Package 'SCOPE'

February 14, 2019

Type Package

Title A normalization and copy number estimation method for single-cell DNA sequencing

Version 0.0.1

Author Rujin Wang, Danyu Lin, Yuchaojiang

Maintainer Rujin Wang <rujin@email.unc.edu>

Description Whole genome single-cell DNA sequencing (scDNA-seq) enables characterization of copy number profiles at the cellular level. This circumvents the averaging effects associated with bulk-tissue sequencing and has increased resolution yet decreased ambiguity in deconvolving cancer subclones and elucidating cancer evolutionary history. ScDNA-seq data is, however, sparse, noisy, and highly variable even within a homogeneous cell population, due to the biases and artifacts that are introduced during the library preparation and sequencing procedure. Here, we propose SCOPE, a normalization and copy number estimation method for scDNA-seq data. The distinguishing features of SCOPE include: (i) utilization of cell-specific Gini coefficients for quality controls and for identification of normal/diploid cells, which are further used as negative control samples in a Poisson latent factor model for normalization; (ii) modeling of GC content bias using an expectation-maximization algorithm embedded in the Poisson generalized linear models, which accounts for the different copy number states along the genome; (iii) a cross-sample iterative segmentation procedure to identify breakpoints that are shared across cells from the same genetic background. We evaluate performance of SCOPE on real scDNA-seq data sets from cancer genomic studies. Compared to existing methods, SCOPE more accurately estimates subclonal copy number aberrations and is shown to have higher correlation with array-based copy number profiles of purified bulk samples from the same patient. We further demonstrate SCOPE on three recently released data sets using the 10X Genomics single-cell CNV pipeline and show that it can reliably recover 1% of the cancer cells from a background of normal.

Depends R (>= 3.4.3)

Imports CODEX2, Rsamtools, GenomicRanges, IRanges, stats, GenomeInfoDb, BSgenome.Hsapiens.UCSC.hg19, graphics, utils

License GPL-2

LazyData true

RoxygenNote 6.1.1

Encoding UTF-8

R topics documented:

Estep.Pois 2

getcoverage.scDNA . . . 3
getfGC.Pois.random . . . 3
getmapp . . . 4
getsampQC . . . 5
logsumexp . . . 6
mapp_hg19 . . . 6
mapp_hg38 . . . 7
Mstep.Pois . . . 7
multi_run.Pois . . . 8
normalize_codex2_ns_noK . . . 9
normalize_codex2_ns_noK_EM_random . . . 10
segmentCBScs . . . 11

Index **12**

`Estep.Pois` *Expectation step for SCOPE normalization*

Description

Expectation step of GC content fitting for SCOPE normalization.

Usage

```
Estep.Pois(Yj, Nj, betatemp, fGCi, vec_pi, min.prop)
```

Arguments

`Yj`	read depth vector for each single cell
`Nj`	a numerirc value of total number of reads per cell
`betatemp`	bin-specific bias estimated using negative control samples
`fGCi`	estimated vector of GC content bias fitting from the M-step `Mstep.Pois`
`vec_pi`	vector of incident rates for CNV events that span bin *i* from the M-step `Mstep.Pois`
`min.prop`	the minmimum of mixture proportion for candidate CNV groups, which serves as a stopping metric for EM algorithm

Value

A list with components

`Z`	Matrix of optimized missing data to be fed into M-step `Mstep.Pois`
`obs_LL`	Observed log-Likelihood
`keep_going`	If normal proportion or a mixture proportion reached the minimum value, `keep_going = FALSE`, and output `NULL`, aka this case won't be involved in optimal number of CNV group selection based upon BIC.

Author(s)

Rujin Wang <rujin@email.unc.edu>

getcoverage.scDNA *Get read coverage from single-cell DNA sequencing*

Description

Get read coverage for each genomic bin across all single cells from scDNA-seq.

