Package 'HiCblock'

April 19, 2018

Type Package

Title Systematic Analysis of Architectural Proteins and Functional Elements in Blocking Long-Range Contacts Between Loci

Version 1.4

Depends R (>= 3.4.0)

Imports IRanges, GenomeInfoDb, methods, GenomicRanges, Matrix, glmnet, rtracklayer, HiTC, MASS, S4Vectors

Date 2018-04-18

Author Raphael Mourad

Maintainer Raphael Mourad <raphael.mourad@ibcg.biotoul.fr>

Description Here we propose a model to systematically analyze the roles of architectural proteins and functional elements in blocking long-range contacts between loci. The proposed model does not rely on topologically associating domain (TAD) mapping from Hi-C data. Instead of testing the enrichment or influence of protein binding at TAD borders, the model directly estimates the blocking effect of proteins on long-range contacts between flanking loci, making the model intuitive and biologically meaningful.

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2018-04-18 22:46:01 UTC

R topics documented:

HiCblock-package *Systematic Analysis of Architectural Proteins and Functional Elements in Blocking Long-Range Contacts Between Loci*

Description

Here we propose a model to systematically analyze the roles of architectural proteins and functional elements in blocking long-range contacts between loci. The proposed model does not rely on topologically associating domain (TAD) mapping from Hi-C data. Instead of testing the enrichment or influence of protein binding at TAD borders, the model directly estimates the blocking effect of proteins on long-range contacts between flanking loci, making the model intuitive and biologically meaningful.

Details

# To install dependencies:
source("https://bioconductor.org/biocLite.R")
biocLite("IRanges","GenomicRanges","GenomeInfoDb","rtracklayer","HiTC")
install.packages(c("methods","MASS","Matrix","glmnet","S4Vectors"))

# To install HiCblock package:
install.packages("HiCblock")

To use the package, there are two steps:
- To assess the blocking effects of features such as architectural proteins, one should first annotate Hi-C matrix with bias data (GC-content, mappability, fragment length) and with feature data (for instance, ChIP-seq data) with the function HiCblockProcData(). Note that if the Hi-C matrix has already been corrected for biases, then no bias data need to be used.
- Then, one should compute the negative binomial or Poisson lasso regression with the function HiCblockModel() using the output from HiCblockProcData().
The blocking effect of a protein (or motif) is the associated beta from the regression.

Choice between Poisson lasso over binomial negative regressions:
Because of Hi-C count overdispersion, we used negative binomial regression as the most appropriate specification of the generalized linear model. However, Poisson regression with lasso shrinkage can also be used. We believe that the choice between both depends mainly on the number of variables to analyze. On the one hand, if there are a few candidate variables (less than 10), it is interesting to estimate beta parameters together with corresponding p-values to assess significance using negative binomial regression. On the other hand, if there are a large number of variables (10 or more), it is more convenient to use Poisson lasso regression in order to select the key variables and to account for correlations among the variables (frequent in ChIP-seq and motif occurrence data).

Author(s)

Raphael Mourad

Maintainer: Raphael Mourad, raphael.mourad@ibcg.biotoul.fr

References

Raphael Mourad and Olivier Cuvier. TAD-free analysis of architectural proteins and insulators. Nucleic Acids Research, Volume 46, Issue 5, 16 March 2018, Pages e27.

