Package 'preseqR'

June 27, 2018

Type Package

Title Predicting Species Accumulation Curves

Version 4.0.0

Date 2018-06-27

Author Chao Deng, Timothy Daley and Andrew D. Smith

Maintainer Chao Deng <chaodeng@usc.edu>

Description Originally as an R version of Preseq <doi:10.1038/nmeth.2375>, the package has extended its functionality to predict the r-species accumulation curve (r-SAC), which is the number of species represented at least r times as a function of the sampling effort. When r = 1, the curve is known as the species accumulation curve, or the library complexity curve in high-throughput genomic sequencing. The package includes both parametric and nonparametric methods, as described by Deng C, et al. (2018) <arXiv:1607.02804v3>.

License GPL-3

Imports polynom, graphics, stats

NeedsCompilation no

Repository CRAN

Date/Publication 2018-06-27 10:52:56 UTC

R topics documented:

`preseqR-package` *Predicting r-species accumulation curves*

Description

The functionality of this package is to predict r-species accumulaiton curves. The method is based on a nonparametric empirical Bayes approach with rational function approximation. The estimator is excellent in accuracy for both large values of r and long-range extrapolations, which are essential to large-scale applications. Some examples are predicting the molecular complexity of sequencing libraries, estimating the minimum sufficient sequencing depths for whole-exome sequencing experiments and optimizing depths for single-cell whole-genome sequencing experiments.

Details

main functions:

preseqR.rSAC

preseqR.rSAC.bootstrap

preseqR.optimal.sequencing

preseqR.rSAC.sequencing.rmdup

preseqR.sample.cov

preseqR.sample.cov.bootstrap

Author(s)

Chao Deng, Timothy Daley, and Andrew D. Smith

Maintainer: Chao Deng <chaodeng@usc.edu>

References

Baker, G. A., & Graves-Morris, P. (1996). Pade approximants (Encyclopedia of Mathematics and its Applications vol 59).

Boneh, S., Boneh, A., & Caron, R. J. (1998). Estimating the prediction function and the number of unseen species in sampling with replacement. Journal of the American Statistical Association, 93(441), 372-379.

Chao, A., & Shen, T. J. (2004). Nonparametric prediction in species sampling. Journal of agricultural, biological, and environmental statistics, 9(3), 253-269.

Cohen Jr, A. C. (1960). Estimating the parameters of a modified Poisson distribution. Journal of the American Statistical Association, 55(289), 139-143.

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature methods, 10(4), 325-327.

Deng C, Daley T & Smith AD (2015). Applications of species accumulation curves in large-scale biological data analysis. Quantitative Biology, 3(3), 135-144. URL http://dx.doi.org/10.1007/s40484-015-0049-7.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?. Biometrika, 63(3), 435-447.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. The annals of Statistics, 1-26.

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Fisher, R. A., Corbet, A. S., and Williams, C. B. ,1943, The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, Journal of Animal Ecology, 12, 42-58.

Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. Biometrika, 43(1-2), 45-63.

Heck Jr, K. L., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology, 1459-1461.

Kalinin V (1965). Functionals related to the poisson distribution and statistical structure of a text. Articles on Mathematical Statistics and the Theory of Probability pp. 202-220.

bbc.rSAC *BBC estimator*

Description

bbc.rSAC predicts the expected number of species represented at least r times in a random sample, based on the initial sample. The estimator was originally proposed by Boneh et al. (1998) for estimating the SAC. We generalize this estimator for predicting the r-SAC.

Usage

```
bbc.rSAC(n, r=1)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
r	A positive integer. Default is 1.

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Boneh, S., Boneh, A., & Caron, R. J. (1998). Estimating the prediction function and the number of unseen species in sampling with replacement. Journal of the American Statistical Association, 93(441), 372-379.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for SAC
```

```
bbc1 <- bbc.rSAC(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
bbc1(c(10, 20))

## construct the estimator for r-SAC
bbc2 <- bbc.rSAC(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
bbc2(c(50, 100))
```

`cs.rSAC` *CS estimator*

Description

`cs.rSAC` predicts the expected number of species represented at least r times in a random sample, based on the initial sample. The estimator was originally proposed by Chao and Shen (2004) for estimating the SAC. We generalize this estimator for predicting the r-SAC.

Usage

```
cs.rSAC(n, r=1, k=10)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A positive integer. Default is 1.
`k`	A cutoff for common species. Default is 10.

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Chao, A., & Shen, T. J. (2004). Nonparametric prediction in species sampling. Journal of agricultural, biological, and environmental statistics, 9(3), 253-269.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for SAC
chao1 <- cs.rSAC(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
chao1(c(10, 20))

## construct the estimator for r-SAC
chao2 <- cs.rSAC(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
chao2(c(50, 100))
```

Dickens *Dickens' vocabulary*

Description

Words frequencies of a collection of Charles Dickens from Project Gutenberg

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of unique words appeared exactly j times in a collection of Charles Dickens.

