Package ‘hicrep’

October 17, 2017

Title Measuring the reproducibility of Hi-C data

Version 1.0.0

Description Hi-C is a powerful technology for studying genome-wide chromatin interactions. However, current methods for assessing Hi-C data reproducibility can produce misleading results because they ignore spatial features in Hi-C data, such as domain structure and distance-dependence. We present a novel reproducibility measure that systematically takes these features into consideration. This measure can assess pairwise differences between Hi-C matrices under a wide range of settings, and can be used to determine optimal sequencing depth. Compared to existing approaches, it consistently shows higher accuracy in distinguishing subtle differences in reproducibility and depicting interrelationships of cell lineages than existing approaches. This R package `hicrep` implements our approach.

biocViews Sequencing, HiC, QualityControl

Depends R (>= 3.4)

Imports stats

License GPL (>= 2.0)

Encoding UTF-8

LazyData true

RoxygenNote 5.0.1

Suggests knitr, rmarkdown, testthat

VignetteBuilder knitr

NeedsCompilation no

Author Tao Yang [aut, cre]

Maintainer Tao Yang <xadmyangt@gmail.com>

R topics documented:

`hicrep-package` *HiCRep pipeline calculates reproducibility of Hi-C intrachromosome data*

Description

The pipelne is a two-step method. The first step is to smooth the Hi-C matrix, and the #' second step is to calculate the stratum-adjusted correlation coefficient (scc). The method also provides the estimation of asymptotic standard deviation of scc.

Details

- Package: hicrep
- Type: Package
- Version: 0.99.6
- Date: 2017-2-5
- License: GPL-2
- LazyLoad: Yes

The main functions are `prep`, `get.scc` and `htrain`. The function `prep` will take the two replicates of $N * (3 + N)$ matrix format as input, and return the vectorized, smoothed or unsmoothed (when smoothing neighborhood size parameter h = 0) Hi-C data, which will subsequently used to compute stratum-adjusted correlation coefficients (scc). The function `get.scc` computes scc and its asymptotic standard deviation, and the function `htrain` estimates optimal smoothing neighborhood size from the input matrices.

Author(s)

Tao Yang Maintainer: Tao Yang <xadmyangt@gmail.com>

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)
data(HiCR2)

#Estimate the optimial smoothing neighborhood size parameter
h_hat <- htrain(HiCR1, HiCR2, 1000000, 5000000, 0:2)
h_hat <- 0
processed <- prep(HiCR1, HiCR2, 1000000, h_hat, 5000000)

scc.out <- get.scc(processed, 1000000, 5000000)
scc.out$scc
scc.out$std
```

---

depth.adj *Sequencing depth adjustment*

Description

Sequencing depth could be a confounding effect when measuring the reproducibility. This function will adjust sequencing depth of a given matrix to a specified total number of reads through random sampling.

Usage

```
depth.adj(d, size, resol, out = 0)
```

Arguments

`d`	a Hi-C matrix needed to be adjusted.
`size`	the size the total number one wants to adjust to.
`resol`	the resolution of the input matrix.
`out`	either 0 or 1. If it is 0, the function returns matrix format; if 1, it returns vection format.

Value

a matrix or vec which has the adjusted total number of reads.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)
#total number of reads
sum(HiCR1[,-c(1:3)])

#Adjust it to 200000 reads, output Hi-C matrix
HiC_R1_200k = depth.adj(HiCR1, 200000, 1000000, out = 0)
```

```
#check total number of reads after adjustment
sum(HiC_R1_200k[,-c(1:3)])

#output vector
HiC_R1_200k = depth.adj(HiCR1, 200000, 1000000, out = 1)
#check total number of reads after adjustment
sum(HiC_R1_200k[,3])
```

---

get.scc *calculate the stratum-adjusted correlation coefficient*

Description

calculate the stratum-adjusted correlation coefficient

Usage

```
get.scc(dat, resol, max)
```

Arguments

dat	A matrix of four columns. The first two are the mid-point coordinates of two interacting bin.
resol	An integer indicating the resolution of the Hi-C matrix.
max	An integer indicating the maximum distance of interaction that is considered.

Details

The function stratifies the Hi-C reads count according to their interacting distance, calculates the Pearson correlation coefficient for each stratum, then aggregrates them using a weighted average.

Value

A list of results including stratum-specific correlation coefficients, weights, stratum-adjusted correlation coefficient (scc), and the asymptotic standard deviation of scc.

- corr A vector that contains the stratum specific Pearson correlation coefficients.
- wei A vector that contains the weights for each stratum.
- scc Stratum-adjusted correlation coefficients.
- std The asymptotic standard deviation of scc.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)
data(HiCR2)
processed <- prep(HiCR1, HiCR2, 1000000, 0, 5000000)

scc.out = get.scc(processed, 1000000, 5000000)
scc.out$scc
scc.out$std
```

HiCR1 *HiCR1*

Description

A Hi-C matrix in the format of $N*(3+N)$, the additional first three columns contains the chromosome information and coordinates of interaction bins. First is chromosome name, the second and third are the mid-points of the contacting bins.

Usage

```
HiCR1
```

Format

An object of class `data.frame` with 52 rows and 55 columns.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

HiCR2 *HiCR2*

Description

A Hi-C matrix in the format of $N*(3+N)$, the additional first three columns contains the chromosome information and coordinates of interaction bins. First is chromosome name, the second and third are the mid-points of the contacting bins.

Usage

