of the same gene are adjacent in the matrix. Des serves as the description of Data.

Details

If there are more than one sample of the same patient and the same tissue type measured by an assay platform (which is actually a rare case), only one of the samples will be kept for data combining.

Example

```
# In this example, we will first download and process various kinds of data of several rectum
# adenocarcinoma (READ) patient samples, and then merge them into single mega table using
# both the Intersect approach and Union approach. The present working directory of R must
# be TCGA-Assembler, i.e. the package folder, to run the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download and process copy number data of six READ patient samples.

filename_READ_CNA <- DownloadCNAData(cancerType = "READ", assayPlatform = "cna_cnv.hg19",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-
6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AF-2692", "TCGA-AG-4021"))

GeneLevelCNA <- ProcessCNAData(inputFilePath = filename_READ_CNA[1], outputFileName =
"READ__genome_wide_snp_6__GeneLevelCNA", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler", refGenomeFile =
"./SupportingFiles/Hg19GenePosition.txt")

# Download and process humanmethylation450 data of six READ patient samples
```

```
filename_READ_Methylation450 <- DownloadMethylationData(cancerType = "READ", assayPlatform = "methylation_450", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-A01W", "TCGA-AG-3731"))

# Acquire humanmethylation27 data of all rectum adenocarcinoma (READ) patient samples.

Methylation450Data <- ProcessMethylation450Data(inputFilePath = filename_READ_Methylation450[1], outputFileName = "READ__humanmethylation450", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler")

# Calculate single-value methylation data.

Methylation450_TSS1500_DHS <- CalculateSingleValueMethylationData(input = Methylation450Data, regionOption = "TSS1500", DHSOption = "DHS", outputFileName = "READ__humanmethylation450__SingleValue", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler", chipAnnotationFile = "./SupportingFiles/MethylationChipAnnotation.rda")

# Download and process miRNA-seq data of six READ patient samples

filename_READ_miRNASeq <- DownloadmiRNASeqData(cancerType = "READ", assayPlatform = "mir_HiSeq.hg19", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AF-2689", "TCGA-AF-2691"))

miRNASeqData <- ProcessmiRNASeqData(inputFilePath = filename_READ_miRNASeq[1], outputFileName = "READ__illuminahiseq_mirnaseq", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler")

# Download and process normalized gene expression data of six READ patient samples

filename_READ_RNASeq <- DownloadRNASeqData(cancerType = "READ", assayPlatform = "gene.normalized_RNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-3732", "TCGA-AG-3742"))

GeneExpData <- ProcessRNASeqData(inputFilePath = filename_READ_RNASeq[1], outputFileName = "READ__illuminahiseq_rnaseqv2__GeneExp", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler", dataType = "geneExp", verType = "RNASeqV2")

# Download and process RPPA protein expression data of six READ patient samples

filename_READ_RPPA <- DownloadRPPAData(cancerType = "READ", assayPlatform = "protein_RPPA", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-3582", "TCGA-AG-4001"))

RPPAData <- ProcessRPPADataWithGeneAnnotation(inputFilePath = filename_READ_RPPA[1], outputFileName = "READ__mda_rppa_core", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler")

# Put multi-modal data in a vector of list objects to be inputted into CombineMultiPlatformData function.

inputDataList <- vector("list", 5)

inputDataList[[1]] <- list(Des = GeneExpData$Des, Data = GeneExpData$Data, dataType = "geneExp")
```

```
inputDataList[[2]] <- list(Des = Methylation450_TSS1500_DHS$Des, Data =
Methylation450_TSS1500_DHS$Data, dataType = "methylation")

inputDataList[[3]] <- list(Des = GeneLevelCNA$Des, Data = GeneLevelCNA$Data, dataType =
"copyNumber")

inputDataList[[4]] <- list(Des = RPPAData$Des, Data = RPPAData$Data, dataType = "protein_RPPA")

inputDataList[[5]] <- list(Des = miRNASeqData$Des, Data = miRNASeqData$Data, dataType =
"miRNAExp")

# Merge multi-platform data using Intersect approach.

MergedData <- CombineMultiPlatformData(inputDataList = inputDataList)

# Merge multi-platform data using Union approach.

MergedData <- CombineMultiPlatformData(inputDataList = inputDataList, combineStyle = "Union")
```

ExtractTissueSpecificSamples

Usage

```
ExtractTissueSpecificSamples(inputData, tissueType, singleSampleFlag, sampleTypeFile =
    "./SupportingFiles/TCGASampleType.txt")
```

Description

This function extracts from the input data matrix a subset of the data belonging to the specified tissue types.

Arguments

inputData: a numeric matrix containing data. Each row is a genomic feature and each column is a sample. The column names of the data matrix must be TCGA sample barcodes, which must start with TCGA- and have a length no shorter than 15 characters. Refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode for details of TCGA barcodes.

tissueType: a character vector indicating the specified tissue types, for which data should be downloaded. Its value can be one or multiple of the tissue type abbreviations, including TP, TR, TB, TRBM, TAP, TM, TAM, THOC, TBM, NB, NT, NBC, NEBV, and NBM. Its default value is NULL, which indicates all above tissue types if available. The tissue type abbreviation table (Table 2) shows the detail tissue type names of all abbreviations.

singleSampleFlag: a logical variable. If it is TRUE, when there are multiple samples of the same patient and the same tissue type, only one sample is kept; if it is FALSE, all samples are kept.

sampleTypeFile: a character string indicating the path of the TCGA sample type file to be used by the function, which is TCGASampleType.txt in the SupportingFiles folder in the package.