References

http://zipfr.r-forge.r-project.org/

Examples

```
##load library
library(preseqR)

##load data
data(Dickens)
```

ds.rSAC *RFA estimator*

Description

ds.rSAC predicts the expected number of species represented at least r times in a random sample, based on the initial sample.

Usage

```
ds.rSAC(n, r=1, mt=20)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
mt	An positive integer constraining possible rational function approximations. Default is 20.
r	A positive integer. Default is 1.

Details

The estimator is based on an empirical Bayes approach using rational function approximation (RFA), as described in the paper in the references section.

ds.rSAC is the fast version of ds.rSAC.bootstrap. The function does not provide the confidence interval. To obtain the confidence interval along with the estimates, one should use the function ds.rSAC.bootstrap.

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for SAC
ds1 <- ds.rSAC(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
ds1(c(10, 20))

## construct the estimator for r-SAC
ds2 <- ds.rSAC(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
ds2(c(50, 100))
```

`ds.rSAC.bootstrap` *RFA estimator with bootstrap*

Description

`ds.rSAC.bootstrap` predicts the expected number of species represented at least r times in a random sample, based on the initial sample.

Usage

```
ds.rSAC.bootstrap(n, r=1, mt=20, times=30, conf=0.95)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A positive integer. Default is 1.
`mt`	An positive integer constraining possible rational function approximations. Default is 20.
`times`	The number of bootstrap samples. Default is 30.
`conf`	The confidence level. Default is 0.95

Details

This is the bootstrap version of ds.rSAC. The bootstrap sample is generated by randomly sampling the initial sample with replacement. For each bootstrap sample, we construct an estimator. The median of estimates is used as the prediction for the number of species represented at least r times in a random sample.

The confidence interval is constructed based on a lognormal distribution.

Value

f	The estimator for the number of species represented at least r times in a random sample. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means a random sample that is twice the size of the initial sample.
se	The standard error for the estimator. The input is a vector of sampling efforts t.
lb	The lower bound of the confidence interval.The input is a vector of sampling efforts t.
ub	The upper bound of the confidence interval.The input is a vector of sampling efforts t.

Author(s)

Chao Deng

References

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
# library(preseqR)

## import data
# data(FisherButterfly)

## construct the estimator for SAC
# ds1 <- ds.rSAC.bootstrap(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
# ds1$f(c(10, 20))
## The standard error of the estiamtes
# ds1$se(c(10, 20))
## The confidence interval of the estimates
# lb <- ds1$lb(c(10, 20))
# ub <- ds1$ub(c(10, 20))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)
```

```
## construct the estimator for SAC
# ds2 <- ds.rSAC.bootstrap(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
# ds2$f(c(50, 100))
## The standard error of the estiamtes
# ds2$se(c(50, 100))
## The confidence interval of the estimates
# lb <- ds2$lb(c(50, 100))
# ub <- ds2$ub(c(50, 100))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)
```

`fisher.alpha` *Parameter alpha in the logseries estimator*

Description

`fisher.alpha` estimates the parameter alpha in the logseries estimator by Fisher, R. A., et al. (1943) based on an initial sample.

Usage

```
fisher.alpha(n)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.

Value

A double, the estimated value of the parameter alpha

Author(s)

Chao Deng

References

Fisher, R., Corbet, A., & Williams, C. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. Journal of Animal Ecology, 12(1), 42-58. doi:10.2307/1411

Examples

```
## load library
library(preseqR)

## import data
data(WillButterfly)

## estimating alpha
fisher.alpha <- fisher.alpha(WillButterfly)
```

`fisher.rSAC` *Logseries estimator*

Description

`fisher.rSAC` estimates the expected number of species represented at least r times in a random sample, based on the initial sample. The estimator was originally proposed by Fisher et al. (1943) for estimating the SAC. We generalize this estimator for predicting the r-SAC.

Usage

```
fisher.rSAC(n, r=1)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A positive integer. Default is 1.

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Fisher, R., Corbet, A., & Williams, C. (1943). The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. Journal of Animal Ecology, 12(1), 42-58. doi:10.2307/1411

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
library(preseqR)

## import data
data(WillButterfly)

## construct the estimator for SAC
fisher1 <- fisher.rSAC(WillButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
fisher1(c(10, 20))

## construct the estimator for r-SAC
fisher2 <- fisher.rSAC(WillButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
fisher2(c(50, 100))
```

FisherButterfly *Fisher's butterfly data*

Description

Frequencies data of butterflies collected in the Malay peninsula was from Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943).

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of butterflies captured j times in the sample.

References

Fisher, R. A., Corbet, A. S., and Williams, C. B. ,1943, The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, Journal of Animal Ecology, 12, 42-58, Table 1,2.

Examples

```
##load library
library(preseqR)

##load data
data(FisherButterfly)
```

`kmer.frac.curve` *Fraction of k-mers observed at least r times*

Description

`kmer.frac.curve` predicts the expected fraction of k-mers observed at least r times in a high-throughput sequencing experiment given the amount of sequencing

Usage

```
kmer.frac.curve(n, k, read.len, seq, r=2, mt=20)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of k-mers observed exactly j times in the initial experiment. The first column must be sorted in an ascending order.
`k`	The number of nucleotides in a k-mer.
`read.len`	The average length of a read.
`seq`	The amount of nucleotides sequenced..
`r`	A positive integer. Default is 1.
`mt`	An positive integer constraining possible rational function approximations. Default is 20.

