Package ‘SCAN.UPC’

October 9, 2013

Type Package

Title Single-channel array normalization (SCAN) and University
Probability of expression Codes (UPC)

Version 2.0.2

Author Stephen R. Piccolo and W. Evan Johnson

Maintainer Stephen R. Piccolo <stephen.piccolo@hsc.utah.edu>

Description SCAN is a microarray normalization method to facilitate personalized-medicine workflows. Rather than processing microarray samples as groups, which can introduce biases and present logistical challenges, SCAN normalizes each sample individually by modeling and removing probe- and array-specific background noise using only data from within each array. SCAN can be applied to one-channel (e.g., Affymetrix) or two-channel (e.g., Agilent) microarrays. The Universal Probability of expression Codes (UPC) method is an extension of SCAN that generates probability-of-expression values. These values can be interpreted as the probability that a given genomic feature (e.g., gene, transcript) is expressed above the background in a given sample. The UPC method can be applied to one-channel or two-channel microarrays as well as to RNA-Seq read counts. Because UPC values are represented on the same scale and have an identical interpretation for each platform, they can be used for cross-platform data integration.)

License MIT

Depends R (>= 2.14.0), Biobase (>= 2.6.0), oligo, Biostrings

Suggests pd.hg.u95a

Imports utils, methods, MASS, tools

biocViews Software, Microarray, Preprocessing, RNAseq, TwoChannel,OneChannel

URL http://bioconductor.org, http://jlab.bu.edu/software/scan-upc

R topics documented:

SCAN *Single-Channel Array Normalization (SCAN)*

Description

This function is used to normalize single-channel expression microarrays via the SCAN method. In raw form, such microarray data come in the form of Affymetrix .CEL files.

Usage

```
SCAN(celFilePattern, outFilePath = NA, convThreshold = 0.01, probeSummaryPackage = NA,
     probeLevelOutDirPath = NA, verbose = TRUE)
```

Arguments

`celFilePattern` Absolute or relative path to the input file to be processed. To process multiple files, wildcard characters can be used (e.g., "*.CEL"). This is the only required parameter.

`outFilePath` Absolute or relative path where the output file will be saved. This is optional.

`convThreshold` Convergence threshold that determines at what point the mixture-model parameters have stabilized. The default value should be suitable in most cases. However, if the model fails to converge, it may be useful to adjust this value. (This parameter is optional.)

`probeSummaryPackage` An R package that specifies alternative probe/gene mappings. This is optional. See note below for more details.

`probeLevelOutDirPath` Absolute or relative path to a directory where probe-level normalized values can be saved. This is optional. By default, the probe-level values will be discarded after they have been summarized. However, if the user has a need to repeatedly process the same file (perhaps to try various probe/gene mappings), this option can be useful because SCAN will retrieve previously normalized values if a probe-level file exists, rather than renormalize the raw data. The user should be aware that probe-level files may consume a considerable amount of disk space.

`verbose` Whether to output more detailed status information as files are normalized. Default is TRUE.

Value

An ExpressionSet object that contains a row for each probeset/gene/transcript and a column for each input file.

Note

By default, SCAN uses the default mappings between probes and genes that have been provided by the manufacturer. However, these mappings may be outdated or may include problematic probes (for example, those that cross hybridize). The default mappings also may produce multiple summary values per gene. Alternative mappings, such as those provided by the BrainArray resource (see http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp), allow SCAN to produce a single value per gene and to use updated gene definitions. Users can specify alternative mappings using the probeSummaryPackage parameter. If specified, this package must conform to the standards of the AnnotationDbi package. The BrainArray packages can be downloaded from http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp. When using BrainArray, be sure to download the R source package for probe-level mappings (see vignette for more information).

Author(s)

Stephen R. Piccolo

References

Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, and Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 2012, 100:6, pp. 337-344.

Examples

```
# Download an example CEL file and save it as a temporary local file
celFilePath = file.path(tempdir(), "Vignette_Example.CEL.gz")
download.file("http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSM555237&format=file&file=GSM555237.CEL.gz", celF

# Normalize a CEL file
normalized = SCAN(celFilePath)

# Normalize a CEL file and save output to a file
normalized = SCAN(celFilePath, "output_file.txt")

## Not run:
# Normalize a CEL file and summarize at the gene level using BrainArray
# mappings for Entrez Gene
probeFilePath = file.path(tempdir(), "hgu95ahsentrezgprobe_15.0.0.tar.gz")
download.file("http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/15.0.0/entrezg.download/hgu9
install.packages(probeFilePath, repos=NULL, type="source")
library(hgu95ahsentrezgprobe)
normalized = SCAN(celFilePath, probeSummaryPackage=hgu95ahsentrezgprobe)

## End(Not run)
```

SCAN_TwoColor *Single-Channel Array Normalization (SCAN) for two-channel Agilent microarrays*

Description

This function is used to normalize two-channel expression microarrays (from Agilent) via the SCAN method. In raw form, such microarray data are tab-separate data files.