Value

A numeric matrix containing the extracted data. Each column in the matrix is a sample and each row is a genomic feature. The column names of the matrix are the TCGA barcodes of samples.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download humanmethylation27 data of six rectum adenocarcinoma (READ) patient samples
```

```
Methylation27RawData <- DownloadMethylationData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"methylation_27", inputPatientIDs = c("TCGA-AG-3583", "TCGA-AG-A032", "TCGA-AF-2692", "TCGA-AG-
4001", "TCGA-AG-3608", "TCGA-AG-3574"))

# Process the downloaded humanmethylation27 data and import the data into R

Methylation27Data <- ProcessMethylation27Data(inputFilePath = Methylation27RawData[1],
outputFileName = "READ__humanmethylation27", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Extract methylation data of primary solid tumor samples.

ExtractedData__TP <- ExtractTissueSpecificSamples(inputData = Methylation27Data$Data, tissueType =
"TP", singleSampleFlag = FALSE, sampleTypeFile = "./SupportingFiles/TCGASampleType.txt")

# Extract methylation data of primary solid tumor samples and solid tissue normal samples.

ExtractedData__TP_NT <- ExtractTissueSpecificSamples(inputData = Methylation27Data$Data,
tissueType = c("TP", "NT"), singleSampleFlag = TRUE)
```

MergeMethylationData

Usage

```
MergeMethylationData(input1, input2, outputFileName, outputFileFolder)
```

Description

This function combines two methylation datasets together by doing the following:

(1) Identify the CpG sites that appear in both datasets, and combine the data of these CpG sites.

(2) Perform quantile normalization on the combined data.

(3) Draw and save a box plot of the combined data before and after normalization for quality control purpose. The picture filenames are composed of outputFileName and "__BeforeNormalizationBoxplot.png" or "__AfterNormalizationBoxplot.png".

(4) Save the normalized combined data as a tab-delimited .txt file. The first four columns are Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. The other columns are data of individual samples. The first row gives TCGA barcodes of samples (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

(5) Save the normalized combined data as an R data file (.rda) that includes two variables. The first variable is Des, which is a character matrix including the four-column description of CpG sites. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in the matrix corresponds to a sample and the column names are TCGA barcodes of samples.

The output .txt file and .rda file contain the same data. Having the same data in two different file formats is for the convenience of using data under different software environments.

Arguments

input1: a list object containing a methylation dataset to be merged. It is a list of two variables. One variable is Des, which is a four-column character matrix including Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. The other variable is Data, a numeric matrix containing methylation data. Each column in Data corresponds to a sample with column names showing the TCGA barcodes of samples. Des serves as the description of Data. This list object of input data can be generated by the ProcessMethylation27Data, ProcessMethylation450Data, or MergeMethylationData function. Another way to generate this list object is to load the .rda data file produced by the ProcessMethylation27Data or ProcessMethylation450Data function, and form a list using the loaded Des and Data.

input2: a list object containing the other methylation dataset to be merged, which should be generated in the same way as input1.

outputFileName: a character string to form the names of output files.

outputFileFolder: a character string indicating the path of the directory in which the output files will be saved.

Value

A list object formed by two variables. The first variable is Des, which is a character matrix including a four-column description of CpG sites, i.e. Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in Data corresponds to a sample and the column names of Data are TCGA barcodes of samples.

Details

MergeMethylationData function can be used to merge only methylation datasets that include Illumina IDs of CpG sites as description. It can NOT be used to merge methylation dataset without Illumina IDs of CpG sites.

Example

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,

# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download humanmethylation27 data of six rectum adenocarcinoma (READ) patient samples

Methylation27RawData <- DownloadMethylationData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"methylation_27", inputPatientIDs = c("TCGA-AG-3583", "TCGA-AG-A032", "TCGA-AF-2692", "TCGA-AG-
4001", "TCGA-AG-3608", "TCGA-AG-3574"))

# Process the downloaded humanmethylation27 data and import the data into R

Methylation27Data <- ProcessMethylation27Data(inputFilePath = Methylation27RawData[1],
outputFileName = "READ__humanmethylation27", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Download humanmethylation450 data of six READ patient samples

Methylation450RawData <- DownloadMethylationData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"methylation_450", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-
6812", "TCGA-AG-A01W", "TCGA-AG-3731"))

# Process the downloaded humanmethylation450 data and import the data into R

Methylation450Data <- ProcessMethylation450Data(inputFilePath = Methylation450RawData[1],
outputFileName = "READ__humanmethylation450", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Merge the humanmethylation27 data and humanmethylation450 data
```

```
Methylation27_450_Merged <- MergeMethylationData(input1 = Methylation27Data, input2 =
Methylation450Data, outputFileName = "READ__humanmethylation27_450_merged", outputFileFolder
= "./ManualExampleData/ProcessedData.TCGA-Assembler")
```

ProcessCNAData

Usage

```
ProcessCNAData(inputFilePath, outputFileName, outputFileFolder, refGenomeFile)
```

Description

This function processes DNA copy number data files acquired by TCGA-Assembler Module A or downloaded from Firehose website, and imports the data into R. It does the following:

(1) Calculate gene-level copy number value, which is the average copy number of the genomic region of a gene.