Details

`kmer.frac.curve` is mainly designed for metagenomics to evaluate how saturated a metagenomic data is.

`kmer.frac.curve` is the fast version of `kmer.frac.curve.bootstrap`. The function does not provide the confidence interval. To obtain the confidence interval along with the estimates, one should use the function `kmer.frac.curve.bootstrap`.

Value

A two-column matrix. The first column is the amount of sequencing in an experiment. The second column is the estimate of the fraction of k-mers observed at least r times in the experiment.

Author(s)

Chao Deng

References

Deng, C and Smith, AD (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804

Examples

```
## load library
library(preseqR)

## import data
data(SRR061157_k31)

## the fraction of 31-mers represented at least 10 times in an experiment when
## sequencing 1M, 10M, 100M, 1G, 10G, 100G, 1T nucleotides
kmer.frac.curve(n=SRR061157_k31, k=31, read.len=100, seq=10^(6:12), r=10, mt=20)
```

`kmer.frac.curve.bootstrap`

Fraction of k-mers observed at least r times with bootstrap

Description

`kmer.frac.curve` predicts the expected fraction of k-mers observed at least r times in a high-throughput sequencing experiment given the amount of sequencing

Usage

```
kmer.frac.curve.bootstrap(n, k, read.len, seq, r=2, mt=20, times=30, conf=0.95)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of k-mers observed exactly j times in the initial experiment. The first column must be sorted in an ascending order.
`k`	The number of nucleotides in a k-mer.
`read.len`	The average length of a read.
`seq`	The amount of nucleotides sequenced.
`r`	A positive integer. Default is 1.
`mt`	An positive integer constraining possible rational function approximations. Default is 20.
`times`	The number of bootstrap samples.
`conf`	The confidence level. Default is 0.95

Details

This is the bootstrap version of `kmer.frac.curve`. The bootstrap sample is generated by randomly sampling the initial sample with replacement. For each bootstrap sample, we construct an estimator. The median of estimates is used as the prediction for the number of species represented at least r times in a random sample.

The confidence interval is constructed based on a lognormal distribution.

Value

A four-column matrix. The first column is the amount of sequencing in an experiment. The second column is the estimate of the fraction of k-mers observed at least r times in the experiment. The third and fourth columns are the lower bounds and the upper bounds of the confidence intervals.

Author(s)

Chao Deng

References

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
# library(preseqR)

## import data
# data(SRR061157_k31)

## the fraction of 31-mers represented at least 10 times in an experiment when
## sequencing 1M, 10M, 100M, 1G, 10G, 100G, 1T nucleotides
# kmer.frac.curve.bootstrap(n=SRR061157_k31, k=31, read.len=100,
#                           seq=10^(6:12), r=10, mt=20)
```

---

preseqR.interpolate.rSAC

Interpolation

Description

Interpolating the number of species represented at least r times in a subsample given an initial sample

Usage

```
preseqR.interpolate.rSAC(n, ss, r=1)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
ss	A positive double equal to the step size between subsamples.
r	A positive integer. Default is 1

Details

The expected number of species represented at least r times in the subsample is estimated based on an expended version of the formula by Heck Jr, KL. et al. (1975).

Value

A two-column matrix for the number of species represented at least r times in a random sample. The first column is the size of the random sample; the second column is the expected number of species represented at least r times in the sample.

NULL if failed.

Author(s)

Chao Deng

References

Heck Jr, K. L., van Belle, G., & Simberloff, D. (1975). Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology, 1459-1461.

Examples

```
## load library
library(preseqR)

## import data
data(Shakespeare)

## The expected number of distinct words represented twice or more in the
## subsample
preseqR.interpolate.rSAC(n=Shakespeare, ss=1e5, r=2)
```

---

preseqR.nonreplace.sampling

Sampling

Description

Generating a histogram by subsampling without replacement.

Usage

```
preseqR.nonreplace.sampling(n, size)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
size	An positive integer representing the size of the subsample.

Details

preseqR.nonreplace.sampling generates a subsample by sampling the initial sample without replacement. sample in R is used to implement the function. We wrap up this function in such a way that both the input and the output are histograms.

