# Package 'SIBERG'

January 20, 2025

**Type** Package

**LazyData** true

**Date** 2022-05-02

**Title** Systematic Identification of Bimodally Expressed Genes Using RNAseq Data

**Version** 2.0.3

**Author** Pan Tong, Kevin R. Coombes

**Maintainer** Kevin R. Coombes <krc@silicovore.com>

**Description** Provides models to identify bimodally expressed genes from RNAseq data based on the Bimodality Index. SIBERG models the RNAseq data in the finite mixture modeling framework and incorporates mechanisms for dealing with RNAseq normalization. Three types of mixture models are implemented, namely, the mixture of log normal, negative binomial, or generalized Poisson distribution. See Tong et al. (2013) <doi:10.1093/bioinformatics/bts713>.

**License** Apache License (== 2.0)

**Imports** mclust

**Suggests** edgeR, doParallel

**URL** http://oompa.r-forge.r-project.org/

**NeedsCompilation** no

**Depends** R (>= 2.10)

**Repository** CRAN

**Date/Publication** 2022-05-03 06:50:05 UTC

## Contents

---

`fitGP` *Fit Generalized Poisson Mixture Model*

---

## Description

The function fits a two-component Generalized Poisson mixture model.

## Usage

```
fitGP(y, d=NULL, inits=NULL, model='V', zeroPercentThr=0.2)
```

## Arguments

| | |
|---|---|
| `y` | A vector representing the RNAseq raw count. |
| `d` | A vector of the same length as y representing the normalization constant to be applied to the data. |
| `inits` | Initial value to fit the mixture model. A vector with elements mu1, mu2, phi1, phi2 and pi1. |
| `model` | Character specifying E or V model. E model fits the mixture model with equal dispersion phi while V model doesn't put any constraint. |
| `zeroPercentThr` | A scalar specifying the minimum percent of zero counts needed when fitting a zero-inflated Generalized Poisson model. This parameter is used to deal with zero-inflation in RNAseq count data. When the percent of zero exceeds this threshold, rather than fitting a 2-component Generalized Poisson mixture, a mixture of point mass at 0 and Generalized Poisson is fitted. |

## Details

This function directly maximize the log likelihood function through optimization. With this function, three models can be fitted: (1) Generalized Poisson mixture with equal dispersion (E model); (2) Generalized Poisson mixture with unequal dispersion (V model); (3) 0-inflated Generalized Poisson model. The 0-inflated Generalized Poisson has the following density function:

$P(Y = y) = \pi D(y) + (1 - \pi)GP(\mu, \phi)$ where D is the point mass at 0 while $GP(\mu, \phi)$ is the density of Generalized Poisson distribution with mean $\mu$ and dispersion $\phi$. The variance is $\phi\mu$.

The rule to fit 0-inflated model is that the observed percentage of count exceeds the user specified threshold. This rule overrides the model argument when observed percentae of zero count exceeds the threshold.

## Value

A vector consisting parameter estimates of mu1, mu2, phi1, phi2, pi1, logLik and BIC. For 0-inflated model, mu1=phi1=0.

## Author(s)

Pan Tong (nickytong@gmail.com), Kevin R Coombes (krc@silicovore.com)

## References

Tong, P., Chen, Y., Su, X. and Coombes, K. R. (2012). Systematic Identification of Bimodally Expressed Genes Using RNAseq Data. Bioinformatics, 2013 Mar 1;29(5):605-13.

## See Also

SIBER fitLN fitNB fitNL

## Examples

```
# artificial RNAseq data from negative binomial distribution
set.seed(1000)
dat <- rnbinom(100, mu=1000, size=1/0.2)
fitGP(y=dat)
```

---

`fitLN` *Fit Log Normal Mixture Model*

---

## Description

The function fits a two-component log normal mixture model.

## Usage

```
fitLN(y, base=10, eps=10, d=NULL, model='E', zeroPercentThr=0.2, logLikToLN=TRUE)
```

## Arguments

| | |
|---|---|
| `y` | A vector representing the RNAseq raw count. |
| `base` | The logarithm base defining the parameter estimates in the logarithm scale. This is also the base of log transformation applied to the data. |
| `eps` | A scalar to be added to the count data to avoid taking logarithm of zero. |
| `d` | A vector of the same length as y representing the normalization constant to be applied to the data. For the LN model, the original data would be devided by this vector. |
| `model` | Character specifying E or V model. E model fits the mixture model with equal variance while V model doesn't put any constraint. |
| `zeroPercentThr` | A scalar specifying the minimum percent of zero counts needed when fitting a zero-inflated log normal model. This parameter is used to deal with zero-inflation in RNAseq count data. When the percent of zero exceeds this threshold, 1-comp mixture LN model is used to estimate mu and sigma from nonzero count. |
| `logLikToLN` | logical indicating if the log likelihod is defined on the transformed value or the orginal value from log noral distribution. |

## Details

The parameter estimates from log normal mixture is obtained by taking logarithm and fit normal mixture. We use mclust package to obtain parameter estimates of normal mixture model. In particular, $log_{base}(\frac{y+eps}{d})$ is used to fit to normal mixture model.