(2) For gene-level copy number data, check and correct the gene identifiers to official gene symbols.

(3) Draw and save a box plot of the gene-level copy number data for quality control purpose. The picture filename is composed of outputFileName and "__boxplot.png".

(4) Save the gene-level copy number data as a tab-delimited .txt file. The first three columns are descriptions of genomic features including gene symbol, chromosome ID, and strand. The other columns are data of individual samples. The first row gives TCGA sample barcode (see reference at https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode). The filename is composed of outputFileName and ".txt".

(5) Save the gene-level copy number data as an R data file (.rda), which includes two variables. The first variable is Des, a character matrix including gene symbol, chromosome ID, and strand in the first, second, and third column, respectively. It describes the second variable, Data, which is a numeric matrix. Each column in Data is a sample and each row corresponds to a gene. The column names of Data are sample barcodes. The filename is composed of outputFileName and ".rda".

The output .txt file and .rda file contain the same data. Having the same data in two different file formats is for the convenience of using data under different software environments.

Arguments

inputFilePath: a character string indicating the path of the input DNA copy number data file acquired by TCGA-Assembler Module A or downloaded from Firehose.

outputFileName: a character string to form the names of output files.

outputFileFolder: a character string indicating the path of the directory to which the output files will be saved.

refGenomeFile: a character string indicating the path of gene genomic position file that will be used. There are two gene position files in the SupportingFiles folder of the package. Use Hg18GenePosition.txt or Hg19GenePosition.txt, if the input copy number data was generated based on reference genome Hg18 or Hg19, respectively.

Value

A list object formed by two variables Des and Data. Des is a character matrix of three columns including gene symbol, chromosome ID, and strand in the first, second, and third column, respectively. It describes the second variable, Data, which is a numeric matrix of gene-level copy number. Each column in Data is a sample and each row corresponds to a gene. The column names of Data are sample barcodes.

Examples:

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download DNA copy number data of six rectum adenocarcinoma (READ) patient samples.

CNARawData <- DownloadCNAData(saveFolderName = "./ManualExampleData/RawData.TCGA-
Assembler", cancerType = "READ", assayPlatform = "cna_cnv.hg19", inputPatientIDs = c("TCGA-EI-6884",
"TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AF-2692", "TCGA-AG-4021"))

# Process the downloaded copy number data and calculate an average copy number for each gene.
# Save results to subfolder ProcessedData.TCGA-Assembler in the ManualExampleData folder.

READ.GeneLevel.CNA <- ProcessCNAData(inputFilePath = CNARawData[1], outputFileName =
"READ__genome_wide_snp_6__GeneLevelCNA", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler", refGenomeFile =
"./SupportingFiles/Hg19GenePosition.txt")

# Process READ copy number data downloaded from Firehose website and calculate average copy
# number value. Save results to ./ManualExampleData/ProcessedData.Firehose/.

READ.GeneLevel.CNA <- ProcessCNAData(inputFilePath =
"./ManualExampleData/RawData.Firehose/READ.snp__genome_wide_snp_6__broad_mit_edu__hg18_
_Level_3__segmented_scna_hg18__seg.seg.txt", outputFileName =
"READ__genome_wide_snp_6__GeneLevelCNA", outputFileFolder =
"./ManualExampleData/ProcessedData.Firehose", refGenomeFile =
"./SupportingFiles/Hg18GenePosition.txt")
```

ProcessMethylation27Data

Usage