Value

A two-column matrix as a subsample. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species represented j times in the subsample.

Author(s)

Chao Deng

References

https://stat.ethz.ch/R-manual/R-patched/library/base/html/sample.html

Examples

```
## load library
library(preseqR)
## import data
data(FisherButterfly)
## generate a subsample of size 1000.
preseqR.nonreplace.sampling(n=FisherButterfly, size=1000)
```

---

preseqR.optimal.sequencing

Optimal amount of sequencing for scWGS

Description

preseqR.optimal.sequencing predicts the optimal amount of sequencing in a single-cell whole-genome sequencing (scWGS) experiment based on a shallow sequencing experiment.

Usage

```
preseqR.optimal.sequencing(n, efficiency=0.05, bin=1e8, r=1, mt=20,
                           times=30, conf=0.95)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
efficiency	The minimum benefit-cost ratio
bin	One unit of sequencing effort. Default is 1e8.
r	A positive integer. Default is 1.
mt	An positive integer constraining possible rational function approximations. Default is 20.
times	The number of bootstrap samples.
conf	The confidence level. Default is 0.95

Details

preseqR.optimal.sequencing predicts the optimal amount of sequencing in a scWGS experiment. The term optimal is interpreted as the maximum amount of sequencing with its benefit-cost ratio greater than a given threshold. The benefit-cost ratio is defined as the probability of a new nucleotide in the genome represented at least r times when one more base is sequenced. In order to improve the numeric stability, we use the mean of new nucleotdies with coverage at least r in one unit of sequencing effort to approximate the ratio. The amount of sequences in one unit of sequencing effort is defined by the variable bin.

Note that the benefit-cost ratio is not monotonic. The ratio first increases and then decrease as the amount of sequencing increase. To predicte the optimal amount of sequencing, we consider only the areas after the peak, where the ratio starts to decrease.

Value

A vector of three dimensions. The first coordinate is the optimal amount of sequencing. The second and the third coordinates are the lower and upper bound of the confidence interval.

Author(s)

Chao Deng

References

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
#library(preseqR)

## import data
# data(SRR611492_5M)
```

```
## the optimal amount of sequencing with the benefit-cost ratio greater than
## 0.05 for r = 4
# preseqR.optimal.sequencing(n=SRR611492_5M, efficiency=0.05, bin=1e8, r=4)
## the optimal amount of sequencing with the benefit-cost ratio greater than
## 0.05 for r = 10
# preseqR.optimal.sequencing(n=SRR611492_5M, efficiency=0.05, bin=1e8, r=10)
```

`preseqR.rSAC` *Best practice for r-SAC – a fast version*

Description

`preseqR.rSAC` predicts the expected number of species represented at least r times in a random sample based on the initial sample.

Usage

```
preseqR.rSAC(n, r=1, mt=20, size=SIZE.INIT, mu=MU.INIT)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`mt`	A positive integer constraining possible rational function approximations. Default is 20.
`r`	A positive integer. Default is 1.
`size`	A positive double, the initial value of the parameter `size` in the negative binomial distribution for the EM algorithm. Default value is 1.
`mu`	A positive double, the initial value of the parameter `mu` in the negative binomial distribution for the EM algorithm. Default value is 0.5.

Details

`preseqR.rSAC` combines the nonparametric approach using the rational function approximation and the parametric approach using the zero-truncated negative binomial (ZTNB). For a given initial sample, if the sample is from a heterogeneous population, the function calls `ds.rSAC`; otherwise it calls `ztnb.rSAC`. The degree of heterogeneity is measured by the coefficient of variation, which is estimated by the ZTNB approach.

`preseqR.rSAC` is the fast version of `preseqR.rSAC.bootstrap`. The function does not provide the confidence interval. To obtain the confidence interval along with the estimates, one should use the function `preseqR.rSAC.bootstrap`.

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for SAC
estimator1 <- preseqR.rSAC(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
estimator1(c(10, 20))

## construct the estimator for r-SAC
estimator2 <- preseqR.rSAC(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
estimator2(c(50, 100))
```

---

preseqR.rSAC.bootstrap

Best practice for r-SAC

Description

$preseqR.rSAC.bootstrap$ predicts the expected number of species represented at least r times in a random sample based on the initial sample.

