HiCPlotter Manual

version 0.5.01

1. Requirements

Python 2.7.*

Numpy, Scipy, Matplotlib.

- Please note: scipy, numpy and matplotlib modules should be installed and updated to current version. Following versions of numpy (1.9.0, 1.9.2), scipy (0.14.0, 0.15.1) and matplotlib (1.3.1, 1.4.3) have been tested successfully.

2. Input data

HiCPlotter accepts the following input formats:

a. **Matrix format** (n-to-n)

Input file contains within-chromosomal interaction values.

Related Parameters:

-cn (--clearNaNs): If matrix contains NaN values, HiCPlotter will convert those to 0's. (Default value: 1 - set it to 0 for no conversion)

-fh (--fileHeader): If matrix file has header lines, HiCPlotter will skip entered number of lines. (Default value: 1 - skipped the first line).

If your file has no header line(s), please use –fh 0, otherwise you will receive the following error:

'unbalanced matrix inputFile! input should be a square matrix'

-ff (--fileFooter): If matrix file has footer lines, HiCPlotter will skip entered number of lines. (Default value: 0 - does not skipped any lines)

b. Triplet sparse format

If the input file is generated through HiC-Pro pipeline, interaction values are stored in a bed-like sparse file format, where first and second columns represent bins of interaction and the third column contains the interaction value. In order to decode bin's actual genomic locations, a bed file should be provided denoting genomic coordinates and the corresponding bin number. For more about this triplet sparse file format please check HiC-Pro website (http://nservant.github.io/HiC-Pro/RESULTS.html).

Related Parameters:

-tri (--tripleColumn): A boolean if the input file is generated by HiC-Pro pipeline (Default value: 0 - set it to 1 for activating).

-bed (--bedFile): A file name for the bin location bed file, required when -tri parameter is used.

```
python HiCPlotter.py -f GSE35156_GSM892306_hESC.40000.matrix -chr chr7 -o Example -r
40000 -tri 1 -bed GSE35156_GSM892306_hESC.40000_ord.bed -n hES
```

c. Random Bins (5C data type)

HiCPlotter can visualize 5C or other matrix files where interacting bins are not separated by equal distances. Currently, HiCPlotter can only visualize the interaction matrix with this file format.

Related Parameters:

-rb (--randomBins): A boolean if the input file bins are not equally distanced (Default value: 0 - set it to 1 for activating).

d. Whole Genome Matrix

HiCPlotter can visualize a multiple-chromosome input file, which contains interactions among different chromosomes or whole genome interactions. Currently, HiCPlotter can only visualize the interaction matrix with this file format. *A dummy parameter should be entered for whole genome plots from matrix files*.
If input format is triple sparse, -chr parameter will be used designate to the end chromosome, such as –chr chr11 will plot interactions starting from chr1 to chr11. Please use –chr chrY for whole genome interaction plots. Note: will skip mitochondrial interactions (chrM).

Related Parameters:
-wg (--wholeGenome): A boolean if the input file contains more than one chromosome interactions (Default value: 0 - set it to 1 for activating).

3. Basic usage

Required parameters:

-f (--file): A file name or a list of file names (separated by space) to be visualized as interaction matrices.

-n (--name): Name of files, labeled on top of each matrix (separated by space if multiple matrices are provided).

-chr (--chromosome): Name of the particular chromosome to be visualized (should contain "chr", such as chr1).
You will receive an error message if bedGraph files (for plotting additional tracks) do not include any lines with the given chromosome name.

-o (--output): A prefix for the output file name.

Example usage:

python HiCPlotter.py -f chr1.txt -chr chr1 -n IMR90 -o example

with multiple matrix files:

python HiCPlotter.py -f IMR.chr1.txt hES.chr1.txt -chr chr1 -n IMR90 hES -o example

Optional parameters:

-s (--start): An integer specifying the starting bin for visualizing the interaction matrix image. (Default value: 0).

-e (--end): An integer specifying the last bin for visualizing the interaction matrix image. (Default value: end bin).

-r (--resolution): An integer (in base pairs) for the space between two consequtive bins inside the interaction matrix input file. (Default value: 100000). Please remember to specify –r parameter if your data resolution is other than default value of 100000, otherwise labels on the plots will be misleading.

-spi (--spine): A boolean whether to remove top and right frame borders for each track. (Default value: 0 - set to 1 for activating).

with start and end values:

python HiCPlotter.py -f chr1.txt -chr chr1 -n IMR90 -o example -s 500 -e 750 -r 40000

4. Track types

a. Matrix

To customize the interaction matrix visualization, following parameter can be utilized:

-mm (--matrixMax): An integer to set highest color scale for the heatmap (Default value is the highest value inside the first matrix file).