```
HiCR2
```

Format

An object of class `data.frame` with 52 rows and 55 columns.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

`htrain` *Train the smoothing parameter (neighborhood size).*

Description

Train the smoothing parameter (neighborhood size).

Usage

```
htrain(R1, R2, resol, max, range)
```

Arguments

`R1`	A Hi-C intra-chromosome matrix.
`R2`	The other intra-chromosome matrix to compare with.
`resol`	An integer indicating the resolution of the Hi-C matrix.
`max`	An integer indicating the maximum distance of interaction that is considered.
`range`	A vector of consecutive integers from which the optimal smoothing parameter is searched, starting from zero (i.g., 0:10).

Details

A fraction (10%) of data are first randomly sampled, then the scc for the sampled data is computed at a series of smoothing parameterts in the ascending order. The samllest `h` at which the increment of scc is less than 0.01 is saved. This procedure is repeat 10 times, and the mode of the 10 `h`'s is outputed as the estimated optimal neighborhood size.

Value

a integer estimated to be the optimal smoothing parameter.

References

HiCRep: HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)
data(HiCR2)
h_hat <- htrain(HiCR1, HiCR2, 1000000, 5000000, 0:2)
```

MatToVec *Convert the HiC matrix format to vector format*

Description

The matrix format is the standard input for the HiCRep reproducibility analysis. It has the dimension of $N * (3 + N)$. The additional first three columns are chromosome name, and mid-point coordinates of two contacting bins. The converted format has three columns. The first two columns are mid-point coordinates of two contacting bins, and the third column is the reads number in each bin.

Usage

```
MatToVec(dat)
```

Arguments

dat	a Hi-C intra-chromosome matrix in the format of $N * N$ (No chromsome name and coordinates columns).

Value

a vectorized Hi-C data. The first two columns are mid-point coordinates of the two contacting bins. The third column is read numbers of the contacts.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)

#re-format the row and column names
resol <- 1000000
ref_Rep1 <- HiCR1[,-c(1,2,3)]
rownames(ref_Rep1) = colnames(ref_Rep1) = HiCR1[,3]-resol/2

vec_HiC_R1 <- MatToVec(ref_Rep1)
head(vec_HiC_R1)
```

prep *Pre-processing the Hi-C matrices*

Description

Format pairs of Hi-C matrices, smooth the matrices with matrix resolution, and maximum distance of interaction considering specified by user, filter out the bins that has no reads in both replciates.

Usage

```
prep(R1, R2, resol, h, max)
```

Arguments

R1	a Hi-C intra-chromosome matrix.
R2	the other intra-chromosome matrix to compare with.
resol	an integer indicating the resolution of the Hi-C matrix.
h	an integer indicating the size of the smoothing neighborhood.
max	an integer indicating the maximum distance of interaction that is considered.

Value

a smoothed (or not when resol = 0), zero-filtered and vectorized Hi-C data. The first two columns are bin start and bin ends, and the last two columns are reads number if replicate 1 and replicate 2 respectively.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)
data(HiCR2)
processed <- prep(HiCR1, HiCR2, 1000000, 0, 5000000)
head(processed)
```

smoothMat *smooth the Hi-C matrix with mean filter*

Description

smooth the Hi-C matrix with mean filter

Usage

```
smoothMat(dat, h)
```

Arguments

`dat`	A $N * N$ Hi-C intra-chromosome matrix.
`h`	The neighborhood size parameter. It is the distance of smoothing target bin to the boundary of the neighborhood in the unit of resolution.

Details

Given a Hi-C $N * N$ matrix, the algorithm scans through each data points (i, j), idnetifies points within its neighborhood of size h (max distance to (i, j) is $h * resolution$), and calculates the mean. The mean is subsequently used as the smoothed value of the point (i, j).

Value

a smoothed (or not when resol = 0), zero-filtered and vectorized Hi-C data.

References

HiCRep: assessing the reproducibility of Hi-C data using a stratum-adjusted correlation coefficient. Tao Yang, Feipeng Zhang, Galip Gurkan Yardimci, Ross C Hardison, William Stafford Noble, Feng Yue, Qunhua Li. bioRxiv 101386; doi: https://doi.org/10.1101/101386.

Examples

```
data(HiCR1)

#re-format the row and column names
resol <- 1000000
ref_Rep1 <- HiCR1[,-c(1,2,3)]
rownames(ref_Rep1) = colnames(ref_Rep1) = HiCR1[,3]-resol/2

smt_HiC_R1 <- smoothMat(ref_Rep1, 1)
dim(smt_HiC_R1)
smt_HiC_R1[1:5,1:5]
```

`vstran` *Variance stablization transformation*

Description

The function finds the rank of the data and rescale the ranks in the range of (0, 1).

Usage

```
vstran(d)
```

Arguments

`d`	a matrix or data.frame that has two columns, each column is the Hi-C read counts data in a replicate.

Details

In Hi-C data, the read counts for contacts with short interaction distances have a much larger dynamic range than those with long interaction distances. To mitigate this difference, we rank the contact counts in each stratum separately, and then normalize the ranks by the total number of observations in each stratum, such that all strata share a similar dynamic range.

Value

a matrix of two columns, each represents a transformed read counts of a replicate.

Examples

```
data(HiCR1)
HiC_R1_vs <- vstran(HiCR1)
head(HiC_R1_vs)
```

Index