Usage

```
getcoverage.scDNA(bambedObj, mapqthres, mask.ref, seq)
```

Arguments

bambedObj	object returned from getbambed
mapqthres	mapping quality threshold of reads
mask.ref	a GRanges object indicating bad regions/bins, such as segmental duplication regions and gaps near telomeres/centromeres, which need to be masked prior to getting coverage
seq	the sequencing method to be used. This should be either "paired-end" or "single-end"

Value

Y	Read depth matrix

Author(s)

Rujin Wang <rujin@email.unc.edu>

getfGC.Pois.random *Get GC content bias fitting using Expectation-Maximization algorithm*

Description

Fit a Poisson generalized linear model to normalize the raw read depth data from single-cell DNA sequencing, without latent factors under the case-control setting. SCOPE implements an EM algorithm with random initialization to unmask null regions. This is a bulit-in function within multi_run.Pois. We don't recommend running this function independently.

Usage

```
getfGC.Pois.random(gcfitj, gctemp, Yj, Nj, betatemp, numGroup,
  verbose.plot = FALSE, gctemp.keep, min.prop)
```

Arguments

gcfitj	vector of bin-specific GC content biases for each single cell
gctemp	a vector giving values of bin-specific GC content after gcfitj outlier removal (for better fitting)
Yj	read depth vector for each single cell
Nj	a numerirc value of total number of reads per cell
betatemp	bin-specific bias estimated using negative control samples
numGroup	a vector of integers indicating number of CNV groups. Use BIC to select optimal number of CNV groups. If numGroup = 1, assume all reads are from normal regions so that EM algorithm is not implemented. Otherwise, we assume there is always a CNV group of heterozygous deletion and a group of null region. The rest groups are representative of different duplication states.
verbose.plot	logical, whether to plot GC content bias fitting results using EM and the choice of optimal CNV group. Default is FALSE
gctemp.keep	a vector giving values of bin-specific GC content without gcfitj outlier removal. For better fitting, extreme GC content biases need to be excluded from EM input. But a post hoc admission of all bins is necessary.
min.prop	the minmimum of mixture proportion for candidate CNV groups, which serves as a stopping metric for EM algorithm

Value

A list with components

BIC	BIC for CNV group selection
logL	Observed log-Likelihood
fGCi	EM estimated vector of GC content bias fitting given gctemp
Z	Matrix of optimized missing data using EM algorithm
vec_pi	Vector of EM estimated incident rate for CNV events that span bin *i*
K	Choice of optimal CNV group number based upon BIC
fGCi.keep	EM estimated vector of GC content bias fitting given gctemp.keep

Author(s)

Rujin Wang <rujin@email.unc.edu>

---

getmapp *Compute mappability*

Description

Compute mappability for each bin. Note that scDNA sequencing is whole-genome amplification and the mappability score is essential to determine variable binning method. Mappability track for 100-mers on the GRCh37/hg19 human reference genome from ENCODE is pre-saved. Compute the mean of mappability scores that overlapped reads map to bins, weighted by the width of mappability tracks on the genome reference. Use liftOver utility to calculate mappability for hg38, which is pre-saved as well.

Usage

```
getmapp(ref, genome = NULL)
```

Arguments

ref	GRanges object returned from getbambed
genome	by default, genome = BSgenome.Hsapiens.UCSC.hg19. To calculate mappability for hg38, specify genome = BSgenome.Hsapiens.UCSC.hg38

Value

mapp	Vector of mappability for each bin/target

Author(s)

Rujin Wang <rujin@email.unc.edu>

---

getsampQC *Get QC metrics for single cells*

Description

Perform QC step on single cells.

Usage

```
getsampQC(bambedObj)
```

Arguments

bambedObj	object returned from getbambed

Value

QCmetric	A matrix containing total number/proportion of reads, total number/proportion of mapped reads, total number/proportion of mapped non-duplicate reads, and number/proportion of reads with mapping quality greater than 20

Author(s)

Rujin Wang <rujin@email.unc.edu>

logsumexp *Logarithm of summation of exponentials*

Description

Computes the logarithm of summation of numeric exponentials.

