Package 'TADCompare'

January 25, 2021

Title TADCompare: Identification and characterization of differential TADs

Version 1.0.0

Author Kellen Cresswell <cresswellkg@vcu.edu>, Mikhail Dozmorov <mikhail.dozmorov@vcuhealth.org>

Maintainer Kellen Cresswell <cresswellkg@vcu.edu>

Description TADCompare is an R package designed to identify and characterize differential Topologically Associated Domains (TADs) between multiple Hi-C contact matrices. It contains functions for finding differential TADs between two datasets, finding differential TADs over time and identifying consensus TADs across multiple matrices. It takes all of the main types of HiC input and returns simple, comprehensive, easy to analyze results.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.0.1

LazyData true

BugReports https://github.com/dozmorovlab/TADCompare/issues

Imports dplyr, PRIMME, cluster, Matrix, magrittr, HiCcompare, ggplot2, tidyr, ggpubr, RColorBrewer, reshape2, cowplot

Suggests BiocStyle, knitr, rmarkdown, microbenchmark, testthat, covr, pheatmap, rGREAT, SpectralTAD

Depends R (>= 4.0)

VignetteBuilder knitr

biocViews Software, HiC, Sequencing, FeatureExtraction, Clustering

git_url https://git.bioconductor.org/packages/TADCompare

git_branch RELEASE_3_12

git_last_commit da4ed3d

git_last_commit_date 2020-10-27

Date/Publication 2021-01-24

R topics documented:

`ConsensusTADs` *Consensus boundary identification*

Description

Consensus boundary identification

Usage

```
ConsensusTADs(cont_mats, resolution, z_thresh = 3, window_size = 15,
  gap_thresh = 0.2)
```

Arguments

`cont_mats`	List of contact matrices in either sparse 3 column, n x n or n x (n+3) form where the first three columns are coordinates in BED format. See "Input_Data" vignette for more information. If an x n matrix is used, the column names must correspond to the start point of the corresponding bin. Required.
`resolution`	Resolution of the data. Used to assign TAD boundaries to genomic regions. If not provided, resolution will be estimated from column names of the first matrix. Default is "auto"
`z_thresh`	Threshold for boundary score. Higher values result in a higher threshold for differential TADs. Default is 3.
`window_size`	Size of sliding window for TAD detection, measured in bins. Results should be consistent Default is 15.
`gap_thresh`	Required % of non-zero entries before a region will be considered non-informative and excluded. Default is .2

Details

Given a list of sparse 3 column, n x n , or n x (n+3) contact matrices, ConsensusTADs provides the set of consensus TAD boundaries across them. Consensus TADs are defined by the consensus boundary score, a score measuring TAD boundary likelihood across all matrices.

Value

A list containing consensus TAD boundaries and overall scores

- Consensus - Data frame containing location of all consensus boundaries. Coordinate is the region of the genome, Sample columns correspond to individual boundary scores. Consensus_Score is consensus boundary score.
- All_Regions - Data frame containing consensus scores for all regions. All columns are identiical to the Consensus object.

Examples

```
# Read in data
data("time_mats")
# Find consensus TAD boundaries
diff_list <- ConsensusTADs(time_mats, resolution = 50000)
```

`DiffPlot` *Visualization of differential TAD boundaries*

Description

Visualization of differential TAD boundaries

Usage

```
DiffPlot(tad_diff, cont_mat1, cont_mat2, resolution, start_coord,
  end_coord, pre_tad = NULL, show_types = TRUE, point_size = 3,
  max_height = 25, rel_heights = c(2, 1), palette = "RdYlBu")
```

Arguments

`tad_diff`	Raw object output by TADCompare. Required.
`cont_mat1`	contact matrix in either sparse 3 column, n x n or n x (n+3) form where the first three columns are coordinates in BED format. See "Input_Data" vignette for more information. If an x n matrix is used, the column names must correspond to the start point of the corresponding bin. Should correspond to the first contact matrix input into TADCompare. Required.
`cont_mat2`	contact matrix in either sparse 3 column, n x n or n x (n+3) form where the first three columns are coordinates in BED format. If an x n matrix is used, the column names must correspond to the start point of the corresponding bin. Should correspond to the second contact matrix input into TADCompare. Required.
`resolution`	Resolution of the data. Required.
`start_coord`	The start coordinate defining a region to plot. Required.
`end_coord`	The end coordinate defining a region to plot. Required.
`pre_tad`	A list of pre-defined TADs for drawing. Must contain two entries with the first corresponding to TADs detected in matrix 1 and the second to those detected in matrix 2. Each entry must contain a BED-like data frame or GenomicRanges object with columns "chr", "start", and "end", corresponding to coordinates of TADs. Must correspond to TADCompare results obtained for the same pre-defined TADs. Optional