-hmc (--heatmapColor): An integer between 0-5 (Greys(0), Reds(1), YellowToBlue(2), YellowToRed(3-default), Hot(4), BlueToRed(5)) for choosing color scale of the heatmap.

-ptr (--plotTriangular): A boolean whether to plot rotated half interaction matrix (Default value: 0 - set to 1 for activating).

```
python HiCPlotter.py -f IMR90.txt hES.txt -chr chr1 -n IMR90 hES -o example -mm 8  -hmc 2 -ptr 1
```

b. Histograms

Histograms are useful to visualize continuous data types such as ChIP-Seq, DNase-Seq, Repli-Seq, RAP-Seq or 4C assay outputs along a given chromosome.

Related parameters:

-hist (--histograms): A file name or a list of file names to be visualized as histograms. Comma separated for multiple histograms of each matrix.

-hl (--histLabels): A name or a list of names to be plotted as labels for histograms. Comma separated for multiple histograms of each matrix.

Optional parameters:

-hm (--histMax): An integer to set highest y-value for the histogram (Default value is the highest value in the plotted range). Comma separated for multiple histograms of each matrix.

-fhist (--fillHist): A boolean whether to fill area under curve for individual histograms (Default value: 0 - set to 1 for activating). Comma separated for multiple histograms of each matrix.

-hc (--histColors): A list of hexadecimal numbers to determine the color for filling area under curve for individual histograms. Comma separated for multiple histograms of each matrix.

```
python HiCPlotter.py -f mES.chr2 -n mES -chr chr2 -r 40000 -o HoxD -hist GSM1334415.bedGraph,GSM1334440.bedGraph,GSM747534.bedGraph -hl Hoxd4ES,Hoxd4-Tail,CTCF-ES -s 1830 -e 1880 -fh 0 -fhist 1,1,0 -hm 2000,2000,50 -hc BCBCBC,FF0000,CC5500
```

c. Bar Plot

Bar plots can be used especially to visualize gene expression levels for a given locus.
Related parameters:

-b (--barPlots): A file name or a list of file names to be visualized as bar plots. Comma separated for multiple tiles of each matrix.

-bl (--barLabels): A name or a list of names to be plotted as labels for bar plots. Comma separated for multiple tiles of each matrix.

Optional parameters:

-bm (--barMax): An integer to set highest y-value for the bar plots (Default value is the highest value in the plotted range). Comma separated for multiple bar plots of each matrix.

-bc (--barColor): An hexadecimal number whether to fill bars with custom colors (Default color: #0099FF and color will be shaded based on maximum number, so color of the bar will be correlated with the height of the bar. If bedGraph file contains color column (RGB format), this parameter will be ignored).

```
python HiCPlotter.py -f mES.chr6 -n mES -chr chr6 -r 40000 -o Digit.vs.GT -s 1295 -e 1338 -fh 0 -b
rnaSeq.bedGraph -bl Expression
```

d. Tiles

Tiles can be used to depict discrete genomic features such as chromatin states/domains, enhancer locations, or structural alterations.
Related parameters:

-t (--tilePlots): A file name or a list of file names to be visualized as tiles. Comma separated for multiple tiles of each matrix.

-tl (--tileLabels): A name or a list of names to be plotted as labels for tiles. Comma separated for multiple tiles of each matrix.

Optional parameters:

-tt (--tileText): A boolean whether to plot texts above each tile (Default value: 0 - set to 1 for activating).

-tc (--tileColor): An hexadecimal number whether to fill tiles with custom colors (Default color: #0099FF. If bedGraph file contains color column (RGB format), this parameter will be ignored).

python HiCPlotter.py -f mES.chr6 -n mES -chr chr6 -r 40000 -o Digit.vs.GT -s 1295 -e 1338 -fh 0 -t primers.bedGraph -tl Enhancers -tt 1

e. Arcs

Arcs represent connectivity between two loci; this type of visualization is useful for assay outputs including 3C or ChIA-Pet.

Related parameters:

-a (--arcPlots): A file name or a list of file names to be visualized as arcs. Comma separated for multiple arcs of each matrix.

-al (--arcLabels): A name or a list of names to be plotted as labels for arcs. Comma separated for multiple arcs of each matrix.