Usage

```
logsumexp(xx)
```

Arguments

xx	a numeric value or vector

Value

A numeric value giving natural logarithm of summation of exponentials

Author(s)

Rujin Wang <rujin@email.unc.edu>

mapp_hg19 *GRanges with mappability scores for hg19*

Description

GRanges object specifying target positions with mappabilities across the whole genome.

Usage

```
data(mapp_hg19)
```

Format

A GRanges object with 21591667 ranges and 1 metadata column of mappability scores

Details

GRanges of mappability track for 100-mers on the GRCh37/hg19 human reference genome from ENCODE.

Value

GRanges object with mappabilities for hg19

Author(s)

Rujin Wang <rujin@email.unc.edu>

References

http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgFileUi?db=hg19&g=wgEncodeMapability

---

mapp_hg38 *GRanges with mappability scores for hg38*

Description

GRanges object specifying target positions with mappabilities across the whole genome.

Usage

```
data(mapp_hg38)
```

Format

A GRanges object with 21584930 ranges and 1 metadata column of mappability scores

Details

Use liftOver utility to convert hg19 coordinates to hg38

Value

GRanges object with mappabilities for hg38

Author(s)

Rujin Wang <rujin@email.unc.edu>

References

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/

Mstep.Pois *Maximization step for SCOPE normalization*

Description

Maximization step of GC content fitting for SCOPE normalization.

Usage

```
Mstep.Pois(Z, gcfitj, gctemp)
```

Arguments

Z	matrix of optimized missing data from the E-step Estep.Pois
gcfitj	vector of bin-specific GC content biases for each single cell
gctemp	a vector giving values of GC content for each bin after quality control

Value

A list with components

vec_pi	Vector of incident rate for CNV events that span bin *i*, which is to be fed into E-step Estep.Pois
fGCi	Estimated vector of GC content bias fitting

Author(s)

Rujin Wang <rujin@email.unc.edu>

---

multi_run.Pois *Get the optimal GC content bias fitting using Expectation-Maximization algorithm*

Description

SCOPE implements an EM algorithm with random initialization to unmask null regions. Adopt BIC to choose optimal GC content bias fitting among multiple runs of getfGC.Pois.random. This is a bulit-in function within normalize_codex2_ns_noK_EM_random.

Usage

```
multi_run.Pois(gcfitj, gctemp, Yj, Nj, betatemp, numGroup, rerun,
  verbose.plot = FALSE, qc.thres = 5e-05, min.prop = 0.002)
```

Arguments

gcfitj	vector of bin-specific GC content biases for each single cell
gctemp	a vector giving values of bin-specific GC content after gcfitj outlier removal (for better fitting)
Yj	read depth vector for each single cell
Nj	a numerirc value of total number of reads per cell
betatemp	bin-specific bias estimated using negative control samples
numGroup	a vector of integers indicating number of CNV groups. Use BIC to select optimal number of CNV groups. If numGroup = 1, assume all reads are from normal regions so that EM algorithm is not implemented. Otherwise, we assume there is always a CNV group of heterozygous deletion and a group of null region. The rest groups are representative of different duplication states.
rerun	specify the number of running EM algorithm with random initialization
verbose.plot	logical, whether to plot the optimal GC content bias fitting results using EM and the choice of optimal CNV group. Default is FALSE
qc.thres	the lower bound of bin-specific GC content bias threshold
min.prop	the minmimum of mixture proportion for candidate CNV groups, which serves as a stopping metric for EM algorithm

Value

A list with components for the optimal GC content bias fitting using EM

BIC	BIC for CNV group selection
logL	Observed log-Likelihood
fGCi	EM estimated vector of GC content bias fitting given gctemp
Z	Matrix of optimized missing data using EM algorithm
vec_pi	Vector of EM estimated incident rate for CNV events that span bin *i*
K	Choice of optimal CNV group number based upon BIC
fGCi.keep	EM estimated vector of GC content bias fitting given gctemp.keep

Author(s)

Rujin Wang <rujin@email.unc.edu>

---

normalize_codex2_ns_noK

Normalization of read depth without latent factors under the case-control setting

---

Description

Assuming that all reads are from diploid regions, fit a Poisson generalized linear model to normalize the raw read depth data from single-cell DNA sequencing, without latent factors under the case-control setting.