show_types	If FALSE only the labels "Differential" and "Non-Differential" will be used. More in-depth differential boundary types will be excluded. Default is TRUE.
point_size	Parameter used to adjust the size of boundary points on heatmap plot. Default is 3.
max_height	Maximum height in bins that should be displayed on the heatmap plot. Default is 25.
rel_heights	Proportion of the size of the heatmap and score panels. Should be a vector containing the relative size of each panel with the heatmap panel coming first and the score panel second. Default is c(2, 1).
palette	Parameter used to adjust color palette. For list of palettes see https://rdrr.io/cran/RColorBrewer/man/C Alternatively, users can define a vector of color names or hex codes. Default is 'RdYlBu'

Details

Given a TADCompare object and two corresponding contact matrices, Diff_Plot provides visualization of user-specified regions of the genome with accompanying differential annotations, TAD scores and differential TAD scores

Value

A ggplot plot containing a visualization of the upper diagonal both contact matrices with types of non-/differential boundaries labeled. The first matrix is shown on top and the second on the bottom. If pre_tad is provided, then the outline of the pre-defined TADs are shown. Individual TAD score and differential TAD scores are shown below the contact matrix plots.

Examples

```
# Read in data
data("rao_chr22_prim")
data("rao_chr22_rep")
# Find differential TAD boundaries
tad_diff <- TADCompare(rao_chr22_prim, rao_chr22_rep, resolution = 50000)
# Create plot
DiffPlot(tad_diff,rao_chr22_prim, rao_chr22_rep, resolution = 50000,
start_coord = 22050000, end_coord = 24150000)
```

GM12878.40kb.raw.chr2 *A subset of chomosome 2 contact matrix, GM12878 cell line.*

Description

A 1001x1001 contact matrix from the GM12878 cell line, chr2:8000000-48000000, 40kb Resolution, data from Schmitt et al. 2016.

Usage

```
GM12878.40kb.raw.chr2
```

Format

A data frame with 1001 rows and 1001 variables:

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87112

`IMR90.40kb.raw.chr2` *A subset of chomosome 2 contact matrix, IMR90 cell line.*

Description

A 1001x1001 contact matrix from the IMR90 cell line, chr2:8000000-48000000, 40kb Resolution, data from Schmitt et al. 2016.

Usage

```
IMR90.40kb.raw.chr2
```

Format

A data frame with 1001 rows and 1001 variables:

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87112

`rao_chr22_prim` *Chromosome 22 combined intrachromosomal primary contact matrix from Rao et al. 2014.*

Description

A 704x704 contact matrix from the GM12878 cell line (50kb Resolution)

Usage

```
rao_chr22_prim
```

Format

A data frame with 704 rows and 704 variables:

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525

rao_chr22_rep *Chromosome 22 combined intrachromosomal replicate contact matrix from Rao et al. 2014.*

Description

A 704x704 contact matrix from the GM12878 cell line (50kb Resolution)

Usage

```
rao_chr22_rep
```

Format

A data frame with 704 rows and 704 variables:

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63525

TADCompare *Differential TAD boundary detection*

Description

Differential TAD boundary detection

Usage

```
TADCompare(cont_mat1, cont_mat2, resolution = "auto", z_thresh = 2,
  window_size = 15, gap_thresh = 0.2, pre_tads = NULL)
```

Arguments

cont_mat1	Contact matrix in either sparse 3 column, n x n or n x (n+3) form where the first three columns are coordinates in BED format. See "Input_Data" vignette for more information. If an n x n matrix is used, the column names must correspond to the start point of the corresponding bin. Required.
cont_mat2	Second contact matrix, used for differential comparison, must be in same format as cont_mat1. Required.
resolution	Resolution of the data. Used to assign TAD boundaries to genomic regions. If not provided, resolution will be estimated from column names of matrix. If matrices are sparse, resolution will be estimated from the column names of the transformed full matrix. Default is "auto"
z_thresh	Threshold for differential boundary score. Higher values result in a higher threshold for differential TAD boundaries. Default is 2.
window_size	Size of sliding window for TAD detection, measured in bins. Results should be consistent regardless of window size. Default is 15.