Usage

```
preseqR.rSAC.bootstrap(n, r=1, mt=20, size=SIZE.INIT, mu=MU.INIT, times=30,
                       conf=0.95)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
r	A positive integer. Default is 1.
mt	An positive integer constraining possible rational function approximations. Default is 20.
times	The number of bootstrap samples.
size	A positive double, the initial value of the parameter size in the negative binomial distribution for the EM algorithm. Default value is 1.
mu	A positive double, the initial value of the parameter mu in the negative binomial distribution for the EM algorithm. Default value is 0.5.
conf	The confidence level. Default is 0.95

Details

This is the bootstrap version of preseqR.rSAC. The bootstrap sample is generated by randomly sampling the initial sample with replacement. For each bootstrap sample, we construct an estimator. The median of estimates is used as the prediction for the number of species represented at least r times in a random sample.

The confidence interval is constructed based on a lognormal distribution.

Value

f	The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.
se	The standard error for the estimator. The input is a vector of sampling efforts t.
lb	The lower bound of the confidence interval.The input is a vector of sampling efforts t.
ub	The upper bound of the confidence interval.The input is a vector of sampling efforts t.

Author(s)

Chao Deng

References

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
# library(preseqR)

## import data
# data(FisherButterfly)

## construct estimator for SAC
# estimator1 <- preseqR.rSAC.bootstrap(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
# estimator1$f(c(10, 20))
## The standard error of the estiamtes
# estimator1$se(c(10, 20))
## The confidence interval of the estimates
# lb <- estimator1$lb(c(10, 20))
# ub <- estimator1$ub(c(10, 20))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)

## construct estimator for r-SAC
# estimator2 <- preseqR.rSAC.bootstrap(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
# estimator2$f(c(50, 100))
## The standard error of the estiamtes
# estimator2$se(c(50, 100))
## The confidence interval of the estimates
# lb <- estimator2$lb(c(50, 100))
# ub <- estimator2$ub(c(50, 100))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)
```

`preseqR.rSAC.sequencing.rmdup`

Predicting r-SAC in WES/WGS

Description

`preseqR.rSAC.sequencing.rmdup` predicts the expected number of nucleotides in the genome sequenced at least r times in a sequencing experiment, based on a shallow sequencing experiment.

Usage

```
preseqR.rSAC.sequencing.rmdup(n_base, n_read, r=1, mt=20, times=30, conf=0.95)
```

Arguments

`n_base`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of nucleotides in the genome sequenced exactly j times in the initial experiment. The first column must be sorted in an ascending order.

n_read	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N'_j, the number of distinct reads with exactly j duplicates in the initial experiment. The first column must be sorted in an ascending order.
r	A positive integer. Default is 1.
mt	An positive integer constraining possible rational function approximations. Default is 20.
times	The number of bootstrap samples. Default is 30.
conf	The confidence level. Default is 0.95

Details

preseqR.rSAC.sequencing.rmdup is designed for sequencing experiments, where duplicate reads are removed. The procedure is commonly used in whole-exome sequencing experiments and sometimes appeared in WGS as well. To use the function, one must have two histograms. The first histogram is the coverage histogram, which is based on distinct reads. The second histogram is the counts of reads with exactly j duplicates.

Value

f	The estimator for the expected number of nucleotides in the genome sequenced at least r times given the amount of sequencing. The input of the estimator is a vector of sequencing efforts t, i.e. the relative amount of sequencing comparing with the amount in the initial experiment. For example, t = 2 means sequencing twice the amount of the initial experiment.
se	The standard error for the estimator. The input is a vector of sequencing efforts t.
lb	The lower bound of the confidence interval.The input is a vector of sequencing efforts t.
ub	The upper bound of the confidence interval.The input is a vector of sequencing efforts t.

Author(s)

Chao Deng

References

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
#library(preseqR)

## import data
# data(SRR1301329_1M_base)
# data(SRR1301329_1M_read)
```

```
# construct the estimator
# estimator1 <- preseqR.rSAC.sequencing.rmdup(
#                 n_base=SRR1301329_1M_base, n_read=SRR5365359_5M_read,
#                 r=4, mt=20, times=100, conf=0.95)
## The number of nucleotides in the genome covered at least 4 times, when the
## amount of sequencing is 10 or 20 times of the intial experiment
## 10 or 20 times of the initial sample
# estimator1$f(c(10, 20))
## The standard error of the estiamtes
# estimator1$se(c(10, 20))
## The confidence interval of the estimates
# lb <- estimator1$lb(c(10, 20))
# ub <- estimator1$ub(c(10, 20))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)

# construct the estimator
# estimator2 <- preseqR.rSAC.sequencing.rmdup(
#                 n_base=SRR1301329_1M_base, n_read=SRR5365359_5M_read,
#                 r=10, mt=20, times=100, conf=0.95)
## The number of nucleotides in the genome covered at least 10 times, when the
## amount of sequencing is 10 or 20 times of the intial experiment
## 10 or 20 times of the initial sample
# estimator2$f(c(10, 20))
## The standard error of the estiamtes
# estimator2$se(c(10, 20))
## The confidence interval of the estimates
# lb <- estimator2$lb(c(10, 20))
# ub <- estimator2$ub(c(10, 20))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)
```

`preseqR.sample.cov` *Predicting generalized sample coverage*

Description

`preseqR.sample.cov` predicts the probability of observing a species represented at least r times in a random sample.

Usage

```
preseqR.sample.cov(n, r=1, mt=20)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.

r	A positive integer. Default is 1.
mt	A positive integer constraining possible rational function approximations. Default is 20.