```
ProcessMethylation27Data(inputFilePath, outputFileName, outputFileFolder, fileSource = "TCGA-Assembler")
```

Description

This function processes HumanMethylation27 BeadChip data file either acquired by TCGA-Assembler Module A or downloaded from Firehose, and imports the data into R. It does the following.

(1) For data files downloaded from Firehose, remove redundant columns in the data. Firehose HumanMethylation27 data file includes replicated columns of probe descriptions, i.e. gene symbol, chromosome ID, genomic coordinate, which are identical for each sample.

(2) Check whether the gene symbols are official HGNC gene symbols. If not, correct them.

(3) When a CpG site corresponds to more than one gene symbol, duplicate the measurements of the CpG site (a row in the data matrix) for each gene symbol.

(4) Draw and save a box plot of the data for quality control purpose. The picture filename is composed of outputFileName and "__boxplot.png".

(5) Save the processed data as a tab-delimited .txt file. The first four columns are Illumina ID of CpG site, gene symbol, chromosome ID, and genome coordinate. The other columns are data of individual samples. The top row gives the TCGA sample barcodes (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

(6) Save the processed data as an R data file (.rda), which includes two variables Des and Data. Des is a character matrix including the four-column description of CpG sites. It serves as the description of Data, which is a numeric matrix. Each column in the data matrix is a sample and the column names are sample barcodes.

The output .txt file and .rda file contain the same data. Having the same data in two different file formats is for the convenience of using data under different software environments.

Arguments

inputFilePath: a character string indicating the path of input data file acquired by TCGA-Assembler Module A or downloaded from Firehose website. It should be a tab-delimited .txt (or .txt) file.

outputFileName: a character string to form the names of output files.

outputFileFolder: a character string indicating the path of the directory to which the output files will be saved.

fileSource: a character string, either TCGA-Assembler or Firehose. TCGA-Assembler, which is the default setting, indicates that the input data file was acquired by TCGA-Assembler Module A and Firehose indicates that the input data file was downloaded from Firehose website.

Value

A list object of two variables, which are Des and Data. Des is a character matrix of the four-column description of CpG sites including Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in the matrix is a sample and the column names are the TCGA sample barcodes.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,

# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download humanmethylation27 data of six rectum adenocarcinoma (READ) patient samples

Methylation27RawData <- DownloadMethylationData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"methylation_27", inputPatientIDs = c("TCGA-AG-3583", "TCGA-AG-A032", "TCGA-AF-2692", "TCGA-AG-
4001", "TCGA-AG-3608", "TCGA-AG-3574"))

# Process the downloaded humanmethylation27 data and save the results to

# ./ManualExampleData/ProcessedData.TCGA-Assembler

Methylation27Data <- ProcessMethylation27Data(inputFilePath = Methylation27RawData[1],
outputFileName = "READ__humanmethylation27", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Process READ HumanMethylation27 data downloaded from Firehose website and

# save the results to ./ManualExampleData/ProcessedData.Firehose

Methylation27Data <- ProcessMethylation27Data(inputFilePath =
"./ManualExampleData/RawData.Firehose/READ.methylation__humanmethylation27__jhu_usc_edu__L
evel_3__within_bioassay_data_set_function__data.data.txt", outputFileName =
"READ__humanmethylation27", outputFileFolder = "./ManualExampleData/ProcessedData.Firehose",
fileSource = "Firehose")
```

ProcessMethylation450Data

Usage

```
ProcessMethylation450Data(inputFilePath, outputFileName, outputFileFolder, fileSource = "TCGA-Assembler")
```

Description

This function processes HumanMethylation450 BeadChip data file either acquired by TCGA-Assembler Module A or downloaded from Firehose website, and imports the data into R. It does the following.

(1) For HumanMethylation450 data file downloaded from Firehose website, remove redundant columns in the data. Firehose HumanMethylation450 data file includes replicated columns of probe descriptions, i.e. gene symbol, chromosome ID, genomic coordinate, which are identical for each sample. This redundant information makes the file size very large, more than 10GB for some cancer types, which produces difficulties when loading the data into software environment for data analysis, such as R.

(2) Check whether the gene symbols are official HGNC gene symbols. If not, correct them.

(3) When a CpG site corresponds to more than one gene symbol, duplicate the measurements of the CpG site (a row in the data matrix) for each gene symbol that it stands for.

(4) Draw and save a box plot of the data for quality control purpose. The picture filename is composed of `outputFileName` and "__boxplot.png".

(5) Save the processed data as a tab-delimited .txt file. The first four columns are Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. The other columns are data of individual samples. And the top row shows the TCGA barcodes of samples (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

(6) Save the processed data as an R data file (.rda), which contains two variables `Des` and Data. `Des` is a character matrix including the four-column description of CpG sites. It serves as the description of the second variable, `Data`, which is a numeric matrix. Each column in `Data` is a sample and the column names of `Data` are TCGA sample barcodes.

The output .txt file and .rda file contain the same data. Having the same data in two different file formats is for the convenience of using data under different software environments.

Arguments

`inputFilePath`: a character string indicating the path of the input data file, either acquired by TCGA-Assembler Module A or downloaded from Firehose website. It should be a tab-delimited .txt (or .txt) file.

`outputFileName`: a character string to form the names of output files.

`outputFileFolder`: a character string indicating the directory in which the output files will be saved.

fileSource: a character string, either TCGA-Assembler or Firehose. TCGA-Assembler, which is the default setting, indicates that the input data file was acquired by TCGA-Assembler Module A, and Firehose indicates that the input data file was downloaded from Firehose website.

Value

A list object formed by Des and Data. Des is a character matrix of four-column descriptions of CpG sites including Illumina ID of CpG site, gene symbol, chromosome ID, and genomic coordinate. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in Data is a sample and the column names of Data are TCGA sample barcodes.

Details

For HumanMethylation450 data file downloaded from Firehose, this function calls GetMethylation450Data function to read in the data file and get rid of the redundant columns of CpG site descriptions. GetMethylation450Data function reads and processes the data file block by block to circumvent potential memory limitation problem caused by large sizes of Firehose data files (>10GB for some cancer types).

