Package ‘DeMixT’

January 31, 2020

Title  Cell type-specific deconvolution of heterogeneous tumor samples with two or three components using expression data from RNAseq or microarray platforms

Version 1.2.3

Date 2020-01-06

Author Zeya Wang <zw17.rice@gmail.com>, Shaolong Cao<scao@mdanderson.org>, Wenyi Wang <wwang7@mdanderson.org>

Maintainer Shaolong Cao<scao@mdanderson.org>, Peng Yang <pyang7@mdanderson.org>

Description DeMixT is a software package that performs deconvolution on transcriptome data from a mixture of two or three components.

LazyData TRUE

Depends R (>= 3.6.0), parallel, Rcpp (>= 1.0.0), SummarizedExperiment, knitr, KernSmooth, matrixcalc

Imports matrixStats, stats, truncdist, base64enc, ggplot2

LinkingTo Rcpp

NeedsCompilation yes

VignetteBuilder knitr

biocViews Software, StatisticalMethod, Classification, GeneExpression, Sequencing, Microarray, TissueMicroarray, Coverage

License GPL-3

RoxygenNote 7.0.2

R topics documented:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeMixT</td>
<td>2</td>
</tr>
<tr>
<td>DeMixT_GS</td>
<td>5</td>
</tr>
<tr>
<td>DeMixT_S1</td>
<td>8</td>
</tr>
<tr>
<td>DeMixT_S2</td>
<td>10</td>
</tr>
<tr>
<td>Optimum_KernelC</td>
<td>12</td>
</tr>
<tr>
<td>simulate_2comp</td>
<td>15</td>
</tr>
<tr>
<td>simulate_3comp</td>
<td>16</td>
</tr>
<tr>
<td>test.data.2comp</td>
<td>17</td>
</tr>
<tr>
<td>test.data.3comp</td>
<td>18</td>
</tr>
</tbody>
</table>

Index 20
DeMixT

**Description**

DeMixT is a software that performs deconvolution on transcriptome data from a mixture of two or three components.

**Usage**

```r
DeMixT(
  data.Y,
  data.N1,
  data.N2 = NULL,
  niter = 10,
  nbin = 50,
  if.filter = TRUE,
  filter.sd = 0.5,
  ngene.selected.for.pi = NA,
  mean.diff.in.CM = 0.25,
  nspikein = NULL,
  gene.selection.method = "GS",
  ngene.Profile.selected = NA,
  tol = 10^(-5),
  output.more.info = FALSE,
  pi01 = NULL,
  pi02 = NULL,
  nthread = parallel::detectCores() - 1
)
```

**Arguments**

- `data.Y` A SummarizedExperiment object of expression data from mixed tumor samples. It is a $G \times M_y$ matrix where $G$ is the number of genes and $M_y$ is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

- `data.N1` A SummarizedExperiment object of expression data from reference component 1 (e.g., normal). It is a $G \times M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1.

- `data.N2` A SummarizedExperiment object of expression data from additional reference samples. It is a $G \times M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2. Component 2 is needed only for running a three-component model.

- `niter` The maximum number of iterations used in the algorithm of iterated conditional modes. A larger value better guarantees the convergence in estimation but increases the running time. The default is 10.

- `nbin` The number of bins used in numerical integration for computing complete likelihood. A larger value increases accuracy in estimation but increases the running time, especially in a three-component deconvolution problem. The default is 50.
DeMixT

if.filter  The logical flag indicating whether a predetermined filter rule is used to select genes for proportion estimation. The default is TRUE.

filter.sd  The cut-off for the standard deviation of lognormal distribution. Genes whose log transferred standard deviation smaller than the cut-off will be selected into the model. The default is 0.5.

gene.selected.for.pi  The percentage or the number of genes used for proportion estimation. The difference between the expression levels from mixed tumor samples and the known component(s) are evaluated, and the most differential expressed genes are selected, which is called S1. It is enabled when if.filter = TRUE. The default is \( \min(1500, 0.3 \times My) \), where \( My \) is the number of mixed sample. Users can also try using more genes, ranging from \( 0.3 \times My \) to \( 0.5 \times My \), and evaluate the outcome.