Details

Suppose a sample is given and one more individual is randomly drawn from the population. preseqR.sample.cov estimates the probability of the species, which represents the individual, has been observed at least r times in the sample. When $r = 1$, the probability is called the sample coverage.

Let N_j be the number of species represented exactly j times in a sample. The probability of observing a species represented at least r times in the sample is estimated as $\sum_{j=r+1}^{\infty} jN_j / \sum_{j=1}^{\infty} jN_j$. The theory is described by Mao and Lindsay (2002). For a random sample where N_j is unknown, a modified rational function approximation is first used to predict the value of N_j. Then the estimates are substituted to obtain an estimator for the probability of observing a species represented at least r times in the sample.

This function is the fast version of preseqR.sample.cov.bootstrap. The function does not provide the confidence interval. To obtain the confidence interval along with the estimates, one should use the function preseqR.sample.cov.bootstrap.

Value

The estimator for the probability of observing a species represented at least r times in a random sample. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. Biometrika, 40(3-4), 237-264.

Mao, C. X. and Lindsay, B. G. (2002). A Poisson model for the coverage problem with a genomic application. Biometrika, 89(3), 669-682.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for the sample coverage
estimator1 <- preseqR.sample.cov(FisherButterfly, r=1)
```

```
## Given a sample that is 10 times or 20 times the size of an initial samples,
## suppose one randomly draws one more individual from the population. The
## value of the function is the probability that the representing species
## has been observed in the sample
estimator1(c(10, 20))

## construct the estimator
estimator2 <- preseqR.sample.cov(FisherButterfly, r=2)
## the probability a species represented at least twice when the sample size
## is 50 times or 100 times of the initial sample
estimator2(c(50, 100))
```

`preseqR.sample.cov.bootstrap`

Predicting generalized sample coverage with bootstrap

Description

`preseqR.sample.cov.bootstrap` predicts the probability of observing a species represented at least r times in a random sample.

Usage

```
preseqR.sample.cov.bootstrap(n, r=1, mt=20, times=30, conf=0.95)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A positive integer. Default is 1.
`mt`	A positive integer constraining possible rational function approximations. Default is 20.
`times`	The number of bootstrap samples. Default is 30.
`conf`	The confidence level. Default is 0.95

Details

This is the bootstrap version of `preseqR.sample.cov`. The bootstrap sample is generated by randomly sampling the initial sample with replacement. For each bootstrap sample, we construct an estimator. The median of estimates is used as the prediction for the number of species represented at least r times in a random sample.

The confidence interval is constructed based on a lognormal distribution.

Value

f	The estimator for the probability of observing a species represented at least r times in a sample as a function of the sample size. The input of the estimator is a vector of sampling efforts t, i.e. the relative sample sizes comparing with the initial sample. For example, t = 2 means a random sample that is twice the size of the initial sample.
se	The standard error for the estimator. The input is a vector of sampling efforts t.
lb	The lower bound of the confidence interval.The input is a vector of sampling efforts t.
ub	The upper bound of the confidence interval.The input is a vector of sampling efforts t.

Author(s)

Chao Deng

References

Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

Deng, C., Daley, T., Calabrese, P., Ren, J., & Smith, A.D. (2016). Estimating the number of species to attain sufficient representation in a random sample. arXiv preprint arXiv:1607.02804v3.

Examples

```
## load library
#library(preseqR)

## import data
#data(FisherButterfly)

## construct the estimator for the sample coverage
# estimator1 <- preseqR.sample.cov.bootstrap(FisherButterfly, r=1)
## Given a sample that is 10 times or 20 times the size of an initial samples,
## suppose one randomly draws one more individual from the population. The
## value of the function is the probability that the representing species
## has been observed in the sample
# estimator1$f(c(10, 20))
## The standard error of the estiamtes
# estimator1$se(c(10, 20))
## The confidence interval of the estimates
# lb <- estimator1$lb(c(10, 20))
# ub <- estimator1$ub(c(10, 20))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)

## construct the estimator
# estimator2 <- preseqR.rSAC.bootstrap(FisherButterfly, r=2)
## the probability when the sample size is 50 times or 100 times of the initial
## sample
# estimator2$f(c(50, 100))
## The standard error of the estiamtes
```

```
# estimator2$se(c(50, 100))
## The confidence interval of the estimates
# lb <- estimator2$lb(c(50, 100))
# ub <- estimator2$ub(c(50, 100))
# matrix(c(lb, ub), byrow=FALSE, ncol=2)
```

`preseqR.simu.hist` *Simulation*

Description

Generating a histogram based on a Poisson mixture model.

Usage

```
preseqR.simu.hist(L=1e8, N, FUN)
```

Arguments

`L`	A positive integer, the number of species in a population.
`N`	A positive interger, the simulated sample size.
`FUN`	An RNG generating non negative real number.