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,

# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download humanmethylation450 data of six rectum adenocarcinoma (READ) patient samples

Methylation450RawData <- DownloadMethylationData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"methylation_450", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-
6812", "TCGA-AG-A01W", "TCGA-AG-3731"))

# Process the downloaded humanmethylation450 data and save the results to

# ProcessedData.TCGA-Assembler subfolder.

Methylation450Data <- ProcessMethylation450Data(inputFilePath = Methylation450RawData[1],
outputFileName = "READ__humanmethylation450", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Process READ HumanMethylation450 data downloaded from Firehose website.

Methylation450Data <- ProcessMethylation450Data(inputFilePath =
"./ManualExampleData/RawData.Firehose/READ.methylation__humanmethylation450__jhu_usc_edu__
Level_3__within_bioassay_data_set_function__data.data.txt", outputFileName =
"READ__humanmethylation450", outputFileFolder = "./ManualExampleData/ProcessedData.Firehose",
fileSource = "Firehose")
```

ProcessmiRNASeqData

Usage

```
ProcessmiRNASeqData(inputFilePath, outputFileName, outputFileFolder, fileSource = "TCGA-Assembler")
```

Description

This function processes miRNASeq data file either acquired by TCGA-Assembler module A or downloaded from Firehose website, and imports the data into R. It does the following.

(1) Save the miRNASeq read count data and reads per million miRNA mapped (RPM) data as two separate tab-delimited .txt files. In each file, the first column is the miRNA name. The other columns are data of individual samples. The first row in the file gives the TCGA sample barcodes (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode). The filenames are composed of outputFileName and either "ReadCount.txt" or "RPM.txt", with double-underscore "__" separating the two.

(2) Save the miRNASeq read count data and RPM data as two separate R data file (.rda), each of which contains two variables. The first variable is Des, which is a single-column character matrix including the miRNA names. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in Data is a sample and each row corresponds to a miRNA. The filenames are composed of outputFileName and either "ReadCount.rda" or "RPM.rda", with double-underscore "__" separating the two parts.

For both read count and RPM, the .txt file and .rda file contain the same data. Having the same data in two different file formats is just for the convenience of using the data in different software environments.

Arguments

inputFilePath: a character string indicating the path of the input miRNAseq data file acquired by TCGA-Assembler module A or downloaded from Firehose website. It should be a tab-delimited .txt (or .txt) file.

outputFileName: a character string to form the names of output data files.

outputFileFolder: a character string indicating the directory to which the output files will be saved.

fileSource: a character string that can be either TCGA-Assembler or Firehose. TCGA-Assembler, which is the default setting, indicates that the input data file was acquired by TCGA-Assembler Module A and Firehose indicates that the input data file was downloaded from Firehose website.

Value

A list object of two variables Des and Data. Des is a single-column character matrix including miRNA names. It serves as the description of the second variable, Data, which is a numeric matrix of RPM values. Each column in Data corresponds to a sample and each row corresponds to a miRNA. And the column names are TCGA sample barcodes.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,

# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download miRNA-seq data of six rectum adenocarcinoma (READ) patient samples

miRNASeqRawData <- DownloadmiRNASeqData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"mir_HiSeq.hg19", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-
6812", "TCGA-AF-2689", "TCGA-AF-2691"))

# Process the downloaded READ miRNA-seq data and save the results to

# the ProcessedData.TCGA-Assembler subfolder.

miRNASeqData <- ProcessmiRNASeqData(inputFilePath = miRNASeqRawData[1], outputFileName =
"READ__illuminahiseq_mirnaseq", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-
Assembler")

# Process READ miRNA-seq data downloaded from Firehose website,

# and save the results to the ProcessedData.Firehose subfolder.

miRNASeqData <- ProcessmiRNASeqData(inputFilePath =
"./ManualExampleData/RawData.Firehose/READ.mirnaseq__illuminaga_mirnaseq__bcgsc_ca__Level_3
__miR_gene_expression__data.data.txt", outputFileName = "READ__illuminaga_mirnaseq",
outputFileFolder = "./ManualExampleData/ProcessedData.Firehose", fileSource = "Firehose")
```

ProcessRNASeqData

Usage

```
ProcessRNASeqData(inputFilePath, outputFileName, outputFileFolder, dataType, verType)
```

Description

This function processes RNASeq or microarray gene expression data files either acquired by TCGA-Assembler Module A or downloaded from Firehose website and imports data into R. Data that can be processed include (1) normalized gene expression data generated by RNASeqV2 pipeline, (2) exon expression data generated by RNASeqV2 pipeline, and (3) microarray gene expression data. RNASeqV2 indicates the RNASeq data processing pipeline used by TCGA. This function does the following processing.

(1) For gene expression data, check whether the gene symbols are official HGNC gene symbols. If not, correct them.

(2) Extract most useful measurements from the input data files. For RNASeqV2 normalized gene expression data, extract the normalized count values. For RNASeqV2 exon expression data, extract the RPKM values.

(3) Draw and save a box plot of gene expression data for quality control purpose. The picture filename is composed of outputFileName and "__boxplot.png".