Examples

```
# Load data
# The Hi-C matrix is at 20kb resolution (low resolution only for example)
data(dataExample)
genomicFeatureList.GR=dataExample$GenomicFeatureList.GR
annotNames=dataExample$AnnotNames
HTCList=dataExample$HTC
distInter=c(100e3,140e3)
IBP=c("BEAF32","dCTCF","dTFIIIC","GAF","SuHw")

# Annotate Hi-C data with genomic features
HRPD=HiCblockProcData(genomicFeatureList.GR, annotNames, HTCList, distInter,verbose=TRUE)

# Model 1
modelBlock1=as.formula(paste0("Count~logDist+len+GC+map+I(-BEAF32)"))
HRM_Block1=HiCblockModel(HRPD,modelBlock1,"BEAF32",regressionMode="NB")
print(HRM_Block1)

# Output from model 1
# Blocking effect (beta) of BEAF-32 is 0.21
#Call:
#glm.nb(formula = model, data = dataGLM, init.theta = 2.956475677,
#    link = log)

#Deviance Residuals:
#    Min       1Q   Median       3Q      Max
#-3.9346  -0.9018  -0.2606   0.4158   5.0515

#Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
#(Intercept) -4.420963   1.102393  -4.010 6.06e-05 ***
#logDist     -1.226667   0.073070 -16.788  < 2e-16 ***
#len          0.662864   0.027635  23.987  < 2e-16 ***
#GC          -0.627353   0.087941  -7.134 9.76e-13 ***
#map          1.258398   0.054150  23.239  < 2e-16 ***
#I(-BEAF32)   0.206394   0.007735  26.683  < 2e-16 ***
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#(Dispersion parameter for Negative Binomial(2.9565) family taken to be 1)

#    Null deviance: 5377.2  on 3375  degrees of freedom
#Residual deviance: 3575.1  on 3370  degrees of freedom
#AIC: 46123

#Number of Fisher Scoring iterations: 1
```

```
#              Theta:  2.9565
#          Std. Err.:  0.0693

# 2 x log-likelihood:  -46109.2770



# Model 2
# Blocking effect (beta) of BEAF-32 is 0.18
vars2=paste(paste0("I(-",IBP,")"),collapse='+')
modelBlock2=as.formula(paste0("Count~logDist+len+GC+map+",vars2))
HRM_Block2=HiCblockModel(HRPD,modelBlock2,IBP,regressionMode="NB")
print(HRM_Block2)

# Output from model 2
#Call:
#glm.nb(formula = model, data = dataGLM, init.theta = 3.138426428,
#    link = log)

#Deviance Residuals:
#    Min       1Q   Median       3Q      Max
#-3.7727  -0.8815  -0.2809   0.4282   4.3644

#Coefficients:
#             Estimate Std. Error z value Pr(>|z|)
#(Intercept) -4.982498   1.074279  -4.638 3.52e-06 ***
#logDist     -1.236704   0.070946 -17.432  < 2e-16 ***
#len          0.685146   0.027209  25.181  < 2e-16 ***
#GC          -0.560153   0.086482  -6.477 9.35e-11 ***
#map          1.310488   0.053692  24.407  < 2e-16 ***
#I(-BEAF32)   0.176687   0.008283  21.332  < 2e-16 ***
#I(-dCTCF)    0.028798   0.013665   2.107  0.03508 *
#I(-dTFIIIC)  0.621412   0.077602   8.008 1.17e-15 ***
#I(-GAF)      0.047998   0.007878   6.092 1.11e-09 ***
#I(-SuHw)     0.072596   0.026326   2.758  0.00582 **
#---
#Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#(Dispersion parameter for Negative Binomial(3.1384) family taken to be 1)

#    Null deviance: 5702.9  on 3375  degrees of freedom
#Residual deviance: 3564.8  on 3366  degrees of freedom
#AIC: 45911

#Number of Fisher Scoring iterations: 1


#              Theta:  3.1384
#          Std. Err.:  0.0738

# 2 x log-likelihood:  -45889.1960
```

```
# Model 3
# Blocking effect (beta) of BEAF-32 is 0.22
HRM_Block3=HiCblockModel(HRPD,NULL,IBP,regressionMode="PoissonLasso")
print(HRM_Block3)

# Output from model 3
#        Variable Coefficient
#logDist  logDist    -1.04649
#len          len     0.56508
#GC            GC    -0.43621
#map          map     0.95623
#BEAF32    BEAF32     0.22077
#dCTCF      dCTCF     0.01680
#dTFIIIC  dTFIIIC     0.73750
#GAF          GAF     0.03878
#SuHw        SuHw     0.12617
```

`annotateHiCBin` *Annotate blocking regions with ChIP-seq data.*

Description

This is an internal function (should not be used). Function to compute the average of the coverage of ChIP-seq peaks within blocking regions.

Usage

```
annotateHiCBin(HiC_bin.GR, GenomicFeature.GR)
```

Arguments

`HiC_bin.GR`	A GRanges object containing the Hi-C blocking regions.
`GenomicFeature.GR`	A GRanges object containing the genomic feature intervals with fold-enrichments (score).

Value

A vector object is returned.

Author(s)

Raphael Mourad

createHiCDataset *Create a dataset for further processing with the R package.*

Description

This is an internal function (should not be used).

Usage

```
createHiCDataset(HTCL, distInter = NULL, verbose = F)
```

Arguments

HTCL	An HTClist object from HiTC R package.
distInter	A numeric vector of two values containing the distance interval used for Hi-C data analysis.
verbose	If verbose is True, then the different processing steps will be displayed.

Value

Return a list of objects.

Author(s)

Raphael Mourad

dataExample *Hi-C matrix and ChIP-seq peaks.*

Description

This dataExample object comprises four objects: GenomicFeatureList.GR, AnnotNames, SeqInfoChr and HTCL. All data are from Drosophila melanogaster Kc167 cells. Only data for chromosome 2L are provided.

GenomicFeatureList.GR is a list of GRanges objects, one for each genomic feature to analyze. ChIP-seq data (object GenomicFeatureList.GR) were downloaded from Gene Expression Omnibus (GEO) accessions GSE30740, GSE42085 and GSE54529. AnnotNames is a vector of the genomic feature names. SeqInfoChr is a Seqinfo object that contains chromosome 2L information.