Usage

```
SCAN_TwoColor(inFilePattern, outFilePath = NA, verbose = TRUE)
```

Arguments

`inFilePattern`	Absolute or relative path to the input file to be processed. To process multiple files, wildcard characters can be used (e.g., "*.txt"). This is the only required parameter.
`outFilePath`	Absolute or relative path where the output file will be saved. This is optional.
`verbose`	Whether to output more detailed status information as files are normalized. Default is TRUE.

Value

A list is returned, containing two elements: a matrix containing normalized data values and a vector of probe names that correspond to each row of the matrix. The matrix will contain two columns—one corresponding to each channel—for each sample. When the array design contains duplicate probe names (this is common for control probes), the vector of probe names will also contain duplicates.

Author(s)

Stephen R. Piccolo

References

Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, and Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 2012, 100:6, pp. 337-344.

Examples

```
## Not run:
# Download an example file and save it as a temporary local file
inFilePath = file.path(tempdir(), "Vignette_Example.txt.gz")
download.file("http://www.ncbi.nlm.nih.gov/geosuppl/?acc=GSM1072833&file=GSM1072833

# Normalize the file and save output to a file
```

```
result = SCAN_TwoColor(inFilePath, "output_file.txt")

## End(Not run)
```

UPC *Universal Probability of expression Codes (UPC) for single-channel microarrays*

Description

This function is used to normalize single-channel expression microarrays via the UPC method. In raw form, such microarray data come in the form of Affymetrix .CEL files.

Usage

```
UPC(celFilePattern, outFilePath = NA, convThreshold = 0.01, probeSummaryPackage = NA,
     probeLevelOutDirPath = NA, verbose = TRUE)
```

Arguments

`celFilePattern` Absolute or relative path to the input file to be processed. To process multiple files, wildcard characters can be used (e.g., "*.CEL"). This is the only required parameter.

`outFilePath` Absolute or relative path where the output file will be saved. This parameter is optional.

`convThreshold` Convergence threshold that determines at what point the mixture-model parameters have stabilized. The default value should be suitable in most cases. However, if the model fails to converge, it may be useful to adjust this value. This parameter is optional.

`probeSummaryPackage` An R package that specifies alternative probe/gene mappings. This parameter is optional. See note below for more details.

`probeLevelOutDirPath` Absolute or relative path to a directory where probe-level normalized values can be saved. This parameter is optional. By default, the probe-level values will be discarded after they have been summarized. However, if the user has a need to repeatedly process the same file (perhaps to try various probe/gene mappings), this option can be useful because UPC will retrieve previously normalized values if a probe-level file exists, rather than renormalize the raw data. The user should be aware that probe-level files may consume a considerable amount of disk space.

`verbose` Whether to output more detailed status information as files are normalized. Default is TRUE.

Value

An ExpressionSet object that contains a row for each probeset/gene/transcript and a column for each input file.

Note

By default, UPC uses the default mappings between probes and genes that have been provided by the manufacturer. However, these mappings may be outdated or may include problematic probes (for example, those that cross hybridize). The default mappings also may produce multiple summary values per gene. Alternative mappings, such as those provided by the BrainArray resource (see http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp), allow UPC to produce a single value per gene and to use updated gene definitions. Users can specify alternative mappings using the probeSummaryPackage parameter. If specified, this package must conform to the standards of the AnnotationDbi package. The BrainArray packages can be downloaded from http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp. When using BrainArray, be sure to download the R source package for probe-level mappings (see vignette for more information).

Author(s)

Stephen R. Piccolo

References

Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, and Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 2012, 100:6, pp. 337-344.

Examples

```
# Download an example CEL file and save it as a temporary local file
celFilePath = file.path(tempdir(), "Vignette_Example.CEL.gz")
download.file("http://www.ncbi.nlm.nih.gov/geo/download/?acc=GSM555237&format=file&file=GSM555237.CEL.gz", celF

# Normalize a CEL file
normalized = UPC(celFilePath)

# Normalize a CEL file and save output to a file
normalized = UPC(celFilePath, "output_file.txt")

## Not run:
# Normalize a CEL file and summarize at the gene level using BrainArray
# mappings for Entrez Gene
probeFilePath = file.path(tempdir(), "hgu95ahsentrezgprobe_15.0.0.tar.gz")
download.file("http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/15.0.0/entrezg.download/hgu9
install.packages(probeFilePath, repos=NULL, type="source")

library(hgu95ahsentrezgprobe)
normalized = UPC(celFilePath, probeSummaryPackage=hgu95ahsentrezgprobe)

## End(Not run)
```

UPC_RNASeq *Universal Probability of expression Codes (UPC) for RNA-Seq data*

Description

This function is used to normalize RNA-Sequencing data via the UPC method. A necessary preliminary step is to generate read counts for each gene (or transcript or exon). A popular approach for accomplishing this is to use the Tophat short-read aligner (http://tophat.cbcb.umd.edu/), followed by application of the htseq-count tool (http://www-huber.embl.de/users/anders/HTSeq/).

