Package 'POWSC'

October 18, 2024

Type Package

Title Simulation, power evaluation, and sample size recommendation for single cell RNA-seq

Version 1.13.0

Description Determining the sample size for adequate power to detect statistical significance is a crucial step at the design stage for high-throughput experiments.
Even though a number of methods and tools are available for sample size calculation for microarray and RNA-seq in the context of differential expression (DE),
this topic in the field of single-
cell RNA sequencing is understudied. Moreover, the unique data characteristics present in scRNA-seq such as sparsity and heterogeneity increase the challenge.
We propose POWSC, a simulation-
based method, to provide power evaluation and sample size recommendation for single-
cell RNA sequencing DE analysis.
POWSC consists of a data simulator that creates realistic expression data, and a power assessor that provides a comprehensive evaluation
and visualization of the power and sample size relationship.

License GPL-2

Encoding UTF-8

LazyData true

biocViews DifferentialExpression, ImmunoOncology, SingleCell, Software

Depends R (>= 4.1), Biobase, SingleCellExperiment, MAST

Imports pheatmap, ggplot2, RColorBrewer, grDevices, SummarizedExperiment, limma

Suggests rmarkdown, knitr, testthat (>= 3.0.0), BiocStyle

VignetteBuilder knitr

RoxygenNote 7.1.1

git_url https://git.bioconductor.org/packages/POWSC

git_branch devel

git_last_commit 2cefe0c

git_last_commit_date 2024-04-30

Repository Bioconductor 3.20

Date/Publication 2024-10-17

Author Kenong Su [aut, cre], Hao Wu [aut]

Maintainer Kenong Su <kenong.su@emory.edu>

Contents

`Est2Phase` *Estimate characterized parameters for a given scRNA-seq data (SingleCellExperiment object or a count matrix).*

Description

These parameters include four gene-wise parameters and two cell-wise parameters.

Usage

```
Est2Phase(sce, low.prob = 0.99)
```

Arguments

`sce`	SingleCellExperiment object with assays(sce)[[1]] is the count matrix or input directly
`low.prob`	lower bound probability for phase I

Value

a list of needed estimated parameters

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(123)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
estParas = Est2Phase(sce)
```

es_mef_sce *sample data for POWSC*

Description

This dataset is obtained from GEO under accession number GSE29087. It is generated to profile the transcriptomes for 92 single cells consisting of mouse embryonic fibroblast (MEF) and embryonic stem (ES) cells (Islam et al., 2011). The average sequencing depth for the dataset is around half a million.

Usage

```
es_mef_sce
```

Format

A singlecellexperiment object contain the expressin data with two cell types

References

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., and Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome research, 21(7), 1160–1167.

plot_POWSC *plot the result use visualization.*

Description

plot the result use visualization.

Usage

```
plot_POWSC(POWSCobj, Form = c("I", "II"), Cell_Type = c("PW", "Multi"))
```

Arguments

POWSCobj	a POWSC object from runPOWSC
Form	choose from "I" or "II".
Cell_Type	choose from "PW" or "Multi"

Value

for multiple comparison cases, return the pheatmap; for the pairwise comparison, return the ggplot object.

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(12)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
est_Paras = Est2Phase(sce)
sim_size = c(100, 200) # A numeric vector
pow_rslt = runPOWSC(sim_size = sim_size, est_Paras = est_Paras,per_DE=0.05, DE_Method = "MAST", Cell_Type = "PW
plot_POWSC(pow_rslt, Form="II", Cell_Type = "PW") # Alternatively, we can use Form="I")
```

`Power_Cont` *Run DE analysis by using MAST. Here we output two result tables corresponding to two forms of DE genes. These parameters include four gene-wise parameters and two cell-wise parameters.*

Description

Run DE analysis by using MAST. Here we output two result tables corresponding to two forms of DE genes. These parameters include four gene-wise parameters and two cell-wise parameters.

Usage

```
Power_Cont(
  DErslt,
  simData,
  alpha = 0.1,
  delta = 0.5,
  strata = c(0, 10, 2^(seq_len(4)) * 10, Inf)
)
```

Arguments

`DErslt`	is from the DE analysis by MAST
`simData`	is the corresponding simulated scRNA-seq dataset (SingCellExperiment)
`alpha`	is the cutoff for the fdr which can be modified
`delta`	or the lfc is the cutoff (=0.5) used to determined the high DE genes for Form II
`strata`	can be modified by the user. By default, it is (0, 10], (10, 20], (20, 40], (40, 80], (80, Inf]

Value

a list of metrics for power analysis such as: stratified targeted power and marginal power.

