Package 'TADreg'

May 20, 2021

Type Package

Title Versatile Regression Framework for TAD Analysis and Prediction.

Version 1.0

Depends R (>= 3.6.3)

Imports BSgenome.Mmusculus.UCSC.mm10, rtracklayer, GenomicRanges, glmnet, HiTC, hicrep, Matrix, glmnet, data.table, mgcv, L0Learn, doMC, GenomeInfoDb, IRanges, MASS, S4Vectors

Date 2021-05-19

Author Raphael Mourad

Maintainer Raphael Mourad <raphael.mourad@univ-tlse3.fr>

Description Here, we propose a versatile regression framework which not only identifies TADs in a fast and accurare manner, but also detects differential TAD borders across conditions for which few methods exist, and predicts 3D genome reorganizarion after chromosomal rearrangement. Moreover, the framework is biologically meaningful, has an intuitive interpretation and is easy to visualize.

License MIT License

NeedsCompilation no

R topics documented:

TADreg-package *Versatile Regression Framework for TAD Analysis and Prediction.*

Description

Here, we propose a versatile regression framework which not only identifies TADs in a fast and accurare manner, but also detects differential TAD borders across conditions for which few methods exist, and predicts 3D genome reorganizarion after chromosomal rearrangement. Moreover, the framework is biologically meaningful, has an intuitive interpretation and is easy to visualize.

Details

The DESCRIPTION file:

Package:	TADreg
Type:	Package
Title:	Versatile Regression Framework for TAD Analysis and Prediction.
Version:	1.0
Depends:	R (>= 3.6.3)
Imports:	BSgenome.Mmusculus.UCSC.mm10, rtracklayer, GenomicRanges, glmnet, HiTC, hicrep, Matrix, glmnet, data
Date:	2021-05-19
Author:	Raphael Mourad
Maintainer:	Raphael Mourad <raphael.mourad@univ-tlse3.fr>
Description:	Here, we propose a versatile regression framework which not only identifies TADs in a fast and accurare manne
License:	MIT License

Index of help topics:

`DIM`	`Differential Insulation Model (DIM)`
`HTCfromCHICdata`	`Function to read capture Hi-C data from Simona Bianco et al. Nat Genet 2018.`
`HTCfromJuicerDump`	`Function to read Hi-C data from Juicebox dump.`
`PIM`	`Prediction Insulation Model (PIM)`
`SIM`	`Sparse Insulation Model (SIM)`
`TADreg-package`	`Versatile Regression Framework for TAD Analysis and Prediction.`
`compSCC`	`Simple function to compute the stratum adjusted correlation (SCC) between two Hi-C matrices.`

To use TADreg, see the R Markdown html file which shows examples to run the different functions (SIM, DIM and PIM) from the package.

Author(s)

Raphael Mourad

Maintainer: Raphael Mourad <raphael.mourad@univ-tlse3.fr>

compSCC *Simple function to compute the stratum adjusted correlation (SCC) between two Hi-C matrices.*

Description

It is useful for benchmarking to compare the observed Hi-C matrix after chromosomal rearrangement and the predicted Hi-C matrix computed using PIM function. An SCC close to one means that predictions of rearranged 3D genome are accurate compared to observed Hi-C data from rearranged 3D genome. An SCC close to zero means the predictions are very inaccurate. Compared to the classical Pearson or Spearson correlation coefficients, SCC removes the distance effect from the Hi-C matrices, allowing to focus on the biological variability, here TADs, sub-TADs, hierarchies of TADs, loops, etc.

Usage

```
compSCC(HTC1, HTC2)
```

Arguments

`HTC1`	Observed Hi-C matrix, for a particular chromosome. It should be stored as an HTCexp object from HiTC R package.
`HTC2`	Predicted Hi-C matrix from PIM function, for a particular chromosome. It should be stored as an HTCexp object from HiTC R package.

Value

The stratum adjusted correlation value.

Author(s)

Raphael Mourad

DIM *Differential Insulation Model (DIM)*

Description

Differential Insulation Model (DIM) is a regression model used to identify differential TAD borders between two different Hi-C experiment matrices (e.g. between two conditions).

Usage

```
DIM(HTC1, HTC2, distMax = NULL, analysis = "border", overlap = 1)
```

Arguments

`HTC1` Hi-C matrix from the first condition, for a particular chromosome. It should be stored as an HTCexp object from HiTC R package.

`HTC2` Hi-C matrix from the second condition, for the same particular chromosome. It should be stored as an HTCexp object from HiTC R package.

`distMax` The maximal distance between two bins that is used to identify TADs. Usually, a distance equal to 10 bins is fine (default value). Setting a too high maximal distance will lead to computational burden.

`analysis` If analysis = "border" (default), differential TAD borders will be assessed. If analysis = "facilitator", differential TAD facilitators will be assessed.

`overlap` To prevent bin uncertainty between conditions (for instance bin i is identified as a border in condition 1 and bin i+1 is found as border in condition 2), only one bin among two consecutive bins was kept (overlap = 1). This avoids considering as differential border two consecutive bins from two conditions (likely a false positive).

