Overview

UniPath provides robust statistical methods to represent every single cell using pathway and gene-set enrichment scores. It can be used with both single cell RNA-seq and single cell ATAC-seq profile with scalability for atlas scale data-sets. UniPath comes with several features like pseudo-temporal ordering using pathway scores and unconventional way of enumerating differences between two cell populations.

Introduction

This vignette gives a detailed account on the workflow of UniPath tool for analyzing single cell expression data and single cell open chromatin profiles in pathway domain. UniPath is a steadfast statistical method for getting important biological insights from single cells characterized in terms of pathway activity scores and studying temporal dynamics. UniPath is a scalable platform allowing pre-processing and analysis of thousands of single cells by exploiting heterogeneity among cells and uncovering biologically relevant pathways. UniPath can help users with accurate identification of cell types, signaling pathways and doublet cells. Besides these, the user can also perform clustering and pseudo temporal ordering of single cells in pathway space. This may allow the analysis of relevant pathways and genes on single cell lineage transitions or potency.

---

binorm Conversion of non-zero gene FPKM value into p-value using each cell mean and standard deviation

---

Description

Produces a matrix of p-values

Details

Based on the assumption that non-zero gene expression FPKM follows log normal distribution, non-zero gene expression data is converted into p-values using mean and standard deviation for individual cells.

Usage

```
binorm(x)
```

Arguments

x	Gene expression matrix

Value

n*p matrix of p-values

combine	Combine p-values using empirical browns method

Description

Produces a matrix of combined p-values

Details

Empirical browns method is used for combining p-values of genes in a gene-set. It combines p-values of genes which are dependent on each other.

Usage

combine(gene_file, expression_matrix,gnames,Pval1, thr=2)

Arguments

gene_file	Pathway annotation file/gene-set file
expression matrix	Gene expression matrix
gnames	Gene names of expression matrix
Pval1	P-values matrix obtained from binorm
thr	Based on threshold provided, those gene sets having number of genes greater than the threshold value will to be considered for covariance matrix calculation and combining of p values. Default value is 2.

Value

n*p matrix of combined p-values

adjust	Adjusting of combine p-values using null model

Description

Adjustment of combined p-values.

Details

Combined p-values are adjusted using null model to get final pathway scores. Null model helps in highlighting cell-type specific pathways.

Usage

adjust(combp,combp_ref)

Arguments

combp	Combined p-value matrix obtained using gene expression matrix
combp_ref	Combined p-value matrix obtained using null model

Value

A list contains:

adjpva
: n*p matrix of absolute p-values

adjpvaraw
: n*p matrix of raw adjusted p-values

adjpvalog
: n*p matrix of adjusted log transformed p-values

---

dist_clust hierarchical clustering of pathway score matrix

Description

Performs hierarchical clustering and gives clusters of samples or cells

Usage

```
dist_clust(pathwayscores,n)
```

Arguments

pathwayscores	Log transformed adjusted p-value matrix
n	Number of clusters required for pseudo temporal ordering

Value

A list contains:

distance
: Distance matrix

clusters
: Number of clusters

---

index Indices of top k nearest neighbor

Description

Produces matrix of indices of k-nearest neighbour

Usage

index(pathwayscores,k)

Arguments

pathwayscores	Log transformed adjusted p-value matrix
k	Number of top k nearest neighbour

Value

Matrix of indices of nearest neighbour

KNN Getting cluster numbers for each of the nearest neighbor of a cell

---

Description

Produces a matrix with cluster number for top nearest neighbors for each of the cell

Usage

KNN(pathwayscores,index,clusters)

Arguments

pathwayscores	Log transformed adjusted p-value matrix
clusters	Clusters obtained from hierarchical clustering

Value

n*p matrix having cluster or class number for each of the top nearest neighbor of individual cell

---

class1 Finding how many times nearest neighbor of cells in each class are belonging to different cluster or class.