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(123)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
```

```
estParas = Est2Phase(sce)
simData = Simulate2SCE(n=500, estParas1 = estParas, estParas2 = estParas)
DErslt = runDE(simData$sce)
Cont_pow = Power_Cont(DErslt, simData)
```

`Power_Disc` *Run DE analysis by using MAST. Here we output two result tables corresponding to two forms of DE genes. These parameters include four gene-wise parameters and two cell-wise parameters.*

Description

Run DE analysis by using MAST. Here we output two result tables corresponding to two forms of DE genes. These parameters include four gene-wise parameters and two cell-wise parameters.

Usage

```
Power_Disc(
  DErslt,
  simData,
  alpha = 0.1,
  delta = 0.1,
  strata = seq(0, 1, by = 0.2)
)
```

Arguments

`DErslt`	is from the DE analysis by MAST
`simData`	is the corresponding simulated scRNA-seq dataset (SingCellExperiment)
`alpha`	is the cutoff for the fdr which can be modified
`delta`	or the zero ratio change is the cutoff (=0.1) used to determined the high DE genes for Form II
`strata`	can be modified by the user. By default, it is (0, 0.2], (0.2, 0.4], (0.4, 0.6], (0.6, 0.8], (0.8, 1]

Value

a list of metrics for power analysis such as: stratified targeted power and marginal power.

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(123)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
estParas = Est2Phase(sce)
simData = Simulate2SCE(n=1000, estParas1 = estParas, estParas2 = estParas)
DErslt = runDE(simData$sce)
Disc_pow = Power_Disc(DErslt, simData)
```

runDE *A wrapper function for calling DE genes. This contains two methods: MAST and SC2P*

Description

A wrapper function for calling DE genes. This contains two methods: MAST and SC2P

Usage

```
runDE(sce, DE_Method = c("MAST", "SC2P"))
```

Arguments

sce	is a simulated scRNA-seq dataset with two-group conditions, e.g., treatment vs control.
DE_Method	is a string chosen from "MAST" or "SC2P".

Value

a list of three tables: the first table summaries the DE result for both forms of DE genes. cont table represents the result for continous case. disc table shows the result for discontinous case.

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(123)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
estParas = Est2Phase(sce)
simData = Simulate2SCE(n=100, estParas1 = estParas, estParas2 = estParas)
sim_sce = simData$sce
DErslt = runDE(sim_sce)
```

runMAST *Run DE analysis by using MAST. Here we output two result tables corresponding to two forms of DE genes. These parameters include four gene-wise parameters and two cell-wise parameters.*

Description

Run DE analysis by using MAST. Here we output two result tables corresponding to two forms of DE genes. These parameters include four gene-wise parameters and two cell-wise parameters.

Usage

```
runMAST(sce)
```

Arguments

sce	is a simulated scRNA-seq dataset with two-group conditions, e.g., treatment vs control.

Value

a list of three tables: the first table summaries the DE result for both forms of DE genes. cont table represents the result for continous case. disc table shows the result for discontinous case.

---

runPOWSC *Estimate characterized parameters for a given scRNA-seq data (SingleCellExperiment object or a count matrix).*

Description

These parameters include four gene-wise parameters and two cell-wise parameters.

Usage

```
runPOWSC(
  sim_size = c(50, 100, 200, 800, 1000),
  per_DE = 0.05,
  est_Paras,
  DE_Method = c("MAST", "SC2P"),
  Cell_Type = c("PW", "Multi"),
  multi_Prob = NULL,
  alpha = 0.1,
  disc_delta = 0.1,
  cont_delta = 0.5
)
```

Arguments

sim_size	a list of numbers
per_DE	the percentage of the DE genes.
est_Paras	the template parameter estimated from one cell type.
DE_Method	is a string chosen from "MAST" or "SC2P".
Cell_Type	is a string corresponding to the 1st scenario: same cell type comparison, and 2nd scenario: multiple cell types.
multi_Prob	is the mixture cell proportions which sum up to 1. If not summing up to 1, then the package will internally do the normalization procedure.
alpha	is the cutoff for the fdr which can be modified
disc_delta	or the zero ratio change is the cutoff (=0.1) used to determined the high DE genes for Form II.
cont_delta	or the lfc is the cutoff (=0.5) used to determined the high DE genes for Form II.

Value

POWSC object

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(12)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
est_Paras = Est2Phase(sce)
sim_size = c(100, 200) # A numeric vector
pow_rslt = runPOWSC(sim_size = sim_size, est_Paras = est_Paras,per_DE=0.05, DE_Method = "MAST", Cell_Type = "PW
```

---

runSC2P *Run DE analysis by using SC2P. Here we output two result tables corresponding to two forms of DE genes.*

Description

Run DE analysis by using SC2P. Here we output two result tables corresponding to two forms of DE genes.

