TopDom Manual

1. Requirement

Source code is implemented by R script. Your machine must be installed appropriate R packages. If R package is not installed on your machine, please visit http://www.r-project.org, download appropriate R version from CRAN, and install it on your machine. Note that this source code does not depend on any open-packages, so user does not need to install additional packages.

2. Download and Install

Please, visit http://zhoulab.usc.edu/TopDom/ and download the zipped file(TopDom.zip) in download section. No installation is required. Just unzip the download file and put the script (TopDom.R) into your working directory. You can check or change your working directory by the following commands.

```
> getwd() # check current working directory
> setwd("[YOUR DESIRABLE WORKING DIRECTORY ADDRESS]") # set your working directory
```

3. Usage

```
> source("TopDom.R") # Read source code from working directory
> TopDom(matrix.file=[matrix file address], window.size=[window.size],
outBinSignal=[.binSignal file address], outDomain=[.domain file address]) # Run TopDom
```

- *matrix.file*

  Input matrix.file is a normalized Hi-C contact matrix(n by n+3, where n is the number of bins for each chromosome) for a chromosome. Each column is separated by tab and the first three columns should include bin information, such as "chromosome", "from.coord", "to.coord". The headers are not allowed. Each row indicates a bin including position information and contact frequencies with the other bins.

  <FORMAT>

```
chr10   0        40000      0     0     0     0 .....
chr10   40000    80000      0     0     0     0 .....
chr10   80000    120000     0     0     0     0 .....
chr10   120000   160000     0     0     0     0 ......
```

- *window.size*

  Window size should be a nonnegative integer number and is used to compute binSignal. Recommended size is any integer number between 5 and 20.

- output file arguments(*outBinSingal, outDomain*)

  TopDom.R produces two types of output files, i.e. '.binSignal' and '.domain' file. If us er wants to keep the result as a file, specify a full file name. Note that default is set to NULL.

4. Output

The results are returned by *binSignal* and domain as list form. If user wants to keep results as as file form, give full file names for arguments, *outBinSingal and outDomain.*

- *binSignal*

The *binSignal* includes mean contact frequency, local extreme, and p-value for every bin. The first four columns represent basic bin information given by matrix file, such as bin id(id), chromosome(chr), start coordination(from.coord), and end coordination(to.coord) for each bin. And the last three columns represent computed values by this program.

<id> bin id

<chr> chromosome

<from.coord> start coordination of bin

<to.coord> end coordination of bin

<local.ext >

-1 : local minima.

-0.5 : gap region.

0 : general bin.

1 : local maxima.

<mean.cf> Average of contact frequencies between lower and upper regions for bin i

<p-value> Computed p-value by Wilcox rank sum test. Read reference for more details.

binSignal format

id	chr	from.coord	to.coord	local.ext	mean.cf	p-value
1	chr10	0	40000	-0.5	0	1
2	chr10	40000	80000	-0.5	0	1
3	chr10	80000	120000	-0.5	0	1
…						
1005	chr10	40160000	40200000	0	16.93	3.07e-01
1006	chr10	40200000	40240000	0	15.21	2.60e-03
1007	chr10	40240000	40280000	-1	13.77	2.97e-07
….						

- domain

Every bin is categorized by basic building block, such as gap, domain, or boundary. Each row indicates a basic building block. The first five columns include the basic information about the block, ‘tag’ column indicates the class of the building block.

<id> identifier of block

<chr> chromosome

<from.id> start bin index of the block
<from.coord> start coordination of the block
<to.id> end bin index of the block
<to.coord> end coordination of the block
<tag> categorized name of the block. Three possible blocks exists, “gap”, “domain”, and “boundary”
<size> size of the block

domain format

chr	from.id	from.coord	to.id	to.coord	tag	size
chr10	1	0	75	3000000	gap	3000000
chr10	76	3000000	112	4480000	domain	1480000
chr10	113	4480000	125	5000000	domain	520000
chr10	126	5000000	146	5840000	domain	840000
chr10	147	5840000	148	5920000	boundary	80000
….						

- bed(from v0.0.2)
  1. chrom
  2. chromStart
  3. chromEnd
  4. name

bed format

chr10	0	3000000	gap
chr10	3000000	4480000	domain
chr10	4480000	5000000	domain
chr10	5000000	5840000	domain
chr10	5840000	5920000	boundary
….			

5. Example Run

```
> source("TopDom.R")
> TopDom(matrix.file="chr10.nij.HindIII.comb.40kb.matrix",
window.size=10)
[1] "############################################################"
[1] "Step 0 : File Read and Matrix Scaling.."
[1] "############################################################"
[1] "-- Matrix Scaling...."
[1] "-- Done!"
[1] "Step 0 : Done !!"
[1] "############################################################"
[1] "Step 1 : Generating binSignals by computing bin-level
contact frequencies"
[1] "############################################################"
[1] "Step 1 : Done !!"
[1] "############################################################"
[1] "Step 2 : Detect TD boundaries based on binSignals"
[1] "############################################################"
[1] "Process Regions from  76 to 552"
[1] "Process Regions from  556 to 1439"
[1] "Process Regions from  1442 to 2028"
[1] "Process Regions from  2038 to 2494"
[1] "Process Regions from  2497 to 3250"
[1] "Step 2 : Done !!"
[1] "############################################################"
[1] "Step 3 : Statistical Filtering of false positive TD
boundaries"
[1] "############################################################"
[1] "-- Compute p-values by Wilcox Rank sum Test"
[1] "Process Regions from  76 to 552"
[1] "Process Regions from  556 to 1439"
[1] "Process Regions from  1442 to 2028"
[1] "Process Regions from  2038 to 2494"
[1] "Process Regions from  2497 to 3250"
[1] "-- Done!"
[1] "-- Filtering False Positives"
[1] "-- Done!"
[1] "Step 3 : Done!"
[1] "Job Complete !"
```

6. Reference

TopDom: An Efficient and Deterministic Method for Identifying Topological Domains in Genomes, Hanjun Shin[1,†], Yi Shi[2,†], Chao Dai[1], Harianto Tjong[1], Ke Gong[1], Frank Alber[1], and Xianghong Jasmine Zhou[1,*]

[1] Molecular and Computational Biology, University of Southern California, Los Angeles, California, 90033, USA

[2] Shanghai Center for Systems Biomedicine, Shanghai Jiaotong University, Shanghai, 200240, China.

* To whom correspondence should be addressed. Tel: 1 (213) 740 7055; Fax: 1 (213) 740 2475; Email: xjzhou@dornsife.usc.edu

† The authors wish it to be known that, in their opinion, the first 2 authors should be regarded as joint First Authors

7. Contact

Xianghong Jasmine Zhou, Professor
Biological Science and Computer Science
University of Southern California
Los Angeles, CA 90089-1340
Phone:(213) 740 7055
Fax:(213) 740 2475
http://zhoulab.usc.edu

Update History

- bed format support section is added(July/08/2016), for v0.0.2