mean.diff.in.CM  Threshold of expression difference for selecting genes in the component merging strategy. We merge three-component to two-component by selecting genes with similar expressions for the two known components. Genes with the mean differences less than the threshold will be selected for component merging. It is used in the three-component setting, and is enabled when if.filter = TRUE. The default is 0.25.

nspikein  The number of spikes in normal reference used for proportion estimation. The default value is \( \min(200, 0.3 \times My) \), where \( My \) the number of mixed samples. If it is set to 0, proportion estimation is performed without any spike in normal reference.

gene.selection.method  The method of gene selection used for proportion estimation. The default method is ‘GS’, which applies a profile likelihood based method for gene selection. If it is set to ‘S1’, the most differential expressed genes are selected.

gene.Profile.selected  The number of genes used for proportion estimation ranked by profile likelihood. The default is \( \min(1500, 0.1 \times My) \), where \( My \) is the number of mixed samples. This is enabled only when gene.selection.method is set to ‘GS’.

tol  The convergence criterion. The default is \( 10^{-5} \).

output.more.info  The logical flag indicating whether to show the estimated proportions in each iteration in the output.

pi01  Initialized proportion for first known component. The default is Null and pi01 will be generated randomly from uniform distribution.

pi02  Initialized proportion for second known component. pi02 is needed only for running a three-component model. The default is Null and pi02 will be generated randomly from uniform distribution.

nthread  The number of threads used for deconvolution when OpenMP is available in the system. The default is the number of whole threads minus one. In our no-OpenMP version, it is set to 1.

Value

pi  A matrix of estimated proportion. First row and second row corresponds to the proportion estimate for the known components and unknown component respectively for two or three component settings, and each column corresponds to one sample.
pi.iter Estimated proportions in each iteration. It is a niter * My * p array, where p is the number of components. This is enabled only when output.more.info = TRUE.

ExprT A matrix of deconvolved expression profiles corresponding to T-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample.

ExprN1 A matrix of deconvolved expression profiles corresponding to N1-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample.

ExprN2 A matrix of deconvolved expression profiles corresponding to N2-component in mixed samples for a given subset of genes in a three-component setting. Each row corresponds to one gene and each column corresponds to one sample.

Mu A matrix of estimated $\mu$ of log2-normal distribution for both known ($\mu_{N1}, \mu_{N2}$) and unknown component ($\mu_T$). Each row corresponds to one gene.

Sigma Estimated $\sigma$ of log2-normal distribution for both known ($\sigma_{N1}, \sigma_{N2}$) and unknown component ($\sigma_T$). Each row corresponds to one gene.

gene.name The names of genes used in estimating the proportions. If no gene names are provided in the original data set, the genes will be automatically indexed.

Author(s)
Zeya Wang, Wenyi Wang

References

See Also
http://bioinformatics.mdanderson.org/main/DeMixT

Examples
# Example 1: simulated two-component data by using GS(gene selection method)
data(test.data.2comp)
# res <- DeMixT(data.Y = test.data.2comp$data.Y,
data.N1 = test.data.2comp$data.N1,#
data.N2 = NULL, nspikein = 50,#
gene.selection.method = 'GS',#
niter = 10, nbin = 50, if.filter = TRUE,#
ngene.selected.for.pi = 150,#
mean.diff.in.CM = 0.25, tol = 10^(-5))# res$pi
# head(res$ExprT, 3)
# head(res$ExprN1, 3)
# head(res$Mu, 3)
# head(res$Sigma, 3)#
# Example 2: simulated two-component data by using S1(gene selection method)
data(test.data.2comp)
# res <- DeMixT(data.Y = test.data.2comp$data.Y,#
data.N1 = test.data.2comp$data.N1,
DeMixT_GS

Estimates the proportions of mixed samples for each mixing component using profile likelihood gene selection

Description

This function is designed to estimate the proportions of all mixed samples for each mixing component with a new proposed profile likelihood based gene selection, which can select most identifiable genes as reference gene sets to achieve better model fitting quality. We first calculated the Hessian matrix of the parameter spaces and then derive the confidence interval of the profile likelihood of each gene. We then utilized the length of confidence interval as a metric to rank the identifiability of genes. As a result, the proposed gene selection approach can improve the tumor-specific transcripts proportion estimation.

