Package 'DeMixT'

March 23, 2018

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

Version 0.1

Date 2018-03-19

Authors Zeya Wang <zw17.rice@gmail.com>, Wenyi Wang <wwang7@@mdanderson.org>

Maintainers Zeya Wang <zw17.rice@gmail.com>, Fan Gao <fgao3@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 (>= 3.2), parallel

NeedsCompilation yes

R topics documented:

`DeMixT` *Deconvolution of heterogeneous tumor samples with two or three components using expression data from RNAseq or microarray platforms*

Description

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

Usage

```
DeMixT(data.Y, data.comp1, data.comp2 = NULL, niter = 10, nbin = 50,
     if.filter = TRUE, num.of.gene.selected.for.pi = 250, mean.diff.in.CM = 0.25,
       tol = 10^(-5), output.more.info = FALSE, nthread = detectCores() - 1)
```

Arguments

data.Y	A matrix of expression data from mixed tumor samples. It is a G by Sy matrix where G is the number of genes and Sy is the number of mixed samples. Samples with the same tissue type should be placed together in columns.
data.comp1	A matrix of expression data from reference component 1 (e.g., normal). It is a G by $S1$ matrix where G is the number of genes and $S1$ is the number of samples for component 1.
data.comp2	A matrix of expression data from additional reference samples. It is a G by $S2$ matrix where G is the number of genes and $S2$ 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 (ICM, Ref[1]). 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.
ngene.selected.for.pi	The percentage or the number of genes used for proportion estimation. The difference between the expression levels from mixted tumor samples and the known component(s) are evaluated, and the most differentially expressed genes are selected. It is enabled when if.filter = TRUE. The default is 250.
mean.diff.in.CM	Threshold of expression difference in 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).
output.iter	The logical flag indicating whether to show the estimated proportions in each iteration in the output.
nthread	The number of threads used for deconvolution when OpenMP is availble 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. $\pi 1$ corresponds to the proportion estimate for the first known component. $\pi 2$ corresponds to the second known component.

pi.iter	Estimated proportions in each iteration. It is a $numberofiterationXSyX1$ array in two-component setting, and a $number_of_iterationXSyX2$ array in three-component setting. This is enabled only when output.more.info = TRUE.
decovExprT	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	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	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	Estimated μ of log2-normal distribution for both known ($MuN1, MuN2$) and unknown component (MuT).
decovSigma	Estimated σ of log2-normal distribution for both known ($SigmaN1, SigmaN2$) and unknown component ($SigmaT$).
gene.name	The names of genes used in estimating the proportions. If no gene names are rpovided in the original data set, the genes will be automatically indexed. This is enabled only when output.more.info = TRUE.

Author(s)

Zeya Wang, Wenyi Wang

References

J. Besag. "On the statistical analysis of dirty pictures". In: Journal of the Royal Statistical Society. Series B (Methodological) (1986), pp. 259–302.

See Also

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

Examples

```
# Example 1: simulated two-component data
 data(test.data1)
 res <- DeMixT(data.Y = test.data1$y, data.comp1 = test.data1$comp1, if.filter = FALSE, output.more.info = TRUE
 res$pi
 head(res$decovExprT, 3)
 head(res$decovExprN1, 3)
 head(res$decovMu, 3)
 head(res$decovSigma, 3)
 res$pi.iter
 res$gene.name

# Example 2: simulated three-component data
# It takes about 15 minutes to finish running
# data(test.data2)
# res <- DeMixT(data.Y = test.data2$y, data.comp1 = test.data2$comp1, data.comp2 = test.data2$comp2, if.filter

# Example 3: three-component mixed cell line data applying component merging strategy
# It takes about 1.5 hours to finish running
```

```
# data(test.data3)
# res <- DeMixT(data.Y = test.data3$y, data.comp1 = test.data3$comp1, data.comp2 = test.data3$comp2, if.filter
```

DeMixT.S1 *Estimates the proportions of mixed samples for each mixing component*

Description

This function is designed to estimate the proportions of all mixed samples for each mixing component with or without component merging.

Usage

```
DeMixT(data.Y, data.comp1, data.comp2 = NULL, niter = 10, nbin = 50,
       if.filter = TRUE, ngene.selected.for.pi = 250, mean.diff.in.CM = 0.25,
       tol = 10^(-5), nthread = detectCores() - 1)
```

Arguments

data.Y	A matrix of expression data from mixed tumor samples. It is a G by Sy matrix where G is the number of genes and Sy is the number of mixed samples. Samples with the same tissue type should be placed together in columns.
data.comp1	A matrix of expression data from reference component 1 (e.g., normal). It is a G by $S1$ matrix where G is the number of genes and $S1$ is the number of samples for component 1.
data.comp2	A matrix of expression data from additional reference samples. It is a G by $S2$ matrix where G is the number of genes and $S2$ 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 (ICM, Ref[1]). 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.
ngene.selected.for.pi	The percentage or the number of genes used for proportion estimation. The difference between the expression levels from mixted tumor samples and the known component(s) are evaluated, and the most differentially expressed genes are selected. It is enabled when if.filter = TRUE. The default is 250.
mean.diff.in.CM	Threshold of expression difference in 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).
nthread	The number of threads used for deconvolution when OpenMP is availble 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. $\pi 1$ corresponds to the proportion estimate for the first known component. $\pi 2$ corresponds to the second known component.
pi.iter	Estimated proportions in each iteration. It is a $numberofiterationXSyX1$ array in two-component setting, and a $number_of_iterationXSyX2$ array in three-component setting. This is enabled only when output.iter = TRUE.
gene.name	The names of genes used in estimating the proportions. If no gene names are rpovided in the original data set, the genes will be automatically indexed.