With this function, three models can be fitted: (1) log normal mixture with equal variance (E model); (2) Generalized Poisson mixture with unequal variance (V model); (3) 0-inflated log normal model. The 0-inflated log normal has the following density function:

$P(Y = y) = \pi D(y) + (1 - \pi)LN(\mu, \sigma)$ where D is the point mass at 0 while $LN(\mu, \sigma)$ is the density of log normal distribution with mean $\mu$ and variance $\sigma^2$.

The rule to fit 0-inflated model is that the observed percentage of count exceeds the user specified threshold. This rule overrides the model argument (E or V) when observed percentae of zero count exceeds the threshold.

## Value

A vector consisting parameter estimates of mu1, mu2, sigma1, sigma2, pi1, logLik and BIC. For 0-inflated model, mu1=sigma1=0.

## Author(s)

Pan Tong (nickytong@gmail.com), Kevin R Coombes (krc@silicovore.com)

## References

Wang, J.,Wen, S., Symmans,W., Pusztai, L., and Coombes, K. (2009). The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer informatics, 7, 199.

Tong, P., Chen, Y., Su, X. and Coombes, K. R. (2012). Systematic Identification of Bimodally Expressed Genes Using RNAseq Data. Bioinformatics, 2013 Mar 1;29(5):605-13.

## See Also

SIBER fitNB fitGP fitNL

## Examples

```
# artificial RNAseq data from negative binomial distribution
set.seed(1000)
dat <- rnbinom(100, mu=1000, size=1/0.2)
fitLN(y=dat)
```

fitNB *Fit Negative Binomial Mixture Model*

## Description

The function fits a two-component Negative Binomial mixture model.

## Usage

```
fitNB(y, d=NULL, inits=NULL, model='V', zeroPercentThr=0.2)
```

## Arguments

| | |
|---|---|
| y | A vector representing the RNAseq raw count. |
| d | A vector of the same length as y representing the normalization constant to be applied to the data. |
| inits | Initial value to fit the mixture model. A vector with elements mu1, mu2, phi1, phi2 and pi1. For 0-inflated model, only mu2, phi2, pi1 are used while the other elements can be arbitrary. |
| model | Character specifying E or V model. E model fits the mixture model with equal dispersion phi while V model doesn't put any constraint. |
| zeroPercentThr | A scalar specifying the minimum percent of zero counts needed when fitting a zero-inflated Negative Binomial model. This parameter is used to deal with zero-inflation in RNAseq count data. When the percent of zero exceeds this threshold, rather than fitting a 2-component negative binomial mixture, a mixture of point mass at 0 and negative binomial is fitted. |

## Details

This function directly maximize the log likelihood function through optimization. With this function, three models can be fitted: (1) negative binomial mixture with equal dispersion (E model); (2) negative binomial mixture with unequal dispersion (V model); (3) 0-inflated negative binomial model. The 0-inflated negative binomial has the following density function:

$P(Y = y) = \pi D(y) + (1 - \pi)NB(\mu, \phi)$ where D is the point mass at 0 while $NB(\mu, \phi)$ is the density of negative binomial distribution with mean $\mu$ and dispersion $\phi$. The variance is $\mu + \phi\mu^2$.

The rule to fit 0-inflated model is that the observed percentage of count exceeds the user specified threshold. This rule overrides the model argument when observed percentae of zero count exceeds the threshold.

## Value

A vector consisting parameter estimates of mu1, mu2, phi1, phi2, pi1, logLik and BIC. For 0-inflated model, mu1=phi1=0.

## Author(s)

Pan Tong (nickytong@gmail.com), Kevin R Coombes (krc@silicovore.com)

## References

Tong, P., Chen, Y., Su, X. and Coombes, K. R. (2012). Systematic Identification of Bimodally Expressed Genes Using RNAseq Data. Bioinformatics, 2013 Mar 1;29(5):605-13.

## See Also

SIBER fitLN fitGP fitNL

## Examples

```
# artificial RNAseq data from negative binomial distribution
set.seed(1000)
dat <- rnbinom(100, mu=1000, size=1/0.2)
fitNB(y=dat)
```

---

# fitNL *Fit Negative Binomial Mixture Model*

---

## Description

The function fits a two-component Negative Binomial mixture model.

## Usage

```
fitNL(y, d=NULL, model='E')
```

## Arguments

| | |
|---|---|
| y | A vector representing the transformed data that follows the normal mixture distribution. |
| d | A vector of the same length as y representing the normalization constant to be applied to the data. |
| model | Character specifying E or V model. E model fits the mixture model with equal variance while V model doesn't put any constraint. |

## Details

This function calls the mclust package to fit the 2-component normal mixture.

## Value

A vector consisting parameter estimates of mu1, mu2, phi1, phi2, pi1, logLik and BIC.