Usage

DeMixT_GS(
  data.Y,  
data.N1,  
data.N2 = NULL,  
niter = 10,  
nbin = 50,  
if.filter = TRUE,  
ngene.Profile.selected = NA,  
ngene.selected.for.pi = NA,  
mean.diff.in.CM = 0.25,  
nspikein = NULL,  
tol = 10^-5,  
pi01 = NULL,  
pi02 = NULL,  
nthread = parallel::detectCores() - 1
)
Arguments

data.Y A SummarizedExperiment object of expression data from mixed tumor samples. It is a $G \times M_y$ matrix where $G$ is the number of genes and $M_y$ is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

data.N1 A SummarizedExperiment object of expression data from reference component 1 (e.g., normal). It is a $G \times M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1.

data.N2 A SummarizedExperiment object of expression data from additional reference samples. It is a $G \times M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2. Component 2 is needed only for running a three-component model.

niter The maximum number of iterations used in the algorithm of iterated conditional modes. A larger value better guarantees the convergence in estimation but increases the running time. The default is 10.

nbin The number of bins used in numerical integration for computing complete likelihood. A larger value increases accuracy in estimation but increases the running time, especially in a three-component deconvolution problem. The default is 50.

if.filter The logical flag indicating whether a predetermined filter rule is used to select genes for proportion estimation. The default is TRUE.

filter.sd The cut-off for the standard deviation of lognormal distribution. Genes whose log transferred standard deviation smaller than the cut-off will be selected into the model. The default is TRUE.

ngene.Profile.selected The number of genes used for proportion estimation ranked by profile likelihood. The default is $\min(1500, 0.1 \times M_y)$, where $M_y$ is the number of mixed samples.

ngene.selected.for.pi The percentage or the number of genes used for proportion estimation. The difference between the expression levels from mixed tumor samples and the known component(s) are evaluated, and the most differential expressed genes are selected, which is called S1. It is enabled when if.filter = TRUE. The default is $\min(1500, 0.3 \times M_y)$, where $M_y$ is the number of mixed sample. Users can also try using more genes, ranging from $0.3 \times M_y$ to $0.5 \times M_y$, and evaluate the outcome.

mean.diff.in.CM Threshold of expression difference for selecting genes in the component merging strategy. We merge three-component to two-component by selecting genes with similar expressions for the two known components. Genes with the mean differences less than the threshold will be selected for component merging. It is used in the three-component setting, and is enabled when if.filter = TRUE. The default is 0.25.

nspikein The number of spikes in normal reference used for proportion estimation. The default value is $\min(200, 0.3 \times M_y)$, where $M_y$ the number of mixed samples. If it is set to 0, proportion estimation is performed without any spike in normal reference.

tol The convergence criterion. The default is $10^{-5}$.

pi01 Initialized proportion for first known component. The default is Null and pi01 will be generated randomly from uniform distribution.
pi02  Initialized proportion for second known component. pi02 is needed only for running a three-component model. The default is Null and pi02 will be generated randomly from uniform distribution.

nthread  The number of threads used for deconvolution when OpenMP is available in the system. The default is the number of whole threads minus one. In our no-OpenMP version, it is set to 1.

Value

pi  A matrix of estimated proportion. First row and second row corresponds to the proportion estimate for the known components and unknown component respectively for two or three component settings, and each column corresponds to one sample.

pi.iter  Estimated proportions in each iteration. It is a $niter \times My \times p$ array, where $p$ is the number of components. This is enabled only when output.more.info = TRUE.

gene.name  The names of genes used in estimating the proportions. If no gene names are provided in the original data set, the genes will be automatically indexed.

Note

A Hessian matrix file will be created in the working directory and the corresponding Hessian matrix with an encoded name from the mixed tumor sample data will be saved under this file. If a user reruns this function with the same dataset, this Hessian matrix will be loaded in place of running the profile likelihood method and reduce running time.

Author(s)

Shaolong Cao, Zeya Wang, Wenyi Wang

References


See Also

http://bioinformatics.mdanderson.org/main/DeMixT

Examples

# Example 1: estimate proportions for simulated two-component data
# with spike-in normal reference
data(test.data.2comp)
res.GS = DeMixT_GS(data.Y = test.data.2comp$data.Y,
data.N1 = test.data.2comp$data.N1,
if.filter = TRUE, ngene.Profile.selected = 150,
mean.diff.in.CM = 0.25, ngene.selected.for.pi = 150,
tol = 10^(-5))

# Example 2: estimate proportions for simulated two-component data
# without spike-in normal reference
DeMixT_S1

Estimates the proportions of mixed samples for each mixing component

Description

This function is designed to estimate the deconvolved expressions of individual mixed tumor samples for unknown component for each gene.