(4) Save the extracted data as a tab-delimited .txt file. For gene expression data, the first two columns are gene symbol and Entrez ID. And the other columns are samples, with TCGA sample barcodes in the top row (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode). For microarray data, Entrez IDs are not available. For exon expression data, the first column is exon ID. And the other columns are samples, with TCGA sample barcodes in the top row.

(5) Save the extracted data as an R data file (.rda), which includes two variables Des and Data. Des is a character matrix including gene symbols and Entrez IDs for gene expression data, and exon IDs for exon expression data. For microarray data, Entrez IDs are not available. Des describes the second variable, Data, which is the numeric data matrix. Each column in Data is a sample and each row corresponds to a gene or exon. The column names of Data are TCGA sample barcodes.

Arguments

inputFilePath: a character string indicating the path of input RNASeq data file acquired by TCGA-Assembler Module A or downloaded from Firehose website. It should be a tab-delimited .txt (or .txt) file.

outputFileName: a character string to form the names of output files.

outputFileFolder: a character string indicating the directory in which the output files will be saved.

dataType: a character string indicating the type of input data file. geneExp indicates that the input data file includes gene expression data. exonExp indicates that the input data file is exon expression data. If verType is Microarray, then dataType must be geneExp.

verType: a character string indicating the assay pipeline used to generate the data. Options include RNASeqV2 and Microarray.

Value

A list object of two variables Des and Data. If dataType is geneExp, Des is a character matrix of two columns including gene symbols and Entrez IDs. For microarray data, Entrez IDs are not available. Des describes the second variable, Data, which is a numeric matrix of gene expressions. Each column in the matrix is a sample and each row corresponds to a gene. The column names of Data are TCGA sample barcodes. If dataType is exonExp, Des is a single-column character matrix of exon IDs. Data is a numeric matrix of exon expressions. Each column in the data matrix is a sample and each row corresponds to an exon.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,

# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download normalized gene expression data of six rectum

# adenocarcinoma (READ) patient samples, which were generated by RNASeqV2 pipeline.

RNASeqRawData <- DownloadRNASeqData(saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform = "gene.normalized_RNAseq", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-3732", "TCGA-AG-3742"))

# Process the downloaded normalized gene expression data and save the results to the

# ProcessedData.TCGA-Assembler subfolder.

GeneExpData <- ProcessRNASeqData(inputFilePath = RNASeqRawData[1], outputFileName = "READ__illuminahiseq_rnaseqv2__GeneExp", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler", dataType = "geneExp", verType = "RNASeqV2")

# Download exon expression data of six rectum

# adenocarcinoma (READ) patient samples, which were generated by RNASeqV2 pipeline.

RNASeqRawData <- DownloadRNASeqData(saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform = "exon_RNAseq", inputPatientIDs = c("TCGA-EI-6884", "TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-3732", "TCGA-AG-3742"))

# Process the downloaded exon expression data and save the results to the

# ProcessedData.TCGA-Assembler subfolder.
```

```
ExonExpData <- ProcessRNASeqData(inputFilePath = RNASeqRawData[1], outputFileName =
"READ__illuminahiseq_rnaseqv2__ExonExp", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler", dataType = "exonExp", verType =
"RNASeqV2")

# Download microarray gene expression data of all READ patient samples.

MicroarrayRawData <- DownloadRNASeqData(saveFolderName =
"./ManualExampleData/RawData.TCGA-Assembler", cancerType = "READ", assayPlatform =
"gene_Array")

# Process the downloaded microarray gene expression data.

GeneExpData <- ProcessRNASeqData(inputFilePath = MicroarrayRawData[1], outputFileName =
"READ__Microarray__GeneExp", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-
Assembler", dataType = "geneExp", verType = "Microarray")
```

ProcessRPPADataWithGeneAnnotation

Usage