HTCL is a HTClist object (HiTC R package) that contains a 20 kb Hi-C contact matrix Hi-C processed from Gene Expression Omnibus (GEO) accession GSE62904. This is a low resolution contact matrix just provided as an example.

Usage

```
data(dataExample)
```

Author(s)

Raphael Mourad

References

Kevin Van Bortle, Michael H. Nichols, Li Li, Chin-Tong Ong, Naomi Takenaka, Zhaohui S. Qin, and Victor G. Corces. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biology, 15(5):R82+, June 2014.

Li Li, Xiaowen Lyu, Chunhui Hou, Naomi Takenaka, Huy Q. Nguyen, Chin-Tong Ong, Caelin Cubenas-Potts, Ming Hu, Elissa P. Lei, Giovanni Bosco, Zhaohui S. Qin and Victor G. Corces. Widespread rearrangement of 3D chromatin organization underlies Polycomb-mediated stress-induced silencing. Molecular Cell, 15:S1097-2765, March 2015.

HiCblockModel *The model.*

Description

The main function to compute the generalized linear model.

Usage

```
HiCblockModel(hrpd, model, facBlock, regressionMode = "NB", scale = F, includeBias = T,
sampleSize = NULL, distInter = NULL)
```

Arguments

hrpd	Preprocessed data. It includes Hi-C interaction pairs that have been annotated with genomic feature information. To compute this object, you should use the function HiCglmiProcData.
model	A formula object (stats R package). See Section HiCglmi-package, for examples. If regressionMode="PoissonLasso", then the variable model should be set to NULL (i.e. model=NULL).
facBlock	A vector of character describing the genomic features. For instance, facModel="BEAF32".
regressionMode	If regressionMode="Poisson", a Poisson regression. If regressionMode="QP", a quasi Poisson regression. If regressionMode="NB", a negative binomial regression. If regressionMode="PoissonLasso", a Poisson lasso regression.
scale	If scale=TRUE, blocking variables are standardized.
includeBias	If includeBias=True, then GC, mappability and fragment length biases are included in the regression (default mode). If includeBias=False, biases are not included in the regression. You can used this option if you are using Hi-C matrix that had been previously corrected for biases (not recommended).
sampleSize	Optional. A numerical value for subsampling Hi-C data.
distInter	Optional. A vector of two numerical values to set a distance range for Hi-C analysis.

Value

If regressionMode is "Poisson", "QP" or "NB", then a summary(glm) object is returned.

Author(s)

Raphael Mourad

References

Submitted.

`HiCblockProcData` *Process data for further generalized linear regression.*

Description

This function is used to annotate the blocking regions, i.e. in-between bin pairs, with genomic feature information such as ChIP-seq peak intervals or DNA motif occurrence intervals.

Usage

```
HiCblockProcData(genomicFeatureList.GR, annotNames, HTCList, distInter,
 overlapmode="signal", verbose = F, includeBias = T)
```

Arguments

`genomicFeatureList.GR`	A list of GRanges objects. Each GRanges object has been built from coordinate data using readGFBed function (for instance ChIP-seq peak coordinates or DNA motif occurrence coordinates).
`annotNames`	A character vector defining the name of each genomic feature. Names should not comprise any special character such as ":/+-*^,;!?" because an internal R formula object is created inside the function.
`HTCList`	A HTClist object (HiTC R package). This object contains Hi-C intrachromosomal matrices.
`distInter`	Optional. A vector of two numerical values to set a distance range for Hi-C analysis.
`overlapmode`	If ChIP-seq peaks are used, overlapmode should be "signal" to compute the average of coverage of the peak fold-enrichment within the blocking region, divided by the length of the blocking region. If DNA motif occurrences are used, overlapmode should be "occurrence" to compute the number of occurrences of DNA motifs within the blocking region, divided by the length of the blocking region.
`verbose`	If true, verbose is output while the program is running.
`includeBias`	If True, biases are included. If False, biases are not included.

Value

An object further used for HiCblockModel function.

Author(s)

Raphael Mourad

`testDistancePowerLawDistrib` *Check log-log linear relation between distance and Hi-C count.*

Description

Function to check the log-log linear relation between distance and Hi-C count. At a particular distance range, the log-log function should be linear. The generalized linear regression with interactions (GLMI) can be used only if this log-log linear relation holds. One could consider the log-log linear relation holds if R2>0.95.

Usage

```
testDistancePowerLawDistrib(HTCL, distInter)
```

Arguments

`HTCL`	A HTClist object (HiTC R package). This is an object to store the Hi-C data.
`distInter`	Optional. A vector of two numerical values to set a distance range for Hi-C analysis.

Value

A list containing the log-log regression information.

Author(s)

Raphael Mourad

Index