Usage

DeMixT_S1(
  data.Y,
  data.N1,
  data.N2 = NULL,
  niter = 10,
  nbin = 50,
  if.filter = TRUE,
  filter.sd = 0.5,
  ngene.selected.for.pi = NA,
  nspikein = NULL,
  mean.diff.in.CM = 0.25,
  tol = 10^(-5),
  pi01 = NULL,
  pi02 = NULL,
  nthread = parallel::detectCores() - 1
)

Arguments

data.Y A SummarizedExperiment object of expression data from mixed tumor samples. It is a $G$ by $M_y$ matrix where $G$ is the number of genes and $M_y$ is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

data.N1 A SummarizedExperiment object of expression data from reference component 1 (e.g., normal). It is a $G$ by $M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1.

data.N2 A SummarizedExperiment object of expression data from additional reference samples. It is a $G$ by $M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2. Component 2 is needed only for running a three-component model.
### DeMixT_S1

- **niter**: The maximum number of iterations used in the algorithm of iterated conditional modes. A larger value better guarantees the convergence in estimation but increases the running time. The default is 10.
- **nbin**: The number of bins used in numerical integration for computing complete likelihood. A larger value increases accuracy in estimation but increases the running time, especially in a three-component deconvolution problem. The default is 50.
- **if.filter**: The logical flag indicating whether a predetermined filter rule is used to select genes for proportion estimation. The default is TRUE.
- **filter.sd**: The cut-off for the standard deviation of lognormal distribution. Genes whose log transferred standard deviation smaller than the cut-off will be selected into the model. The default is 0.5.
- **ngene.selected.for.pi**: The percentage or the number of genes used for proportion estimation. The difference between the expression levels from mixed tumor samples and the known component(s) are evaluated, and the most differential expressed genes are selected, which is called S1. It is enabled when if.filter = TRUE. The default is \( \min(1500, 0.3 \times M_y) \), where \( M_y \) is the number of mixed sample. Users can also try using more genes, ranging from \( 0.3 \times M_y \) to \( 0.5 \times M_y \), and evaluate the outcome.
- **nspikein**: The number of spikes in normal reference used for proportion estimation. The default value is \( \min(200, 0.3 \times M_y) \), where \( M_y \) the number of mixed samples. If it is set to 0, proportion estimation is performed without any spike in normal reference.
- **mean.diff.in.CM**: Threshold of expression difference for selecting genes in the component merging strategy. We merge three-component to two-component by selecting genes with similar expressions for the two known components. Genes with the mean differences less than the threshold will be selected for component merging. It is used in the three-component setting, and is enabled when if.filter = TRUE. The default is 0.25.
- **tol**: The convergence criterion. The default is \( 10^{-5} \).
- **pi01**: Initialized proportion for first known component. The default is Null and pi01 will be generated randomly from uniform distribution.
- **pi02**: Initialized proportion for second known component. pi02 is needed only for running a three-component model. The default is Null and pi02 will be generated randomly from uniform distribution.
- **nthread**: The number of threads used for deconvolution when OpenMP is available in the system. The default is the number of whole threads minus one. In our no-OpenMP version, it is set to 1.

### Value

- **pi**: A matrix of estimated proportion. First row and second row corresponds to the proportion estimate for the known components and unknown component respectively for two or three component settings, and each column corresponds to one sample.
- **pi.iter**: Estimated proportions in each iteration. It is a \( niter \times Ny \times p \) array, where \( p \) is the number of components. This is enabled only when output.more.info = TRUE.
- **gene.name**: The names of genes used in estimating the proportions. If no gene names are provided in the original data set, the genes will be automatically indexed.
DeMixT_S2

Deconvolves expressions of each individual sample for unknown component

Description

This function is designed to estimate the deconvolved expressions of individual mixed tumor samples for unknown component for each gene.
Usage

DeMixT_S2(
  data.Y,
  data.N1,
  data.N2 = NULL,
  givenpi,
  nbin = 50,
  nthread = parallel::detectCores() - 1
)