Details

`preseqR.simu.hist` uses a mixture of Poisson distributions to generate a sample, which size is defined by the variable N. The statistical assumption is that for each species the number of individuals captured in a sample follows a Poisson process. The Poisson rates among species are generated by a given function `FUN` per unit of sampling effort.

`FUN` must take an argument indicating the number of random numbers generated and return a vector of generated numbers.

Value

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.

Author(s)

Chao Deng

Examples

```
## load library
library(preseqR)
## construct a RNG
f <- function(n) {
  rgamma(n, shape=0.5, scale=1)
}
## sample 10,000 individuals
preseqR.simu.hist(L=1e5, N=10000, f)
```

`preseqR.ztnb.em` *Fitting a zero-truncated negative binomial distribution*

Description

`preseqR.ztnb.em` fits a zero-truncated negative binomial (ZTNB) distribution to the initial sample. Since the species with zero observations are missed in the sample, an EM algorithm is used to estimate the parameters assuming the number of individuals for each species follows a Negative Binomial distribution with the zero counts as a missing latent data.

Usage

```
preseqR.ztnb.em(n, size = SIZE.INIT, mu = MU.INIT)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`size`	A positive double setting the initial value of the parameter `size` in a negative binomial distribution for the EM algorithm. Default value is 1.
`mu`	A positive double setting the initial value of the parameter `mu` in a negative binomial distribution for the EM algorithm. Default value is 0.5.

Details

See the supplement of Daley and Smith (2013).

Value

`size`	The estimate of the parameter `size` in the negative binomial.
`mu`	The estimate of the parameter `mu` in the negative binomial.
`loglik`	Log-likelihood under estimated ZTNB.

Author(s)

Chao Deng

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## print the parameters of a fitting negative binomial distribution
preseqR.ztnb.em(FisherButterfly)
```

Shakespeare *Shakespeare's word type frequencies*

Description

The Shakespeare's word type frequencies data was from Efron, B., & Thisted, R. (1976).

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of unique words appeared j times in Shakespeare's work.

References

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know?. Biometrika, 63(3), 435-447.

Examples

```
##load library
library(preseqR)

##load data
data(Shakespeare)
```

SRR061157_k31 *k-mer counts of a metagenomic data*

Description

The k-mer counts are based on a metagenome sequencing data from Human Microbiome Project with the accession number SRR061157. Only forward reads are used to generate the k-mer counts.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of 31-mers observed exactly j times.

References

Human Microbiome Project (https://hmpdacc.org/).

Examples

```
##load library
library(preseqR)

##load data
data(SRR061157_k31)
```

SRR1301329_1M_base *Coverage histogram of a WES data*

Description

The coverage histogram is based on a whole-exome sequencing (WES) data from Simons Foundation Autism Research Initiative with the accession number SRR1301329. One million reads are randomly sampled from the raw data to generate this coverage histogram.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of nucleotides in the genome covered exactly j times.

References

Simons Foundation Autism Research Initiative (https://www.sfari.org/).

Examples

```
##load library
library(preseqR)

##load data
data(SRR1301329_1M_base)
```

SRR1301329_1M_read *Read counts of a WES data*

Description

The read counts are based on a whole-exome sequencing (WES) data from Simons Foundation Autism Research Initiative with the accession number SRR1301329. One million reads are randomly sampled from the raw data to generate the read counts.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of reads observed exactly j times in the data.

References

Simons Foundation Autism Research Initiative (https://www.sfari.org/).

Examples

```
##load library
library(preseqR)

##load data
data(SRR1301329_1M_read)
```

SRR1301329_base *Coverage histogram of a WES data*

Description

The coverage histogram is based on a whole-exome sequencing (WES) data from Simons Foundation Autism Research Initiative with the accession number SRR1301329. Only forward reads are used to generate the coverage histogram.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of nucleotides in the genome covered exactly j times.

References

Simons Foundation Autism Research Initiative (https://www.sfari.org/).

Examples

```
##load library
library(preseqR)

##load data
data(SRR1301329_base)
```

SRR1301329_read *Read counts of a WES data*

Description

The read counts are based on a whole-exome sequencing data from Simons Foundation Autism Research Initiative with the accession number SRR1301329. Only forward reads are used to generate the read counts.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of reads observed exactly j times in the data.

References

Simons Foundation Autism Research Initiative (https://www.sfari.org/).

Examples

```
##load library
library(preseqR)

##load data
data(SRR1301329_read)
```

SRR611492 *Coverage histogram of a scWGS data*

Description

The coverage histogram is based on a single-cell whole-genome sequencing data (scWGS) through MALBAK protocol. The accession number of the raw data is SRR1301329. Only forward reads are used to generate the coverage histogram.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of nucleotides in the genome covered exactly j times.

