Package ‘DeMixT’

October 10, 2017

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

Version 0.1

Date 2016-12-16

Author Zeya Wang, Wenyi Wang

Maintainer Zeya Wang <zw17@rice.edu>,Shaolong Cao <SCao@mdanderson.org>

Description

We develop a three-component deconvolution model, DeMixT, for expression data from a mixture of cancerous tissues, infiltrating immune cells and tumor microenvironment. DeMixT is a frequentist-based method and fast in yielding accurate estimates of cell proportions and compartment-specific expression profiles for two-component and three-component deconvolution problem. Our method promises to provide deeper insight into cancer biomarkers and assist in the development of novel prognostic markers and therapeutic strategies.

LazyData TRUE

Depends R (>= 3.2), parallel

NeedsCompilation yes

R topics documented:

DeMixT 1
DeMixT.S1 4
DeMixT.S2 7
Optimum.KernelC 9

Index **11**

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

Description

This is the main function to run the deconvolution analysis in a setting of two or three components.

Usage

```
DeMixT(inputdata, groupid, niter = 10, ninteg1 = 50,ninteg2 = 50, filter.option = 1,
filter.criteria1 = c(0.5,0.5), filter.criteria2 = c(250,250), filter.criteria3 = 0.25,
if.filter = FALSE, tol = 10^(-5), sg0 = 0.5^2, mu0= 0.0, nthread = detectCore() - 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 $G \times S$ matrix where G is the number of genes and S 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).

`niter` The maximum number of iterations used in the algorithm of iterated conditional modes. A larger value can better guarantee the convergence in estimation. The default number is 10.

`ninteg1` The number of bins used in numerical integration for computing complete likelihood when proportions are estimated. 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 lengthen the running time. The default value is 50.

`ninteg2` The number of bins used in numerical integration for computing complete likelihood when individual expressions are deconvolved. The default value is 50.

`filter.option` The option is used to process or filter zero count from input expression matrix. If it is set to 1, all those genes containing zero count in sample will be removed. If it is set to 2, input data matrix will be added 1 to vanish zero count. This option is used for RNA-seq data. Default is 1.

`filter.criteria1`
Filtering threshold used to select genes with small sample standard deviations in reference samples. This threshold is set on the level of log2-transformed data. A vector of length 1 is given for two-component deconvolution and length 2 for three-component deconvolution. A larger value will relax the gene filtering for proportion estimation but might introduce bias. A smaller value will take fewer genes for proportion estimation. User should take care of the choice of this value because too stringent criteria can cause no gene to be input for estimation. It is only enabled when if.filter = TRUE. Default is 0.5 for each component.

`filter.criteria2`
The percentage or the number of genes that will be used for proportion estimation after gene selection. It is a vector of length 2 for three-component setting and length 1 for two-component setting. In the two-component setting, a rank of fold change of mixed to reference is used to select most differentially expressed genes at the average level. In the three-component setting, this is used within our component merging strategy. The first element is set to give the number of genes selected in the first step when we merge three-component into two-component,

and the second element of this vector is set to give the number of genes selected in the second step, where all three components are being estimated. It is enabled when if.filter = TRUE. Default value is 250.

`filter.criteria3`
: Threshold of mean expression difference given to select genes with close expressions between two known components. Genes with mean expression difference between two known components below this value will be included for deconvolution in the first step. This option is set in our first step of component merging strategy when we merge three-component to two-component by selecting genes with very close expressions for the two known components. More stringent value can merge from three-component to two-component more strictly but also cause fewer genes or no gene left for estimation. It is only enabled for the case of three-component. Default is 0.25.

`if.filter`
: Option to control if a pre-determined filtering rule is used to select genes for proportion estimation.

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

`sg0`
: A single initial value of σ_T^2 in log2-normal distribution for each gene. The default value is 0.5^2.

`mu0`
: A single initial value of μ_T in log2-normal distribution for each gene. It should be a postive value. If it is assigned with any non-positive value, the initial value will use an estimator of method of moments by assuming π satisfying a Dirichlet distribution. The default value is 0.0.

`nthread`
: The number of threads used for deconvolution. The default is the number of whole CPUs minus one.

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), and the third component would be $1 - \pi_1 - \pi_2$.

`pi_iteration`
: A list of estimated proportions for the first and second known component in each iteration. The first element in the list provides estimates in each iteration for the first component and the second element provides values for the second component. In each of these two elements, each row corresponds to each gene and each column corresponds to each iteratoin, and the third component would be $1 - \pi_1 - \pi_2$.

`decovExprT`
: A matrix of deconvolved expression profiles corresponding to unknown T-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample. Gene names and sample names are attached in the output.

`decovExprN1`
: A matrix of deconvolved expression profiles corresponding to known N1-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample. Gene names and sample names are attached in the ouptut.

`decovExprN2`
: A matrix of deconvolved expression profiles corresponding to known N2-component in mixed samples for a given subset of genes. Each row corresponds to one gene and each column corresponds to one sample. Gene names and sample names are attached in the output.

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

See Also

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

Examples