Arguments

data.Y
A SummarizedExperiment object of expression data from mixed tumor samples. It is a \( G \) by \( M_y \) matrix where \( G \) is the number of genes and \( M_y \) is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

data.N1
A SummarizedExperiment object of expression data from reference component 1 (e.g., normal). It is a \( G \) by \( M_1 \) matrix where \( G \) is the number of genes and \( M_1 \) is the number of samples for component 1.

data.N2
A SummarizedExperiment object of expression data from additional reference samples. It is a \( G \) by \( M_2 \) matrix where \( G \) is the number of genes and \( M_2 \) is the number of samples for component 2. Component 2 is needed only for running a three-component model.
givenpi
A vector of proportions for all mixed tumor samples. In two-component analysis, it gives the proportions of the unknown reference component, and in three-component analysis, it gives the proportions for the two known components.
nbin
Number of bins used in numerical integration for computing complete likelihood. A larger value increases accuracy in estimation but increases the running time, especially in a three-component deconvolution problem. The default is 50.
nthread
The number of threads used for deconvolution when OpenMP is available in the system. The default is the number of whole threads minus one. In our no-OpenMP version, it is set to 1.

Value

decovExprT
A matrix of deconvolved expression profiles corresponding to T-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample.

decovExprN1
A matrix of deconvolved expression profiles corresponding to N1-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample.

decovExprN2
A matrix of deconvolved expression profiles corresponding to N2-component in mixed samples for a given subset of genes in a three-component setting. Each row corresponds to one gene and each column corresponds to one sample.

decovMu
A matrix of estimated \( \mu \) of log2-normal distribution for both known (\( \mu N_1, \mu N_2 \)) and unknown component (\( \mu T \)). Each row corresponds to one gene.

decovSigma
Estimated \( \sigma \) of log2-normal distribution for both known (\( \sigma N_1, \sigma N_2 \)) and unknown component (\( \sigma T \)). Each row corresponds to one gene.
Optimum_KernelC

Author(s)
Zeya Wang, Wenyi Wang

References

See Also
http://bioinformatics.mdanderson.org/main/DeMixT

Examples
# Example 1: two-component deconvolution given proportions
data(test.data.2comp)
givenpi <- c(t(as.matrix(test.data.2comp$pi[-2,])))
res.S2 <- DeMixT_S2(data.Y = test.data.2comp$data.Y,
data.N1 = test.data.2comp$data.N1,
data.N2 = NULL,
givenpi = givenpi,
nbin = 50)

# Example 2: three-component deconvolution given proportions
# data(test.data.3comp)
# givenpi = c(t(test.data.3comp$pi[-3,]))
# res <- DeMixT_S2(data.Y = test.data.3comp$data.Y,
# data.N1 = test.data.3comp$data.N1,
# data.N2 = test.data.3comp$data.N2,
# givenpi = givenpi,
# nbin = 50)

Optimum_KernelC

Kernel function for optimizing parameters and hidden variables in DeMixT

Description
This function is invoked by DeMixT_GS or DeMixT_S1 and DeMixT_S2 to finish parameter estimation by iterated conditional mode algorithm and reconstitute gene expression profile of all components.

Usage
Optimum_KernelC(
  inputdata,
  groupid,
  nspikein,
  setting.pi,
  givenpi,
  givenpiT,
  niter,
ninteg, 
tol,  
sg0 = 0.5^2,  
mu0 = 0,  
pi01 = NULL,  
pi02 = NULL,  
nthread = 1 
)

Arguments

inputdata  A matrix of expression data (e.g. gene expressions) from reference (e.g. normal) and mixed samples (e.g. mixed tumor samples). It is a \( G \times M \) matrix where \( G \) is the number of genes and \( M \) is the number of samples including reference and mixed samples. Samples with the same tissue type should be placed together in columns (e.g. cbind(normal samples, mixed tumor samples).

groupid  A vector of indicators to denote if the corresponding samples are reference samples or mixed tumor samples. DeMixT is able to deconvolve mixed tumor samples with at most three components. We use 1 and 2 to denote the samples referencing the first and the second known component in mixed tumor samples. We use 3 to indicate mixed tumor samples prepared to be deconvolved. For example, in two-component deconvolution, we have c(1,1,...,3,3) and in three-component deconvolution, we have c(1,1,...,2,2,...,3,3).