Description

Produces a matrix with counts of cells belonging same cells have top k nearest neighbor

Usage

class1(clusters,KNN)

Arguments

clusters	Clusters obtained from hierarchical clustering
KNN	Matrix with cluster number for top k nearest neighbors for each of the cell

Value

n*n matrix with number of times cells in same class have top k neighbors in other classes

---

distance Shrinked distance matrix based on two level of shrinkage

---

Description

Two level shrinkage of distance matrix based on nearest neighbour indices and belongingness of cells to same class

Usage

distance(dist,class,clusters)

Arguments

dist	Distance matrix used for hierarchical clustering
class	Matrix with number of times cells in same class have top k neighbors in other classes
clusters	Clusters obtained from hierarchical clustering

Value

Shrinked distance matrix

---

minimum_spanning_tree Construction of minimum spanning tree

Description

Finds minimum spanning tree by creating adjacency graph using shrinked distance matrix

Usage

minimum_spanning_tree(distance)

Arguments

distance	shrinked distance matrix

Value

Minimum spanning tree

---

mst.plot.mod Plotting minimum spanning tree using netbiov package in R

Usage

```
UniPath::mst.plot.mod()
```

---

temporaldif Differential pathways

Description

Differential pathways based on Wilcoxon rank sum test

Usage

```
temporaldif(data, group)
```

Arguments

data	Adjusted raw p-values matrix
group	Group of cell types among which differential pathway analysis needs to be performed

Value

A list containing p-value based on Wilcoxon rank sum test, fold change based on mean and median

---

gradient Creates gradient of colors

Description

Creating gradient of colors for showing continuum of single pathway on minimum spanning tree

Usage

gradient(pathwayfile,term)

Arguments

pathwayfile	Adjusted p-values matrix
term	Pathway term for which gradient needs to be plotted

Value

Gradient of colors for specific pathway term

---

makecolor Creates gradient of colors

Description

Creating gradient of colors for showing continuum of co-occurrence of pathways on minimum spanning tree

Usage

makecolor (score1, score2)

Arguments

score1	scores of first pathway term
score2	scores of second pathway term

Value

Matrix of gradient of colors for two pathway terms to be visualized on minimum spanning tree

---

Counttofpkm Conversion of read count data to fpkm

Description

Converts raw read count data into fpkm

Usage

Counttofpkm (countMatrix,length)

Arguments

countMatrix	Raw read count matrix
length	length of genes. Order of length of genes and row names of count data should be same

Value

n*p matrix fpkm values

---

difcoccur Differential cooccurrence of pathways among two group of cells

---

Description

Differential cooccurrence pathway analysis

Usage

difcoccur(data , group)

Arguments

data	Adjusted raw p-values matrix
group	Group 1 for cells of interest and rest in group 2

Value

A list contains:

pval

n*n matrix of p-values of pathway pairs

dif

n*n matrix of p-values of pathway pairs

drimpute Imputation

Description

Imputation of scATAC-seq profiles

Usage

drimpute(countFile)

Arguments

countFile scATAC-seq count matrix

Value

Imputed count matrix

global_access Calculating global accessibility score

Description

Computes global accessibility scores based on bulk open chromatin profiles

Usage

global_access**(**testfile,referencefile,globalaccess_scores)

Arguments

testfile Peak list of test data
referencefile Peak list of reference data
globalaccess_scores pre-calculated global accessibility scores

Value

Matrix of global accessibility scores for test data

nearest_gene Generation of foreground file

Description

Produces a four column file having genes along with their genomic distances.

Usage

nearest_gene (arg1,arg2,arg3,arg4)

Arguments

arg1	nearestGenes.pl script
arg2	genomic coordinate file
arg3	human reference genome file
arg4	output file

Value

Matrix having genes along with their genomic distances

---

runGO Calculating pathway enrichment scores for scATAC-seq profiles

---

Description

Produces matrix of hypergeometric and binomial test based pathway enrichment scores respectively.

Details

UniPath can also help is calculating pathway enrichment scores using scATAC-seq profiles using this function. It gives users to calculate enrichment scores using hypergeometric and binomial test with two options for normalization of data to highlight cell type specific enhancers. User can choose option 1 for normalization using global accessibility scores or option number 2 for local accessibility scores-based normalization of data.

Usage

runGO(gmtFile,BGfile,countFile,method,globalaccessibility_scores,FGfile,promoters= FALSE, dist=1000000,threshold=1.25)

Arguments

gmtFile	gene set file or cell marker-based file
BGfile	background file
countFile	scATAC-seq count matrix

method	If method is chosen as 1, data normalization is performed using global accessibility scores. If selected method is 2 then local accessibility score-based normalization is performed
globalaccessibility_scores	global accessibility scores
FGfile	foreground file
Promoters	whether promoters to be used or not for conversion of scATAC-seq profiles to pathway scores. Default is false
dist	distance to be used for considering nearest gene to a peak. Default is 1000000
threshold	Peaks above the given threshold value are chosen to be in a foreground set

Value

A list contains:

hypergeometric

n`*p matrix of p-values based on hypergeometric test

binomial

n*p matrix of p-values based on binomial test