Usage

```
normalize_codex2_ns_noK(Y_qc, gc_qc, K = 1, norm_index, N)
```

Arguments

Y_qc	read depth matrix after quality control
gc_qc	vector of GC content for each bin after quality control
K	default is K = 1
norm_index	indices of normal/diploid cells
N	library size factor, which is computed from the genome-wide read depth data

Value

A list with components

Yhat	A list of normalized read depth matrix
fGC.hat	A list of estimated GC content bias matrix
beta.hat	A list of estimated bin-specific bias vector

Author(s)

Rujin Wang <rujin@email.unc.edu>

normalize_codex2_ns_noK_EM_random

Normalization of read depth without latent factors using Expectation-Maximization algorithm under the case-control setting

Description

Fit a Poisson generalized linear model to normalize the raw read depth data from single-cell DNA sequencing, without latent factors under the case-control setting. Model GC content bias using an expectation-maximization algorithm, which accounts for the different copy number states.

Usage

```
normalize_codex2_ns_noK_EM_random(Y_qc, gc_qc, K = 1, norm_index, N,
  numGroup, qc.thres = 5e-05, min.prop = 0.002)
```

Arguments

Y_qc	read depth matrix after quality control
gc_qc	vector of GC content for each bin after quality control
K	default is K = 1
norm_index	indices of normal/diploid cells
N	library size factor, which is computed from the genome-wide read depth data
numGroup	a vector of integers indicating number of CNV groups. Use BIC to select optimal number of CNV groups. If numGroup = 1, assume all reads are from normal regions so that EM algorithm is not implemented. Otherwise, we assume there is always a CNV group of heterozygous deletion and a group of null region. The rest groups are representative of different duplication states.
qc.thres	the lower bound of bin-specific GC content bias threshold
min.prop	the minmimum of mixture proportion for candidate CNV groups, which serves as a stopping metric for EM algorithm

Value

A list with components

Yhat	A list of normalized read depth matrix with EM
fGC.hat	A list of EM estimated GC content bias matrix
beta.hat	A list of EM estimated bin-specific bias vector

Author(s)

Rujin Wang <rujin@email.unc.edu>

segmentCBScs *Cross-sample segmentation*

Description

SCOPE offers a cross-sample Poisson likelihood-based recursive segmentation, enabling shared breakpoints across cells from the same genetic background.

Usage

```
segmentCBScs(Y, Yhat, sampname, ref, lmax, mode, segment.CODEX2 = FALSE)
```

Arguments

Y	raw read depth matrix after quality control procedure
Yhat	normalized read depth matrix
sampname	vector of sample names
ref	GRanges object after quality control procedure
lmax	maximum CNV length in number of bins returned
mode	format of returned copy numbers, which can be either "integer" or "fraction". "integer" is recommended for scDNA-seq data.
segment.CODEX2	logical, whether to perform individual segmentation. Default is FALSE.

Value

A list with components

poolcall	Cross-sample CNV callings indicating shared breakpoints
finalcall	Final cross-sample segmented callset of CNVs with genotyping results
finalcall_CODEX2	Final individual segmented callset of CNVs with genotyping results
image.orig	A matrix giving logarithm of normalized z-scores
image.seg	A matrix of logarithm of estimated copy number over 2

Author(s)

Rujin Wang <rujin@email.unc.edu>

Index

∗Topic **datasets**
mapp_hg19, 6
mapp_hg38, 7

Estep.Pois, 2, *7*, *8*

getbambed, *3*, *5*
getcoverage.scDNA, 3
getfGC.Pois.random, 3, *8*
getmapp, 4
getsampQC, 5

logsumexp, 6

mapp_hg19, 6
mapp_hg38, 7
Mstep.Pois, *2*, 7
multi_run.Pois, *3*, 8

normalize_codex2_ns_noK, 9
normalize_codex2_ns_noK_EM_random, *8*, 10

segmentCBScs, 11