```
ProcessRPPADataWithGeneAnnotation(inputFilePath, outputFileName, outputFileFolder)
```

Description

This function processes RPPA data files either acquired by TCGA-Assembler Module A or downloaded from Firehose website, and imports the data into R. It does the following.

(1) Split the gene symbol and protein antibody name into two separate columns.

(2) Check whether the gene symbols are official HGNC gene symbols. If not, correct them.

(3) When a protein is encoded by more than one gene, duplicate the measurement of the protein (a row in the data matrix) for each gene.

(4) Draw and save a box plot picture of the data for quality control purpose. The picture filename is composed of outputFileName and "__boxplot.png".

(5) Save protein expression data as a tab-delimited .txt file. The first column gives gene symbols. The second column gives protein antibody names. The other columns are data of individual samples. The first row gives TCGA sample barcodes (refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode).

(6) Save the processed data as an R data file (.rda), which includes two variables Des and Data. Des is a character matrix including two columns, i.e. gene symbol and protein antibody name. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in the data matrix is a sample and column names are TCGA sample barcodes.

The output .txt file and .rda file contain the same protein expression data. Having the same data in two different file formats is just for the convenience of using the data under different software environments.

Arguments

inputFilePath: a character string indicating the path of the input RPPA protein expression data file acquired by TCGA-Assembler Module A or downloaded from Firehose website. It should be a tab-delimited .txt (or .txt) file.

outputFileName: a character string to form the names of output data files and box plot picture file.

outputFileFolder: a character string indicating the path of the directory to which the output files will be saved.

Value

A list object of two matrix variables. The first variable is Des, which is a character matrix including two columns, i.e. gene symbol and protein antibody name. It serves as the description of the second variable,

Data, which is a numeric matrix. Each column in the data matrix is a sample and the column names are TCGA sample barcodes.

Details

Before applying the function, check the input data file to see whether gene symbols corresponding to the same protein antibody are separated by a single white space. Sometimes the gene symbols are not correctly separated due to errors inherited from the original TCGA data files.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,

# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Download RPPA protein expression data of six rectum adenocarcinoma (READ) patient samples

# and save the acquired data in ./ManualExampleData/RawData.TCGA-Assembler

RPPARawData <- DownloadRPPAData(saveFolderName = "./ManualExampleData/RawData.TCGA-
Assembler", cancerType = "READ", assayPlatform = "protein_RPPA", inputPatientIDs = c("TCGA-EI-6884",
"TCGA-DC-5869", "TCGA-G5-6572", "TCGA-F5-6812", "TCGA-AG-3582", "TCGA-AG-4001"))

# Process the downloaded protein expression data and save the results in

# ./ManualExampleData/ProcessedData.TCGA-Assembler

RPPAData <- ProcessRPPADataWithGeneAnnotation(inputFilePath = RPPARawData[1], outputFileName
= "READ__mda_rppa_core", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-
Assembler")

# Process RPPA protein expression data file downloaded from Firehose website.

RPPAData <- ProcessRPPADataWithGeneAnnotation(inputFilePath =
"./ManualExampleData/RawData.Firehose/READ.RPPA_AnnotateWithGene.txt", outputFileName =
"READ__mda_rppa_core", outputFileFolder = "./ManualExampleData/ProcessedData.Firehose")
```

ProcessSomaticMutationData

Usage

```
ProcessSomaticMutationData(inputFilePath, outputFileName, outputFileFolder)
```

Description

This function processes Somatic Mutation data files acquired by TCGA-Assembler Module A, and imports the data into R. It does the following.

(1) Check whether the gene symbols are official HGNC gene symbols. If not, correct them.

(2) Transform the mutation data into matrix format, where each row is a mutation and each column is a patient tumor/normal sample pair. The data matrix includes binary 0/1 elements, where 1 indicates a tumor sample has a mutation while 0 indicates not.

(3) Save the mutation data as a tab-delimited .txt file and an R data file (.rda), both with "mutationLevel" as the suffix of filenames. In the .txt file, the first 22 columns are mutation descriptions and the mutation data start from the 23rd column. The .rda file includes two variables Des and Data. Des is a character matrix including the 22 columns of mutation descriptions. It serves as the description of the second variable, Data, which is the mutation data.

(4) Generate gene-level somatic mutation data. In the data matrix, each row is a gene and each column is a patient tumor/normal sample pair. The matrix element is the total number of somatic mutations in the genomic region of a gene in a tumor sample.

(5) Save the gene-level mutation data as a tab-delimited .txt file and an R data file (.rda), both with "geneLevel" as the suffix of filenames. In the .txt file, the first column gives gene symbols; the second column gives classification information of variants in a gene. The other columns are data of individual tumor samples. The .rda file includes two variables Des and Data. Des is a character matrix including two columns, i.e. gene symbol and classification information of variants. It serves as the description of the second variable, Data, which is a numeric matrix of counts of somatic mutations at gene level. Each column in the data matrix is a tumor/normal sample pair and column names are TCGA sample barcodes.

Arguments

inputFilePath: a character string indicating the path of the input somatic mutation data file acquired by TCGA-Assembler Module A. It should be a tab-delimited .txt file.

outputFileName: a character string to form the names of output data files.

outputFileFolder: a character string indicating the path of the directory in which the output files will be saved.

Value

A list object of two matrix variables representing the gene-level mutation data. The first variable is Des, which is a character matrix including two columns, i.e. gene symbol and classification information of variants in a gene. It serves as the description of the second variable, Data, which is a numeric matrix of counts of somatic mutations at gene level. Each column in the data matrix is a sample and the column name is TCGA sample barcode.

Details

This function transforms the mutation data into matrix format and save them with "mutationLevel" as the suffix of filenames. It further generates gene-level mutation data and save them, where the data are indicated by "geneLevel" as the suffix of filenames.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Acquire somatic mutation data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_somatic <- DownloadSomaticMutationData(cancerType = "BRCA", assayPlatform =
"somaticMutation_DNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-E2-A1IU", "TCGA-A7-A13F", "TCGA-BH- A0BZ", "TCGA-BH-A18N", "TCGA-BH-
A18Q", "TCGA-BH-A18R" ))

# Process the downloaded BRCA somatic mutation data.

SomaticMutationData <- ProcessSomaticMutationData(inputFilePath = filename_BRCA_somatic[1],
outputFileName = "BRCA_MutationData", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")

# Acquire somatic mutation data of glioblastoma multiforme (GBM) patient samples.

filename_GBM_somatic <- DownloadSomaticMutationData(cancerType = "GBM", assayPlatform =
"somaticMutation_DNAseq", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Process the downloaded GBM somatic mutation data.