nspikein  The number of spikes in normal reference used for proportion estimation. The default value is \( \min(200, 0.3 \times M_y) \), where \( M_y \) the number of mixed tumor samples. If it is set to 0, proportion estimation is performed without any spike in normal reference.

setting.pi  If it is set to 0, then deconvolution is performed without any given proportions; if set to 1, deconvolution with given proportions for the first and the second known component is run; if set to 2, deconvolution is run with given tumor proportions. This option helps to perform deconvolution in different settings. In estimation of component-specific proportions, we use a subset of genes; so when it is required to deconvolve another subset of genes, we just easily plug back our estimated proportions by setting this option to 1. In our two-step estimation strategy in a three-component setting, this option is set to 2 to implement the second step.

givenpi  \( ST \)-Vector of proportions. Given the number of mixed tumor samples is \( M_y(M_y < M) \), \( ST \) is set to \( 2 \times M_y \) in a three-component setting and \( M_y \) in a two-component setting. When setting.pi is 1, it is fixed with the given proportions for the first and the second known component of mixed tumor samples, or for one unknown component when there is just one type of reference tissues. It has the form of Vector \( PiN1-1, PiN1-2, ..., PiN1-M_y, PiN2-1, PiN2-2, ..., PiN2-M_y \).

givenpiT  \( ST \)-Vector of proportions. When setting.pi is set to 2, givenpiT is fixed with given proportions for unknown component of mixed tumor samples. This option is used when we adopt a two-step estimation strategy in deconvolution. It has the form of Vector \( PiT-1, PiT-2, ..., PiT-M_y \). If option is not 2, this vector can be given with any element.

niter  The number of iterations used in the algorithm of iterated conditional modes. A larger value can better guarantee the convergence in estimation but increase the computation time.
Optimum_KernelC

ninteg 
The number of bins used in numerical integration for computing complete likelihood. A larger value can increase accuracy in estimation but also increase the running time. Especially in three-component deconvolution, the increase of number of bins can greatly lengthen the running time.

tol 
The convergence criterion. The default is 10^(-5).

sg0 
Initial value for σ^2. The default is 0.5^2.

mu0 
Initial value for μ. The default is 0.

pi01 
Initialized proportion for first known component. The default is Null and pi01 will be generated randomly from uniform distribution.

pi02 
Initialized proportion for second known component. pi02 is needed only for running a three-component model. The default is Null and pi02 will be generated randomly from uniform distribution.

nthread 
The number of threads used for deconvolution when OpenMP is available in the system. The default is the number of whole threads minus one. In our no-OpenMP version, it is set to 1.

Value

pi 
Matrix of estimated proportions for each known component. The first row corresponds to the proportion estimate of each sample for the first known component (groupid = 1) and the second row corresponds to that for the second known component (groupid = 2).

decovExpr 
A matrix of deconvolved expression profiles corresponding to unknown (e.g. tumor) component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample.

decovMu 
Estimated Mu of log2-normal distribution for tumor component.

decovSigma 
Estimated Sigma of log2-normal distribution for tumor component.

pi1 
An M x I matrix of estimated proportions for each iteration, where I is the number of iteration, for the first known component.

pi2 
An M x I matrix of estimated proportions for each iteration, where I is the number of iteration, for the second known component.

Author(s)

Zeya Wang, Wenyi Wang

References


See Also

http://bioinformatics.mdanderson.org/main/DeMiXT
**Examples**

```r
# Example 1: simulated two-component data
data(test.data.2comp)
data.N1 <- SummarizedExperiment::assays(test.data.2comp$data.N1)[[1]]
data.Y <- SummarizedExperiment::assays(test.data.2comp$data.Y)[[1]]
inputdata <- cbind(data.N1, data.Y)
groupid <- c(rep(1, ncol(data.N1)), rep(3, ncol(data.Y)))
# Simulator <- c(rep(1, ncol(data.N1)), rep(3, ncol(data.y))
# nspikein <- 0
# Optimum_KernelC(inputdata, groupid,
#   nspikein = nspikein, setting.pi = 0,
#   givenpi = rep(0, 2 * ncol(data.y)),
#   niter = 10, ninteg = 30, tol = 10^(-4))
```

---

**simulate_2comp**  
*Function to simulate two-component test data*

**Description**

Function to simulate two-component test data for DeMixT.