```
#first example for simulated two-component data
 data(simul2)
 inputdata <- as.matrix(simul2)
 groupid <- c(rep(1, 5), rep(3, 10));
 res <- DeMixT(inputdata, groupid)

## Those two examples below will take around hours for finishing the running.
## Be cautious if you want to test those examples.
#second example for simulated three-component data
 data(simul3)
 inputdata <- as.matrix(simul3)
 groupid <- c(rep(1, 5), rep(2,5), rep(3, 5));
 res <- DeMixT(inputdata, groupid)

#third example of applying component merging strategy for three-component
#in proportion estimation to a mixed cell line data example (this example
# takes a longer time to be finished.)
 data(cell_line_mix)
 inputdata <- as.matrix(cell_line_mix)
 groupid <- c(rep(1, 6), rep(2,6), rep(3, 8)); # input as 2-component
 res <- DeMixT(inputdata, groupid, if.filter = TRUE)
```

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.S1(inputdata, groupid, niter = 10, ninteg = 50, filter.option = 1,
filter.criteria1 = c(0.5,0.5), filter.criteria2 = c(250,250), filter.criteria3 = 0.25,
if.filter = FALSE,tol=10^(-5), sg0 = 0.5^2, mu0= 0.0, nthread=detectCore() - 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 $G \times S$ matrix where G is the number of genes and S 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).

`niter` The maximum number of iterations used in the algorithm of iterated conditional modes. A larger value can better guarantee the convergence in estimation. The default number is 10.

`ninteg` The number of bins used in numerical integration for computing complete likelihood when proportions are estimated. 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 lengthen the running time. The default value is 50.

`filter.option` The option is used to process or filter zero count from input expression matrix. If it is set to 1, all those genes containing zero count in sample will be removed. If it is set to 2, input data matrix will be added 1 to vanish zero count. This option is used for RNA-seq data. Default is 1.

`filter.option` The option is used to process or filter zero count from input expression matrix. If it is set to 1, all those genes containing zero count in sample will be removed. If it is set to 2, input data matrix will be added 1 to vanish zero count. This option is used for RNA-seq data. Default is 1.

`filter.criteria1`
Filtering threshold used to select genes with small sample standard deviations in reference samples. This threshold is set on the level of log2-transformed data. A vector of length 1 is given for two-component deconvolution and length 2 for three-component deconvolution. A larger value will relax the gene filtering for proportion estimation but might introduce bias. A smaller value will take fewer genes for proportion estimation. User should take care of the choice of this value because too stringent criteria can cause no gene to be input for estimation. It is only enabled when if.filter = TRUE. Default is 0.5 for each component.

`filter.criteria2`
The percentage or the number of genes that will be imported for proportion estimation after gene selection. It is a vector of length 2 for three-component setting and length 1 for two-component setting. In the two-component setting, a rank of ratio of mixed to reference is used to select most differentially expressed genes at the average level. In the three-component setting, this is used within our two-step estimation strategy. The first element is set to give the number of genes selected in the first step when we degenerate three-component to two-component, and the second element of this vector is set to give the number of genes selected in the second step. It is enabled when if.filter = TRUE. Default value is 250.

`filter.criteria3`
: Threshold of mean expression difference given to select genes with close expressions between two known components. Genes with mean expression difference between two known components below this value will be included for deconvolution in the first step. This option is set in our first step of two-step estimation strategy when we degenerate three-component to two-component by selecting genes with very close expressions for the two known components. More stringent value can make the degeneration from three-component to two-component more robust but also cause fewer genes or even no gene left for estimation. It is only enabled for the case of three-component. Default is 0.25.

`if.filter`
: Option to control if a pre-determined filtering rule is used to select genes for proportion estimation.

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

`sg0`
: A single initial value of σ_T^2 in log2-normal distribution for each gene. The default value is 0.5^2.

`mu0`
: A single initial value of μ_T in log2-normal distribution for each gene. It should be a postive value. If it is assigned with any non-positive value, the initial value will use an estimator of method of moments by assuming π satisfying a Dirichlet distribution. The default value is 0.0.

`nthread`
: The number of threads used for deconvolution. The default is the number of whole CPUs minus one.

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).

`pi_iteration`
: A list of estimated proportions for the first and second known component in each iteration. The first element in the list provides estimates in each iteration for the first component and the second element provides values for the second component. In each of these two elements, each row corresponds to each gene and each column corresponds to each iteratoin.

Author(s)

Zeya Wang, Wenyi Wang

See Also

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

Examples