Optional parameter(s):

-ac (--arcColors): A hexadecimal number whether to fill area under the arcs with custom colors (If bedGraph file contains color column (RGB format), this parameter will be ignored).

python HiCPlotter.py -f mES.chr3 -n mES -chr chr3 -o Bhlhe22 -r 40000 -s 400 -e 500 -a mESC_SMC_ChIPPet.bed -al SMC -ac B4B4B4 -fh 0

f. Epilogos

Epilogos is developed in Manolis Kellis' lab for visualization and analysis of chromatin state model data in various cell types. More about epilogos, check: http://compbio.mit.edu/epilogos/#

You can download the epilogos data from:
http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/epilogos/

Related parameters:

-ep (--epiLogos): A path for epilogos file.

Optional parameter(s):

-im (--imputed): A boolean whether the epilogos file belongs to imputed chromatin state models (Default value: 0 - set to 1 for activating).

python HiCPlotter.py -f IMR90.txt hES.txt -chr chr1 -n IMR90 hES -o exp -ep qcat -im 1

Currently color of each states for Epilogos plotting is hard-coded in HiCPlotter, therefore please use qcat files in imputed or observed folders. In an upcoming release a json file reader will be implemented for reading Epilogos files to overcome this issue.

g. Domains

Different chromosomal domains can be plotted with HiCPlotter, such as TADs or Arrowhead domains.

Related parameters:

-ptd (--plotTadDomains): A boolean whether to plot TADs called by an algorithm (deprecated) implemented in HiCPlotter (Default value: 0 - set to 1 for activating).

-pcd (--plotCustomDomains): A boolean whether to plot custom domains (Default value: 0 - set to 1 for activating).

-pcdf (--plotCustomDomainsFile): A file name or a list of file names to be visualized as domains, if -pcd parameter is used.

-pptd (--plotPublishedTadDomains): A boolean whether to plot TAD domains from Dixon et, al 2012 (Default value: 0 - set to 1 for activating). Human TADs are for hg19 and mouse domains are for mm9 genome assembly.

-ppto (--plotPublishedTadDomainsOrganism): A boolean whether to plot published TADs from Dixon et, al 2012 for human or mouse genomes (Default value: 1 for human - set to 0 for mouse).

-pdb (--plotDomainsAsBars): A boolean whether to plot domains as bars instead of triangles (Default value: 0 - set to 1 for activating).

5. Annotating Hi-C matrix

a. Highlighting

Users can highlight specific regions on the interaction matrix to augment readability of the plot.

Related parameters:

-high (--highlights): A boolean whether to highlight certain genomic regions on the plot including additional tracks (Default value: 0 - set to 1 for activating).

-hf (--highFile): A bed file name if –high parameter is set to 1 for highlighting certain genomic regions.

b. Adding circles

Looping between distant chromatin loci can also be annotated for pre-selected loci on the interaction matrix heatmap.

Related parameters:

-peak(--peakFiles): A file name or a list of file names to be visualized as circles on the interaction matrix heatmap. Color of the each circle can be determined inside the bedGraph file.

6. TAD-calling (deprecated)

HiCPlotter contains a TAD-calling algorithm based on insulation score calculation: for a given window (-w) sum of interaction off from the diagonal will be calculated. Local minima will be determined for a given TAD range (-td).

Related parameters:

-w (--window): An integer in bins to calculate insulation scores off from the diagonal (Default value: 5 – based on the resolution of matrix file, bins corresponding to 500-800Kb range is suggested).

-td (--tadRange): An integer in bins to calculate local minima in the insulation scores (Default value: 8 – based on the resolution of matrix file, bins corresponding to 800-2000Kb range is suggested).

-pi (--plotInsulation): A boolean whether to plot insulation scores calculated for the given window and tadRange parameters (Default value: 0 - set to 1 for activating).

We stopped improving TAD-calling algorithm in HiCPlotter, user are still welcomed to plot domain calls with the in-built function however we suggest users check more up-to-date TAD-callers.

Some of the available TAD-callers:

Dixon et, al. Directionality-index approach:
http://bioinformatics-renlab.ucsd.edu/collaborations/sid/domaincall_software.zip

Crane et,al Insulation-score approach:
https://github.com/blajoie/crane-nature-2015

TADbit:
https://github.com/3DGenomes/tadbit

Armatus:
http://www.cs.cmu.edu/~ckingsf/software/armatus/

Matryoshka:
https://github.com/COMBINE-lab/matryoshka

TADtree:
http://compbio.cs.brown.edu/projects/tadtree/