Gene (or transcript or exon) values can be converted to UPCs using this function. The tab-separated data file should contain a row for each gene. The first column should contain the gene ID. The second column should contain the read counts (not RPKM/FPKM values). For example:

AAB 31 AAC 255

Most users will want to correct for length and GC content. In this case, a tab-separated annotation file that specifies length and number of GC bases should be included. The first column should contain the gene ID. The second column should contain the length of the gene. The third column should specify the number of number of G or C bases in the gene.

AAB 1767 640 AAC 2644 1039

Usage

```
UPC_RNASeq(inFilePattern, annotationFilePath, outFilePath = NA, modelType = "nn", convThreshold = 0.01
```

Arguments

`inFilePattern` Absolute or relative path to the input file to be processed. To process multiple files, wildcard characters can be used (e.g., "*.txt"). This is the only required parameter.

`annotationFilePath`
Absolute or relative path where the annotation file is located. This parameter is optional.

`outFilePath` Absolute or relative path where the output file will be saved. This is optional.

`modelType` Various models can be used for the mixture model to differentiate between active and inactive probes. The default is the normal-normal model (“nn”), which uses the normal distribution. Other available options are log-normal (“ln”) and negative-binomial (“nb”).

`convThreshold` Convergence threshold that determines at what point the mixture-model parameters have stabilized. The default value should be suitable in most cases. However, if the model fails to converge, it may be useful to adjust this value. (This parameter is optional.)

`verbose` Whether to output more detailed status information as files are normalized. Default is TRUE.

Value

An `ExpressionSet` object that contains a row for each probeset/gene/transcript and a column for each input file.

Author(s)

Stephen R. Piccolo

References

Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, and Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 2012, 100:6, pp. 337-344.

`UPC_TwoColor` *Universal Probability of expression Codes (UPC) for two-channel microarrays*

Description

This function is used to normalize two-channel expression microarrays (from Agilent) using the Universal Probability of expression Codes (UPC) approach. In raw form, such microarray data come in the form of tab-separate data files.

Usage

```
UPC_TwoColor(inFilePattern, outFilePath = NA, modelType="nn", convThreshold=0.01, verbose = TRUE)
```

Arguments

`inFilePattern` Absolute or relative path to the input file to be processed. To process multiple files, wildcard characters can be used (e.g., "*.txt"). (This is the only required parameter.)

`outFilePath` Absolute or relative path where the output file will be saved. (This parameter is optional.)

`modelType` Various models can be used for the mixture model to differentiate between active and inactive probes. The default is the normal-normal model (“nn”), which uses the normal distribution. Other available options are log-normal (“ln”) and negative-binomial (“nb”).

`convThreshold` Convergence threshold that determines at what point the mixture-model parameters have stabilized. The default value should be suitable in most cases. However, if the model fails to converge, it may be useful to adjust this value. (This parameter is optional.)

`verbose` Whether to output more detailed status information as files are processed. Default is TRUE.

Value

A list is returned, containing two elements: a matrix containing UPC values and a vector of probe names that correspond to each row of the matrix. The matrix will contain two columns—one corresponding to each channel—for each sample. When the array design uses duplicate probe names (this is common for control probes), the vector of probe names will also contain duplicates.

Note

By default, `UPC` uses the default mappings between probes and genes that have been provided by the manufacturer. However, these mappings may be outdated or may include problematic probes (for example, those that cross hybridize). The default mappings also may produce multiple summary values per gene. Alternative mappings, such as those provided by the BrainArray resource (see `http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/genomic_curated_CDF.asp`), allow `UPC` to produce a single value per gene and to use updated gene definitions. Users can specify alternative mappings using the `probeSummaryPackage` parameter. If specified, this package must conform to the standards of the `AnnotationDbi` package. The BrainArray packages can be downloaded from `http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF/CDF_download.asp`. When using BrainArray, be sure to download the R source package for probe-level mappings (see vignette for more information).

Author(s)

Stephen R. Piccolo

References

Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, and Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics*, 2012, 100:6, pp. 337-344.

Examples

```
## Not run:
# Download an example file and save it as a temporary local file
inFilePath = file.path(tempdir(), "Vignette_Example.txt.gz")
download.file("http://www.ncbi.nlm.nih.gov/geosuppl/?acc=GSM1072833&file=GSM1072833

# Normalize the file and save output to a file
result = UPC_TwoColor(inFilePath, "output_file.txt")

## End(Not run)
```

Index