SomaticMutationData <- ProcessSomaticMutationData(inputFilePath = filename_GBM_somatic[1],
outputFileName = " GBM_MutationData", outputFileFolder =
"./ManualExampleData/ProcessedData.TCGA-Assembler")
```

ProcessCPTACData

Usage

```
ProcessCPTACData(inputFilePath, outputFileName, outputFileFolder)
```

Description

This function processes CPTAC proteome data files acquired by TCGA-Assembler Module A. It only works on proteome data acquired using function DownloadCPTACData with assay platform proteome_iTRAQ. This function does the following.

(1) Check whether the gene symbols are official HGNC gene symbols. If not, correct them.

(2) For breast invasive carcinoma (BRCA) and ovarian serous cystadenocarcinoma (OV), separate the data into two data matrices, i.e. Log Ratio data and Unshared Log Ratio data. For colorectal cancer, separate the data into two data matrices, i.e. Spectral Counts data and Unshared Spectral Counts data.

(3) Draw and save boxplot pictures of the data for quality control purpose. The picture filename is composed of outputFileName and "__boxplot.png". Boxplot of data with only unshared peptides is labeled with "unsharedPeptide".

(4) Save the processed data as two tab-delimited .txt files, one for data including all peptides of proteins, while the other including only unshared peptides of proteins with the filename indicated by "unsharedPeptide". In the data files, the 1st row is the column names and TCGA sample barcodes, while each of the other rows corresponds to a protein. The 1st column shows the gene symbol that encodes the protein. The 2nd column is gene description. The 3rd column is the organism. The 4th column is the chromosome ID. The 5th column is the genomic location of the gene. And starting from the 6th column, each column corresponds to a sample. For details of the iTRAQ technology, please refer to https://en.wikipedia.org/wiki/Isobaric_tag_for_relative_and_absolute_quantitation. For TCGA sample barcodes, refer to https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode.

(5) Save the processed data as two R data files (.rda), one for data including all peptides of proteins, while the other including only unshared peptides of proteins with the filename indicated by "unsharedPeptide". In the .rda file, there are two variables Des and Data. Des is a character matrix including five columns, i.e. gene symbol, gene description, organism, chromosome ID, and genomic location. It serves as the description of the second variable, Data, which is a numeric matrix. Each column in the data matrix is a sample and column names are TCGA sample barcodes.

Arguments

inputFilePath: a character string indicating the path of the input CPTAC proteome data file acquired by the function DownloadCPTACData with assay platform proteome_iTRAQ. It should be a tab-delimited .txt file.

outputFileName: a character string to form the names of output data files and box plot picture files.

outputFileFolder: a character string indicating the path of the directory to which the output files will be saved.

Value

A list object of two elements, i.e. allPeptides and unsharedPeptides, each of them is a list object of two matrix variables, which are Des and Data. allPeptides is the data considering all peptides and unsharedPeptides is the data considering only peptides uniquely mapping to one protein. The Des variables are character matrices including five columns, i.e. gene symbol, gene description, organism, chromosome ID and genomic location. They serve as the descriptions of the corresponding Data variables that are numeric data.

Details

This function processes CPTAC proteome data acquired using function DownloadCPTACData with assay platform proteome_iTRAQ and generates data tables for data of all peptides and data of unshared peptides, separately. For each case, the data is saved in a .txt file and an .rda file. The difference of the data files generated in the two cases is the filename, "allPeptides" for all peptides data while "unsharedPeptides" for unshared peptides data.

Examples

```
# The present working directory of R must be TCGA-Assembler, i.e. the package folder,
# for running the examples.

rm(list = ls())      # Clear workspace

source("Module_A.R")   # Load Module A functions

source("Module_B.R")   # Load Module B functions

# Acquire CPTAC proteome data of six breast invasive carcinoma (BRCA) patient samples.

filename_BRCA_iTRAQ <- DownloadCPTACData(cancerType= c("BRCA"), assayPlatform =
"proteome_iTRAQ", saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler",
inputPatientIDs = c("TCGA-A2-A0CM", "TCGA-A7-A13F", "TCGA-BH-A0BZ", "TCGA-BH-A18N", "TCGA- BH-
A18Q", "TCGA-BH-A18R"))

# Process the downloaded data.

CPTACData <- ProcessCPTACData(inputFilePath = filename_BRCA_iTRAQ[1], outputFileName =
"BRCA_iTRAQData", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler")

# Acquire CPTAC proteome data of ovarian serous cystadenocarcinoma (OV) patient samples.

filename_OV_iTRAQ <- DownloadCPTACData(cancerType = c("OV"), assayPlatform = "proteome_iTRAQ",
saveFolderName = "./ManualExampleData/RawData.TCGA-Assembler")

# Process the OV proteome data generated by Johns Hopkins University (JHU).

CPTACData <- ProcessCPTACData(inputFilePath = filename_OV_iTRAQ[1], outputFileName =
"OV_iTRAQData", outputFileFolder = "./ManualExampleData/ProcessedData.TCGA-Assembler")
```