```
#first example of estimating proportions for simulated three-component data
 data(simul2)
 inputdata <- as.matrix(simul2)
 groupid <- c(rep(1, 5), rep(3, 10))
 res <- DeMixT.S1(inputdata, groupid)

## Those two examples below will take around hours for finishing the running.
```

```
# Be cautious if you want to test those examples.
# second example of estimating proportions for simulated three-component data
 data(simul3)
 inputdata <- as.matrix(simul3)
 groupid <- c(rep(1, 5), rep(2,5), rep(3, 5))
 res <- DeMixT.S1(inputdata, groupid)

#third example of applying two-step estimation strategy for three-component
# to a mixed #cell line data example (this example takes a longer time to be finished.)
 data(cell_line_mix)
 inputdata <- as.matrix(cell_line_mix)
 groupid <- c(rep(1, 6), rep(2,6), rep(3, 8))  # input as 2-component
 res <- DeMixT.S1(inputdata, groupid, filter.criteria1 = c(0.5,0.5),
 filter.criteria2 = c(250,250), filter.criteria3 = 0.25, if.stage = TRUE)
```

`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(inputdata, groupid, givenpi,ninteg = 50, filter.option = 1,
nthread=detectCore() - 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 $G \times S$ matrix where G is the number of genes and S 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).

`givenpi` A vector of proportions for all admixed samples, with size of the number of admixed samples for two component and two times the number of admixed samples for three component. It is fixed with given proportions for the first and the second known component of mixed tumor samples, or just for one known component when there is just one type of reference tissues. It has the form of Vector $(\pi^1_{N_1}, \pi^2_{N_1}, ..., \pi^{S_T}_{N_1}, \pi^1_{N_2}, \pi^2_{N_2}, ..., \pi^{S_T}_{N_2})$. $S_T(S_T < S)$ is the number of mixed tumor samples.

`ninteg` The number of bins used in numerical integration for computing complete likelihood when expressions are deconvolved. The default value is 50.

`filter.option` The option is used to process or filter zero count from input expression matrix. If it is set to 1, all those genes containing zero count in sample will be removed. If it is set to 2, input data matrix will be added 1 to vanish zero count. This option is used for RNA-seq data. Default is 1.

`nthread` The number of threads used for deconvolution. The default is the number of whole CPUs minus one.

Value

`decovExprT` A matrix of deconvolved expression profiles corresponding to unknown 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 known 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 known N2-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 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

See Also

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

Examples