Author(s)

Zeya Wang, Wenyi Wang

References

J. Besag. "On the statistical analysis of dirty pictures". In: Journal of the Royal Statistical Society. Series B (Methodological) (1986), pp. 259–302.

See Also

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

Examples

```
# Example 1: estimate proportions for simulated two-component data
 data(test.data1)
 res <- DeMixT.S1(data.Y = test.data1$y, data.comp1 = test.data1$comp1, if.filter = FALSE)

# Example 2: estimate proportions for simulated three-component data
# This example takes 10 minutes to finish running
# data(test.data2)
# res <- DeMixT.S1(data.Y = test.data2$y, data.comp1 = test.data2$comp1, data.comp2 = test.data2$comp2, if.filt

# Example 3: estimate proportions for simulated three-component  mixed cell line data
# This example takes 1 hour to finish running
# data(test.data2$comp3)
# res <- DeMixT.S1(data.Y = test.data3$y, data.comp1 = test.data3$comp1, data.comp2 = test.data3$comp2)
```

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.comp1, data.comp2 = NULL, givenpi, nbin = 50, nthread = detectCores() - 1)
```

Arguments

data.Y	A matrix of expression data from mixed tumor samples. It is a G by Sy matrix where G is the number of genes and Sy is the number of mixed samples. Samples with the same tissue type should be placed together in columns.
data.comp1	A matrix of expression data from reference component 1 (e.g., normal). It is a G by $S1$ matrix where G is the number of genes and $S1$ is the number of samples for component 1.
data.comp2	A matrix of expression data from additional reference samples. It is a G by $S2$ matrix where G is the number of genes and $S2$ 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 known reference component, and in three-component analysis, it gives the proportions for the two known components.
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.
nthread	The number of threads used for deconvolution when OpenMP is availble 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	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	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	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	Estimated μ of log2-normal distribution for both known ($MuN1, MuN2$) and unknown component (MuT).
decovSigma	Estimated σ of log2-normal distribution for both known ($SigmaN1, SigmaN2$) and unknown component ($SigmaT$).

Author(s)

Zeya Wang, Wenyi Wang

References

J. Besag. "On the statistical analysis of dirty pictures". In: Journal of the Royal Statistical Society. Series B (Methodological) (1986), pp. 259–302.

See Also

http://bioinformatics.mdanderson.org/main/DeMix:Overview

Examples

```
# Example 1: two-component deconvolution given proportions
  data(test.data1)
  givenpi <- c(t(as.matrix(test.data1$truth[-2,])))
  res <- DeMixT.S2(data.Y = test.data1$y, data.comp1 = test.data1$comp1, givenpi = givenpi)

# Example 2: three-component deconvolution given proportions
# This example takes 10 minutes to finish running
#  data(test.data2)
#  givenpi <- c(t(test.data2$truth[-3,]))
#  res <- DeMixT.S2(data.Y = test.data2$y, data.comp1 = test.data2$comp1, data.comp2 = test.data2$comp2, givenp
```

---

test.data1 *simulated two-component test data*

Description

simulated two-component test data used in function DeMixT

Usage

```
test.data1, test.data1$y, test.data1$comp1, test.data1$truth
```

Format

A list containing two matrices

`y` a matrix of expression data from mixed tumor samples

`comp1` a matrix of expression data from reference component 1

`truth` a matrix of true proportions, i.e., π_1 and $1 - \pi_1$

Examples

```
data(test.data1)
test.data1
```

test.data2 *simulated three-component test data*

Description

simulated three-component test data used in function DeMixT

Usage

```
test.data2, test.data2$y, test.data2$comp1, test.data2$comp2, test.data2$truth
```

Format

A list containing three matrices

`y` a matrix of expression data from mixed tumor samples

`comp1` a matrix of expression data from reference component 1

`comp2` a matrix of expression data from reference component 2

`truth` a matrix of true proportions, i.e., π_1, π_2, and $1 - \pi_1 - \pi_2$

Examples

```
data(test.data2)
test.data2
```

test.data3 *three-component mixed cell line test data*

Description

three-component mixed cell line test data used in function DeMixT

Usage

```
test.data3, test.data3$y, test.data3$comp1, test.data3$comp2
```

Format

A list containing three matrices

`y` a matrix of expression data from mixed tumor samples

`comp1` a matrix of expression data from reference component 1

`comp2` a matrix of expression data from reference component 2

Examples

```
data(test.data3)
test.data3
```

Index