**Usage**

```r
simulate_2comp(G = 500, My = 100, M1 = 100, output.more.info = FALSE)
```

**Arguments**

- **G**  
  Number of genes for simulation.
- **My**  
  Number of mixture tumor samples for simulation.
- **M1**  
  Number of normal reference for simulation.
- **output.more.info**  
  The logical flag indicating whether to show True.data.T and True.data.N1 in the output. The default is FALSE.

**Value**

- **pi**  
  A matrix of estimated proportion. First row and second row correspond to the proportion estimate for the known components and unknown component respectively for two or three component settings. Each column corresponds to one sample.
- **Mu**  
  Simulated Mu of log2-normal distribution for both known (MuN1) and unknown component (MuT).
- **Sigma**  
  Simulated Sigma of log2-normal distribution for both known (SigmaN1) and unknown component (SigmaT).
- **data.Y**  
  A SummarizedExperiment object of expression data from mixed tumor samples. It is a G by My matrix where G is the number of genes and My is the number of mixed samples. Samples with the same tissue type should be placed together in columns.
- **data.N1**  
  A SummarizedExperiment object of expression data from reference component 1 (e.g., normal). It is a G by M1 matrix where G is the number of genes and M1 is the number of samples for component 1.
A SummarizedExperiment object of simulated tumor expression data. It is a $G$ by $M_y$ matrix, where $G$ is the number of genes and $M_y$ is the number of mixed samples. This is enabled only when `output.more.info = TRUE`.

A SummarizedExperiment object of simulated true expression data for reference component 1 (e.g., normal). It is a $G$ by $M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1. This is enabled only when `output.more.info = TRUE`.

```
examples

test.data = simulate_2comp(G = 500, My = 100, M1 = 100)
test.data$pi
test.data$mu

test.data$sigma
```

---

**simulate_3comp**

*Function to simulate three-component mixed cell line test data*

**Description**

Function to simulate three-component mixed cell line test data used in DeMixT function.

**Usage**

```
simulate_3comp(
  G1 = 675,
  G2 = 25,
  My = 20,
  M1 = 100,
  M2 = 100,
  output.more.info = FALSE
)
```

**Arguments**

- `G1`: Number of genes, where $\mu_{N1}$ is close to $\mu_{N2}$.
- `G2`: Number of genes, where $\mu_{N1}$ is not close to $\mu_{N2}$.
- `My`: Number of mixture tumor samples for simulation.
- `M1`: Number of first known reference for simulation.
- `M2`: Number of second known reference for simulation.
- `output.more.info`: The logical flag indicating whether to show `True.data.T`, `True.data.N1` and `True.data.N2` in the output. The default is `FALSE`.

**Value**

- `pi`: A matrix of estimated proportion. First row and second row corresponds to the proportion estimate for the known components and unknown component respectively for two or three component settings. Each column corresponds to one sample.
$\Mu$ Simulated $\Mu$ of log2-normal distribution for both known ($\MuN1, \MuN2$) and unknown component ($\MuT$).

$\Sigma$ Simulated $\Sigma$ of log2-normal distribution for both known ($\SigmaN1, \SigmaN2$) and unknown component ($\SigmaT$).

data.Y A SummarizedExperiment object of simulated expression data from mixed tumor samples. It is a $G \times M_y$ matrix where $G$ is the number of genes and $M_y$ is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

data.N1 A SummarizedExperiment object of simulated expression data from reference component 1 (e.g., normal). It is a $G \times M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1.

data.N2 A SummarizedExperiment object of expression data from additional reference samples. It is a $G \times M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2.

True.data.T A SummarizedExperiment object of simulated tumor expression data. It is a $G \times M_y$ matrix, where $G$ is the number of genes and $M_y$ is the number of mixed samples. This is enabled only when `output.more.info = TRUE`.

True.data.N1 A SummarizedExperiment object of simulated true expression data for reference component 1 (e.g., stroma). It is a $G \times M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1. This is enabled only when `output.more.info = TRUE`.

True.data.N2 A SummarizedExperiment object of simulated true expression data for reference component 2 (e.g., immune). It is a $G \times M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2. This is enabled only when `output.more.info = TRUE`.