References

Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science, 338(6114), 1622-1626.

Examples

```
##load library
library(preseqR)

##load data
data(SRR611492)
```

SRR611492_5M *Coverage histogram of a scWGS data*

Description

The coverage histogram is based on a single-cell whole-genome sequencing (scWGS) data through MALBAK protocol. The accession number of the raw data is SRR1301329. Five million reads are randomly sampled from the raw data to generate this coverage histogram.

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of nucleotides in the genome covered exactly j times.

References

Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science, 338(6114), 1622-1626.

Examples

```
##load library
library(preseqR)

##load data
data(SRR1301329_5M)
```

`Twitter` *Social network*

Description

Following relationships of Twitter's social network

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of users with exactly j followers.

References

Zafarani R, Liu H (2009) Social computing data repository at ASU.

Examples

```
##load library
library(preseqR)

##load data
data(Twitter)
```

`WillButterfly` *Fisher's butterfly data*

Description

Frequencies data of butterflies collected in the Malay peninsula was from Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943).

Details

A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is n_j, the number of butterflies captured exactly j times in the sample.

References

Fisher, R. A., Corbet, A. S., and Williams, C. B. ,1943, The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population, Journal of Animal Ecology, 12, 42-58, Table 3.

Examples

```
##load library
library(preseqR)

##load data
data(WillButterfly)
```

`ztnb.rSAC` *ZTNB estimator*

Description

`ztnb.rSAC` predicts the expected number of species represented at least r times in a random sample, based on the initial sample.

Usage

```
ztnb.rSAC(n, r=1, size=SIZE.INIT, mu=MU.INIT)
```

Arguments

`n`	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
`r`	A positive integer. Default is 1.
`size`	A positive double, the initial value of the parameter `size` in the negative binomial distribution for the EM algorithm. Default value is 1.
`mu`	A positive double, the initial value of the parameter `mu` in the negative binomial distribution for the EM algorithm. Default value is 0.5.

Details

The statistical assumption is that for each species the number of individuals in a sample follows a Poisson distribution. The Poisson rate `lambda` are numbers generated from a gamma distribution. So the random variable `X`, which is the number of species represented x (x > 0) times in the sample, follows a zero-truncated negative binomial distribution. The unknown parameters are estimated by the function `preseqR.ztnb.em` based on the initial sample. Using the estimated distribution, we calculate the expected number of species represented at least r times in a random sample. Details of the estimation procedure can be found in the supplement of Daley T. and Smith AD. (2013).

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Daley, T., & Smith, A. D. (2013). Predicting the molecular complexity of sequencing libraries. Nature methods, 10(4), 325-327.

Deng C, Daley T & Smith AD (2015). Applications of species accumulation curves in large-scale biological data analysis. Quantitative Biology, 3(3), 135-144.

See Also

preseqR.ztnb.em

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for SAC
ztnb1 <- ztnb.rSAC(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
ztnb1(c(10, 20))

## construct the estimator for r-SAC
ztnb2 <- ztnb.rSAC(FisherButterfly, r=2)
## The number of species represented at least twice in a sample,
## when the sample size is 50 or 100 times of the initial sample
ztnb2(c(50, 100))
```

---

ztp.rSAC *ZTP estimator*

Description

ztp.rSAC predicts the expected number of species represented at least r times in a random sample, based on the initial sample.

Usage

```
ztp.rSAC(n, r=1)
```

Arguments

n	A two-column matrix. The first column is the frequency $j = 1, 2, \ldots$; and the second column is N_j, the number of species with each species represented exactly j times in the initial sample. The first column must be sorted in an ascending order.
r	A positive integer. Default is 1.

Details

The statistical assumption is that for each species the number of individuals in a sample follows a Poisson distribution. The Poisson rate `lambda` is the same among all species. So the random variable `X`, which is the number of species represented x (x > 0) times, follows a zero-truncated Poisson distribution. The unknown parameters are estimated by Cohen (1960). Based on the estimated distribution, we calculate the expected number of species in a random sample.

Value

The estimator for the r-SAC. The input of the estimator is a vector of sampling efforts t, i.e., the relative sample sizes comparing with the initial sample. For example, $t = 2$ means a random sample that is twice the size of the initial sample.

Author(s)

Chao Deng

References

Cohen, A. Clifford. "Estimating the parameter in a conditional Poisson distribution." Biometrics 16, no. 2 (1960): 203-211.

Examples

```
## load library
library(preseqR)

## import data
data(FisherButterfly)

## construct the estimator for SAC
ztp1 <- ztp.rSAC(FisherButterfly, r=1)
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
ztp1(c(10, 20))

## construct the estimator for r-SAC
ztp2 <- ztp.rSAC(FisherButterfly, r=2)
```

```
## The number of species represented at least once in a sample,
## when the sample size is 10 or 20 times of the initial sample
ztp2(c(50, 100))
```

Index