```
#first example given for deconvolving expressions given proportions
# under the setting of two-component deconvolution
data(simul2)
data(TPi2)
inputdata = as.matrix(simul2)
givenpi <- c(t(as.matrix(TPi2[-2,])))
groupid <-  c(rep(1, 5), rep(3, 10))
res <- DeMixT.S2(inputdata, groupid, givenpi)

#second example given for deconvolving expressions given proportions
# under the setting of three-component deconvolution
data(simul3)
data(TPi3)
inputdata = as.matrix(simul3)
givenpi <- c(t(as.matrix(TPi3[-3,])))
groupid <-  c(rep(1, 5), rep(2, 5), rep(3, 5))
res <- DeMixT.S2(inputdata, groupid, givenpi)
```

`Optimum.KernelC` *Kernel function to call the C code used for parameter estimation in DeMixT*

Description

This function is used by DeMixT.S1 and DeMixT.S2 to perform parameter estimation and expression deconvolution and invokes a set of C codes to perform the majority of the computation.

Usage

```
Optimum.KernelC(inputdata, groupid, nhavepi, givenpi, givenpiT, niter, ninteg, tol,
sg0 = 0.5^2, mu0= 0.0, nthread = detectCores() - 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 $G \times S$ matrix where G is the number of genes and S 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).

`nhavepi` 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 proportions of T-component for component merging. This option helps to perform deconvolution in different settings.

`givenpi` A vector of proportions for all admixed samples, with size of the number of admixed samples for two component and two times the number of admixed samples for three component. It is fixed with given proportions for the first and the second known component of mixed tumor samples, or just for one known component when there is just one type of reference tissues. It has the form of Vector$(\pi^1_{N_1}, \pi^2_{N_1}, ..., \pi^{S_T}_{N_1}, \pi^1_{N_2}, \pi^2_{N_2}, ..., \pi^{S_T}_{N_2})$. $S_T(S_T < S)$ is the number of mixed tumor samples.

`givenpiT` A vector of proportions for all admixed samples used for component merging, with size of the number of admixed samples. When nhavepi 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$(\pi^1_T, \pi^2_T, \cdots, \pi^{S_T}_T)$. If option is not 2, this vector can be given with any element.

`niter` The maximum number of iterations used in the algorithm of iterated conditional modes. A larger value can better guarantee the convergence in estimation.

`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 lengthen the running time.

`tol` The convergence criterion.

`sg0` A single initial value of σ_T^2 in log2-normal distribution for each gene. The default value is 0.5^2.

`mu0` A single initial value of μ_T in log2-normal distribution for each gene. It should be a postive value. If it is assigned with any non-positive value, the initial value will use an estimator of method of moments by assuming π satisfying a Dirichlet distribution. The default value is 0.0.

`nthread` The number of threads used for deconvolution. The default is the number of whole CPUs minus one.

Value

`obj_val` Final negative log-likelihood value at convergence.

`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).

`decovExprT` A matrix of deconvolved expression profiles corresponding to unknown 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 known 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 known N2-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 μ of log2-normal distribution for tumor component.

`decovSigma` Estimated σ of log2-normal distribution for tumor component

`pi1` An $S_T \times I$ Matrix of estimated proportions for each iteration $i \in \{1, \cdots, I\}$ for the first known component. I is the number of iterations. S_T is the number of admixed samples.

`pi2` An $S_T \times I$ Matrix of estimated proportions for each iteration $i \in \{1, \cdots, I\}$ for the second known component. I is the number of iterations. S_T is the number of admixed samples.

Author(s)

Zeya Wang

See Also

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

Index

∗Topic **DeMixT.Kernel**
Optimum.KernelC, 9
∗Topic **DeMixT.S1**
DeMixT.S1, 4
∗Topic **DeMixT.S2**
DeMixT.S2, 7
∗Topic **DeMixT**
DeMixT, 1
DeMixT.S1, 4
DeMixT.S2, 7
Optimum.KernelC, 9

DeMixT, 1
DeMixT.S1, 4
DeMixT.S2, 7

Optimum.KernelC, 9