**Examples**

```r
test.data = simulate_3comp(G1 = 675, G2 = 25, M_y = 20, M_1 = 100, M_2 = 100)
test.data$pi
test.data$\Mu
test.data$\Sigma
```

**Description**

A list of simulated two-component test data used in DeMixT function. Expression data with 500 genes and 100 samples are simulated.

**Usage**

`test.data.2comp`

**Format**

An object of class `list` of length 5.
test.data.3comp

Description

A list of simulated three-component mixed cell line test data used in DeMixT function. Expression data with 700 genes and 20 samples are simulated, where 675 genes’ $\text{Mu}_1$ is close to $\text{Mu}_2$.

Usage

test.data.3comp

Format

An object of class list of length 6.

Value

A list with 6 elements (3 more elements when output.more.info = TRUE), which are

$\pi$

A matrix of estimated proportion. First row and second row corresponds to the proportion estimate for the known components and unknown component respectively for two or three component settings. Each column corresponds to one sample.

$\text{Mu}$

Simulated $\text{Mu}$ of log2-normal distribution for both known ($\text{Mu}_1$) and unknown component ($\text{Mu}_T$).

$\text{Sigma}$

Simulated $\text{Sigma}$ of log2-normal distribution for both known ($\text{Sigma}_1$) and unknown component ($\text{Sigma}_T$).

$\text{data.Y}$

A SummarizedExperiment object of expression data from mixed tumor samples. It is a $G \times M_y$ matrix where $G$ is the number of genes and $M_y$ is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

$\text{data.N1}$

A SummarizedExperiment object of expression data from reference component 1 (e.g., normal). It is a $G \times M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1.

True.data.T

A SummarizedExperiment object of simulated tumor expression data. It is a $G \times M_y$ matrix, where $G$ is the number of genes and $M_y$ is the number of mixed samples. This is shown only when output.more.info = TRUE.

True.data.N1

A SummarizedExperiment object of simulated true expression data for reference component 1 (e.g., normal). It is a $G \times M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1. This is shown only when output.more.info = TRUE.
Simulated $\mu$ of log2-normal distribution for both known ($\mu_{N1}, \mu_{N2}$) and unknown component ($\mu_T$).

Simulated $\sigma$ of log2-normal distribution for both known ($\sigma_{N1}, \sigma_{N2}$) and unknown component ($\sigma_T$).

A SummarizedExperiment object of simulated expression data from mixed tumor samples. It is a $G$ by $M_y$ matrix where $G$ is the number of genes and $M_y$ is the number of mixed samples. Samples with the same tissue type should be placed together in columns.

A SummarizedExperiment object of simulated expression data from reference component 1 (e.g., normal). It is a $G$ by $M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1.

A SummarizedExperiment object of expression data from additional reference samples. It is a $G$ by $M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2.

A SummarizedExperiment object of simulated tumor expression data. It is a $G$ by $M_y$ matrix, where $G$ is the number of genes and $M_y$ is the number of mixed samples. This is shown only when output.more.info = TRUE.

A SummarizedExperiment object of simulated true expression data for reference component 1 (e.g., stroma). It is a $G$ by $M_1$ matrix where $G$ is the number of genes and $M_1$ is the number of samples for component 1. This is shown only when output.more.info = TRUE.

A SummarizedExperiment object of simulated true expression data for reference component 2 (e.g., immune). It is a $G$ by $M_2$ matrix where $G$ is the number of genes and $M_2$ is the number of samples for component 2. This is shown only when output.more.info = TRUE.
Index

*Topic DeMixT_GS
  DeMixT_GS, 5

*Topic DeMixT_S1
  DeMixT_S1, 8

*Topic DeMixT_S2
  DeMixT_S2, 10

*Topic DeMixT
  DeMixT, 2

*Topic Optimum_KernelC
  Optimum_KernelC, 12

*Topic datasets
  test.data.2comp, 17
  test.data.3comp, 18

*Topic simulate_3comp
  simulate_3comp, 16

DeMixT, 2
DeMixT_GS, 5
DeMixT_S1, 8
DeMixT_S2, 10

Optimum_KernelC, 12

simulate_2comp, 15
simulate_3comp, 16

test.data.2comp, 17
test.data.3comp, 18