Value

A GRanges object containing the bin genomic coordinates and the corresponding beta values. Bins with beta.diff < 0 correspond to TAD borders that are gained/reinforced in the second condition compared to the first condition, whereas bins with beta.diff > 0 correspond to TAD borders that are lost/weakened in the second condition compared to the first condition.

Author(s)

Raphael Mourad

`HTCfromCHICdata` *Function to read capture Hi-C data from Simona Bianco et al. Nat Genet 2018.*

Description

Function used to read capture Hi-C data from the article "Polymer physics predicts the effects of structural variants on chromatin architecture", Simona Bianco et al., Nature Genetics 2018.

Usage

```
HTCfromCHICdata(file_CHIC)
```

Arguments

`file_CHIC` The file path to the capture Hi-C data.

Value

A HTCexp object containing the capture Hi-C data.

Author(s)

Raphael Mourad

HTCfromJuicerDump *Function to read Hi-C data from Juicebox dump.*

Description

First, dump (extract) a Hi-C data matrix from a Juicebox .hic file using Juicebox dump tool (https://github.com/aidenlab/juicer Extraction). Then, use this function HTCfromJuicerDump to load the Hi-C matrix and convert it to HTCexp format.

Usage

```
HTCfromJuicerDump(file_juicer_dump, resolution, chr, assembly, sparse = T)
```

Arguments

`file_juicer_dump`
File path to the dumped Hi-C matrix.

`resolution` Hi-C data matrix resolution (bin size).

`chr` Chromosome.

`assembly` Genome assembly.

`sparse` Whether the dumped Hi-C matrix is in dense (sparse = F) or sparse format (sparse = F).

Value

The corresponded Hi-C matrix stored as an HTCexp object from HiTC R package.

PIM *Prediction Insulation Model (PIM)*

Description

Prediction Insulation Model (PIM) is a regression model used to predict 3D genome reorganization after a chromosomal rearrangement.

Usage

```
PIM(HTC, structVar.GR, typeVar, output = "asMutant", model = "glmlasso",
distMax = 5e+05, parallel = T, noise = 0)
```

Arguments

`HTC` A wild-type Hi-C matrix (no chromosomal rearrangement) for a particular chromosome. It should be stored as an HTCexp object from HiTC R package. Here to improve predictions, you might prefer to focus on a particular region surrounding the chromosomal rearrangement coordinates (for instance -/+10 bins around), and not to give as input the whole chromosomal matrix.

`structVar.GR` The chromosomal rearrangement coordinates, stored a GRanges object. It should be larger than a bin (ideally > 5 bins).

`typeVar` The type of chromosomal rearrangement : "deletion" or "inversion".

`output` The output matrix format. If output = "asMutant", it means that the predicted Hi-C matrix will have modified genomic bins due to the chromosomal rearrangement. If output = "asWT", it means that the predicted Hi-C matrix will have the same genomic bins as the original wild-type Hi-C matrix.

`model` By default, the model = "glmlasso" should be used. Other models produce less accurate predictions.

`distMax` The maximal distance between two bins that is used to identify TADs. Here to improve predictions, if you used as suggested above only a Hi-C matrix corresponding to a particular region surrounding the chromosomal rearrangement, then it should be better to run computations on a large distance.

`parallel` If parallel = True, parallel computing will be done on multiple cores using the parallel R library.

`noise` By default, noise is not added to the predicted Hi-C matrix (noise = 0). But one can add some noise (noise > 0) to mimic experimental Hi-C data.

Value

A predicted Hi-C matrix after chromosomal rearrangement for a particular chromosome. It is stored as an HTCexp object from HiTC R package.

Author(s)

Raphael Mourad

`SIM` *Sparse Insulation Model (SIM)*

Description

Sparse Insulation Model (SIM) is a regression model that is used to map topologically associating domain (TAD) borders and facilitators. TADs are defined as regions in-between two consecutive TAD borders. SIM is based on a regression framework that generalizes the insulation score by estimating a relative score and adding a sparsity constrain.

Usage

```
SIM(HTC, distMax = NULL, penalty = "L0", prefilter = T)
```

Arguments

`HTC` A Hi-C matrix for a particular chromosome. It should be stored as an HTCexp object from HiTC R package.

`distMax` The maximal distance between two bins that is used to identify TADs. Usually a distance equal to 10 bins is fine (default value). Setting a too high maximal distance will lead to computational burden.

`penalty` The penalty (regularization) applied to the regression estimation: "none" (no penalty, classical regression), "L1" (lasso regression) and "L0" (default: L0 regression).

`prefilter` If the number of bins in the matrix is too big (eg > 2000), the L0 regression might fail to process all the bins (variables), due to computational burden. In this case, one can use a prefilter step based on lasso regression to remove bins with abs(betas)<0.2, and then to run L1 regression. Used by default.

Value

A GRanges object containing the Hi-C bin genomic coordinates and the corresponding beta values. Bins with betas < 0 correspond to TAD borders, whereas bins with betas > 0 correspond to TAD facilicators.

Author(s)

Raphael Mourad

Index