gap_thresh — Required % of non-zero interaction frequencies for a given bin to be included in the analysis. Default is .2

pre_tads — A list of pre-defined TADs for testing. Must contain two entries with the first corresponding to TADs detected in matrix 1 and the second to those detected in matrix 2. Each entry must contain a BED-like data frame or GenomicRanges object with columns "chr", "start", and "end", corresponding to coordinates of TADs. If provided, differential TAD boundaries are defined only at these coordinates. Optional.

Details

Given two sparse 3 column, n x n , or n x (n+3) contact matrices, TADCompare identifies differential TAD boundaries. Using a novel boundary score metric, TADCompare simultaneously identifies TAD boundaries (unless provided with the pre-defined TAD boundaries), and tests for the presence of differential boundaries. The magnitude of differences is provided using raw boundary scores and p-values.

Value

A list containing differential TAD characteristics

- TAD_Frame - Data frame containing any bin where a TAD boundary was detected. Boundary refers to the genomic coordinates, Gap_Score refers to the orresponding differential boundary score. TAD_Score1 and TAD_Score2 are boundary scores for cont_mat1 and cont_mat2. Differential is the indicator column whether a boundary is differential. Enriched_In indicates which matrix contains the boundary. Type is the specific type of differential boundary.
- Boundary_Scores - Boundary scores for the entire genome.
- Count_Plot - Stacked barplot containing the number of each type of TAD boundary called by TADCompare

Examples

```
# Read in data
data("rao_chr22_prim")
data("rao_chr22_rep")
# Find differential TADs
diff_frame <- TADCompare(rao_chr22_prim, rao_chr22_rep, resolution = 50000)
```

TimeCompare *Time-varying TAD boundary analysis*

Description

Time-varying TAD boundary analysis

Usage

```
TimeCompare(cont_mats, resolution, z_thresh = 2, window_size = 15,
  gap_thresh = 0.2, groupings = NULL)
```

Arguments

cont_mats	List of contact matrices in either sparse 3 column, n x n or n x (n+3) form where the first three columns are coordinates in BED format. See "Input_Data" vignette for more information. If an n x n matrix is used, the column names must correspond to the start point of the corresponding bin. Required.
resolution	Resolution of the data. Used to assign TAD boundaries to genomic regions. If not provided, resolution will be estimated from column names of the first matrix. Default is "auto".
z_thresh	Threshold for boundary score. Higher values result in a more stringent detection of differential TADs. Default is 3.
window_size	Size of sliding window for TAD detection, measured in bins. Results should be consistent. Default is 15.
gap_thresh	Required % of non-zero entries before a region will be considered non-informative and excluded. Default is .2
groupings	Variable for defining groups of replicates at a given time point. Each group will be combined using consensus boundary scores. It should be a vector of equal length to cont_mats where each entry is a label corresponding to the group membership of the corresponding matrix. Default is NULL, implying one matrix per time point.

Details

Given a list of sparse 3 column, n x n, or n x (n+3) contact matrices representing different time points, TimeCompare identifies all TAD boundaries. Each TAD boundary is classified into six categories (Common, Dynamic, Early/Late Appearing and Early/Late Disappearing), based on how it changes over time.

Value

A list containing consensus TAD boundaries and overall scores

- TAD_Bounds - Data frame containing all regions with a TAD boundary at one or more time point. Coordinate corresponds to genomic region, sample columns correspond to individual boundary scores for each sample, Consensus_Score is the consensus boundary score across all samples. Category is the differential boundary type.
- All_Bounds - Data frame containing consensus scores for all regions
- Count_Plot - Plot containing the prevelance of each boundary type

Examples

```
# Read in data
data("time_mats")
# Find time varying TAD boundaries
diff_list <- TimeCompare(time_mats, resolution = 50000)
```

time_mats *Chromosome 22 time-varying contact matrices from Rao et al. 2017.*

Description

Four 704x704 contact matrices representing 20, 40, 60, 180 minutes since auxin treatment and removal from the HCT-116 cell line (50kb Resolution)

Usage

```
time_mats
```

Format

A data frame with 704 rows and 704 variables:

Source

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104334

Index