## Author(s)

Pan Tong (nickytong@gmail.com), Kevin R Coombes (krc@silicovore.com)

## References

Wang, J.,Wen, S., Symmans,W., Pusztai, L., and Coombes, K. (2009). The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data. Cancer informatics, 7, 199.

Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. Journal of the american statistical association, 97(458), 611:631.

Tong, P., Chen, Y., Su, X. and Coombes, K. R. (2012). Systematic Identification of Bimodally Expressed Genes Using RNAseq Data. Bioinformatics, 2013 Mar 1;29(5):605-13.

## See Also

SIBER fitLN fitNB fitGP

## Examples

```
# artificial microarray data from normal distribution
set.seed(1000)
dat <- rnorm(100, mean=5, sd=1)
fitNL(y=dat)
```

---

`SIBER` *Fit Mixture Model on The RNAseq Data and Calculates Bimodality Index*

---

## Description

The function fits a two-component mixture model and calculate BI from the parameter estimates.

## Usage

```
SIBER(y, d=NULL, model=c('LN', 'NB', 'GP', 'NL'), zeroPercentThr=0.2, base=exp(1), eps=10)
```

## Arguments

| | |
|---|---|
| `y` | A vector representing the RNAseq raw count or the transformed values if model=NL. |
| `d` | A vector of the same length as y representing the normalization constant to be applied to the data. |
| `model` | Character string specifying the mixture model type. It can be any of LN, NB, GP and NL. |

`zeroPercentThr` A scalar specifying the minimum percent of zero to detect using log normal mixture. This parameter is used to deal with zero-inflation in RNAseq count data. When the percent of zero exceeds this threshold, 1-comp mixture LN model is used to estimate mu and sigma from nonzero count. This parameter is relevant only if model='LN'.

`base` The logarithm base defining the parameter estimates in the logarithm scale from LN model . It is relevant only if model='LN'.

`eps` A scalar to be added to the count data when model='LN'. This parameter is relevant only when model='LN'.

## Details

SIBER proceeds in two steps. The first step fits a two-component mixture model. The second step calculates the Bimodality Index corresponding to the assumed mixture distribution. Four types of mixture models are implemented: log normal (LN), Negative Binomial (NB), Generalized Poisson (GP) and normal mixture (NL). The normal mixture model was developed to identify bimodal genes from microarray data in Wang et al. It is incorporated here in case the user has already transformed the RNAseq data.

Behind the scene, SIBER calls the fitNB, fitGP, fitLN and fitNL function with model=E depending on which distribution model is specified. When the observed percentage of count exceeds the user specified threshold zeroPercentThr, the 0-inflated model overrides the E model and will be fitted.

Type vignette('SIBER') in the R console to pull out the user manual in pdf format.

## Value

A vector consisting estimates of mu1, mu2, sigma1, sigma2, p1, delta and BI.

## Author(s)

Pan Tong (nickytong@gmail.com), Kevin R Coombes (krc@silicovore.com)

## References

Wang J, Wen S, Symmans WF, Pusztai L, Coombes KR.
*The bimodality index: a criterion for discovering and ranking bimodal signatures from cancer gene expression profiling data.*
Cancer Inform. 2009 Aug 5;7:199-216.

Tong P, Chen Y, Su X, Coombes KR.
*SIBER: systematic identification of bimodally expressed genes using RNAseq data.*
Bioinformatics. 2013 Mar 1;29(5):605-13.

## See Also

fitLN fitNB fitGP fitNL

## Examples

```
# artificial RNAseq data from negative binomial distribution
set.seed(1000)
dat <- rnbinom(100, mu=1000, size=1/0.2)
# fit SIBER with the 4 mixture models
SIBER(y=dat, model='LN')
SIBER(y=dat, model='NB')
SIBER(y=dat, model='GP')
SIBER(y=log(dat+1), model='NL')
```

---

`simDat` *Simulated Data From 2-component Mixture Models*

---

## Description

Data from 2-component mixture models (NB, GP and LN) is simulated with the true parameters given for testing and illustration purpose.

## Usage

```
data(simDat)
```

## Format

The data frame contains the following data objects:

**parList** A list of true parameters. There are three named elements (NB, GP and LN) corresponding to the parameters used to simulate gene expression data from NB, GP and LN mixture models. Each element is a 6 by 5 matrix giving the true parameters generating the simulated data.

**dataList** A list of matrices for simulated gene expression data. There are three named elements (NB, GP and LN) corresponding to the simulate gene expression data from NB, GP and LN mixture models. Each element is a 6 by 200 matrix. That is, 6 genes (rows) are simulated with 200 samples (columns). The first 3 genes in each matrix are from 2-component mixture model while the last 3 genes are from 0-inflated models.

## Author(s)

Pan Tong (nickytong@gmail.com), Kevin R Coombes (krc@silicovore.com)

## See Also

SIBER fitNB fitGP fitLN fitNL

## Examples

```
library(SIBERG)
data(simDat)
sapply(parList, dim)
sapply(dataList, dim)
```

# Index