Usage

```
runSC2P(sce)
```

Arguments

sce	is a simulated scRNA-seq dataset with two-group conditions, e.g., treatment vs control.

Value

a list of three tables: the first table summaries the DE result for both forms of DE genes. cont table represents the result for continous case. disc table shows the result for discontinous case.

---

sce *sample data for GSE67835*

Description

This dataset is obtained from GEO under accession number GSE67835. It includes 7 patients about 466 cells to capture the cellular complexity of the adult and fetal human brain at a whole transcriptome level. Healthy adult temporal lobe tissue was obtained from epileptic patients during temporal lobectomy for medically refractory seizures.

Usage

```
sce
```

Format

A singlecellexperiment object contain the expressions data with three cell types about patient AS_7

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67835

References

Darmanis, Spyros, et al. "A survey of human brain transcriptome diversity at the single cell level." Proceedings of the National Academy of Sciences 112.23 (2015): 7285-7290.

`Simulate2SCE` *Simulate the data for two-group comparison; e.g., treatment v.s. control It simulates the DE changes in two forms corresponding two types of DE genes*

Description

Simulate the data for two-group comparison; e.g., treatment v.s. control It simulates the DE changes in two forms corresponding two types of DE genes

Usage

```
Simulate2SCE(n = 100, perDE = 0.05, estParas1, estParas2)
```

Arguments

`n`	the number of total cells for two groups
`perDE`	percentage of DE genes
`estParas1`	the set of parameters corresponding to cell type I
`estParas2`	the set of parameters corresponding to cell type II

Value

a list of metrics recording the changes in the generated data: such as the DE gene indices for Form I and II DE genes, and simulated expression data in singlecellexperiment format.

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(123)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
estParas = Est2Phase(sce)
simData = Simulate2SCE(n=100, estParas1 = estParas, estParas2 = estParas)
```

`SimulateMultiSCEs` *Simulate the data for multiple-group comparisons; e.g., different cell types in blood It simulates the DE changes in two forms corresponding two types of DE genes*

Description

Simulate the data for multiple-group comparisons; e.g., different cell types in blood It simulates the DE changes in two forms corresponding two types of DE genes

Usage

```
SimulateMultiSCEs(
  n = 1000,
  estParas_set,
  multiProb,
  delta1 = 0.1,
  delta2 = 0.5
)
```

Arguments

`n`	the number of total cells for multiple groups; e.g., 1000, 2000, and etc.
`estParas_set`	a set of parameters corresponding to different cell types.
`multiProb`	a vector of probilities correponding to each cell type. It is not necessary to sum up to 1 because POWSC will normalize internally.
`delta1`	the minimum of expression change used to determine the Form I DE.
`delta2`	the minimum of log fold change used to determine the Form II DE.

Value

a list of simulated datasets. Each dataset corresponds to a pair-wise comparison including a series of metrics such as the DE gene indices for Form I and II DE genes, and simulated expression data in singlecellexperiment format.

Examples

```
data("es_mef_sce")
set.seed(123)
rix = sample(1:nrow(es_mef_sce), 500)
es_mef_sce = es_mef_sce[rix, ]
sce1 = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
estParas1 = Est2Phase(sce1)
sce2 = es_mef_sce[, colData(es_mef_sce)$cellTypes == "stemCell"]
estParas2 = Est2Phase(sce2)
estParas_set = list(celltype1 = estParas1, celltype2 = estParas1, celltype3 =estParas2)
multiProb = c(0.2, 0.3, 0.5)
simData = SimulateMultiSCEs(n=200, estParas_set = estParas_set, multiProb = multiProb)
```

summary_POWSC *summary of the result*

Description

summary of the result

Usage

```
summary_POWSC(POWSCobj, Form = c("I", "II"), Cell_Type = c("PW", "Multi"))
```

Arguments

POWSCobj	a POWSC object from runPOWSC
Form	choose from "I" or "II".
Cell_Type	choose from "PW" or "Multi".

Value

return the summary of the power including stratified, marginal, and overall power.

Examples

```
data("es_mef_sce")
sce = es_mef_sce[, colData(es_mef_sce)$cellTypes == "fibro"]
set.seed(12)
rix = sample(1:nrow(sce), 500)
sce = sce[rix, ]
est_Paras = Est2Phase(sce)
sim_size = c(100, 200) # A numeric vector
pow_rslt = runPOWSC(sim_size = sim_size, est_Paras = est_Paras,per_DE=0.05, DE_Method = "MAST", Cell_Type = "PW
summary_POWSC(pow_rslt, Form="II", Cell_Type = "PW")
```

Index