GenomeFlow User Manual

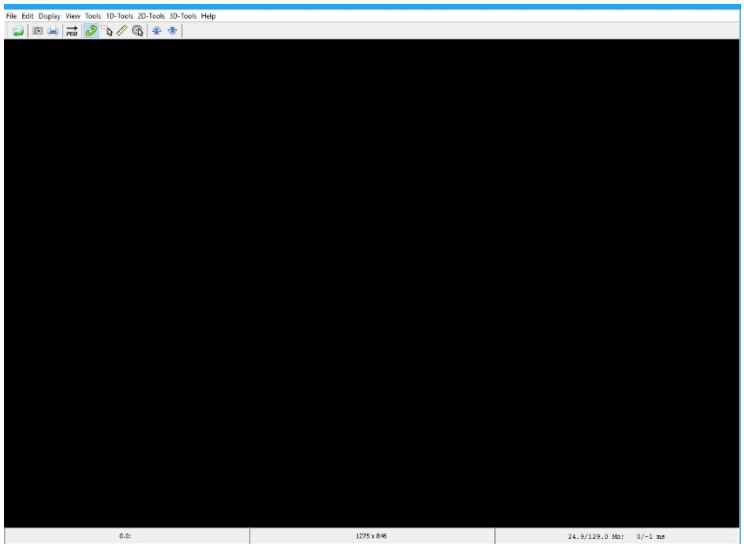
Todo:

- Re-organize functions according to their orders in the menu

Contents

1. Visualize 2D Dataset

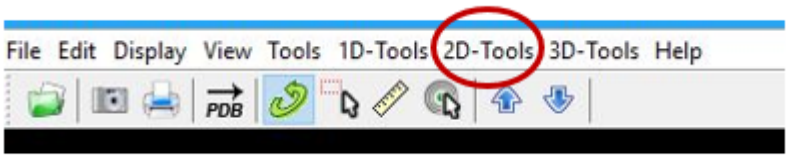
a) GenomeFlow Home Screen

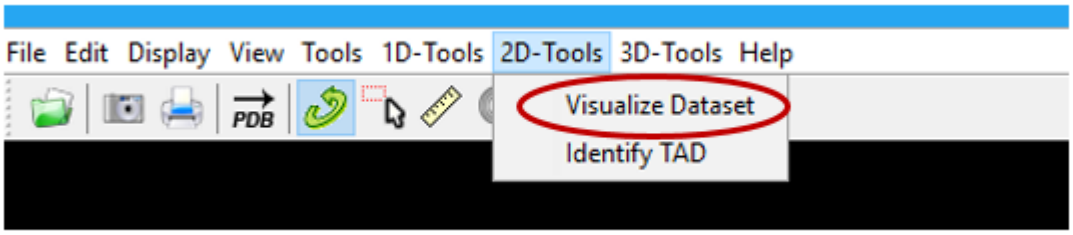


b) Navigation to 2D Visualization window

To use the 2D functions, Navigate to the 2D-Tools menu in the Menu bar.

Select the Submenu, Visualize Dataset

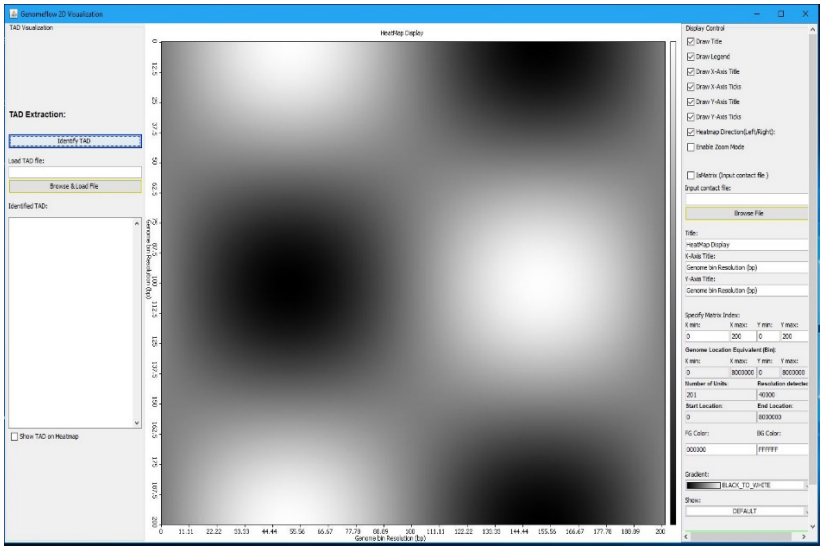




c) Graphical User Interface window

The 2D Graphical User Interface (GUI) window below pops up once you click on the Visualize Dataset

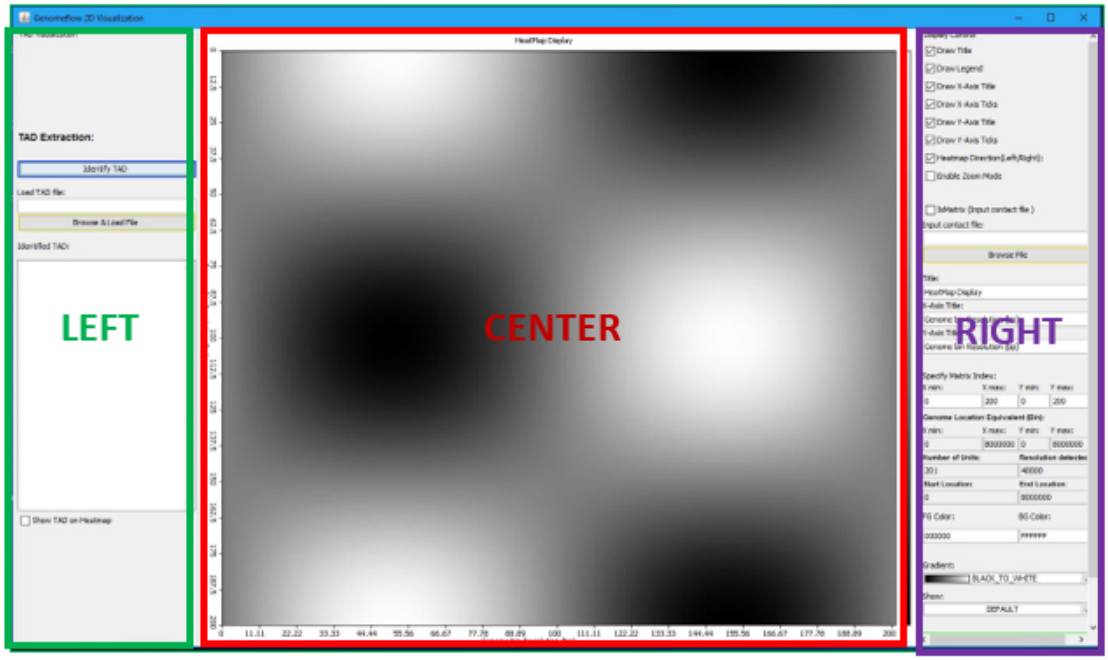
Sub-menu.



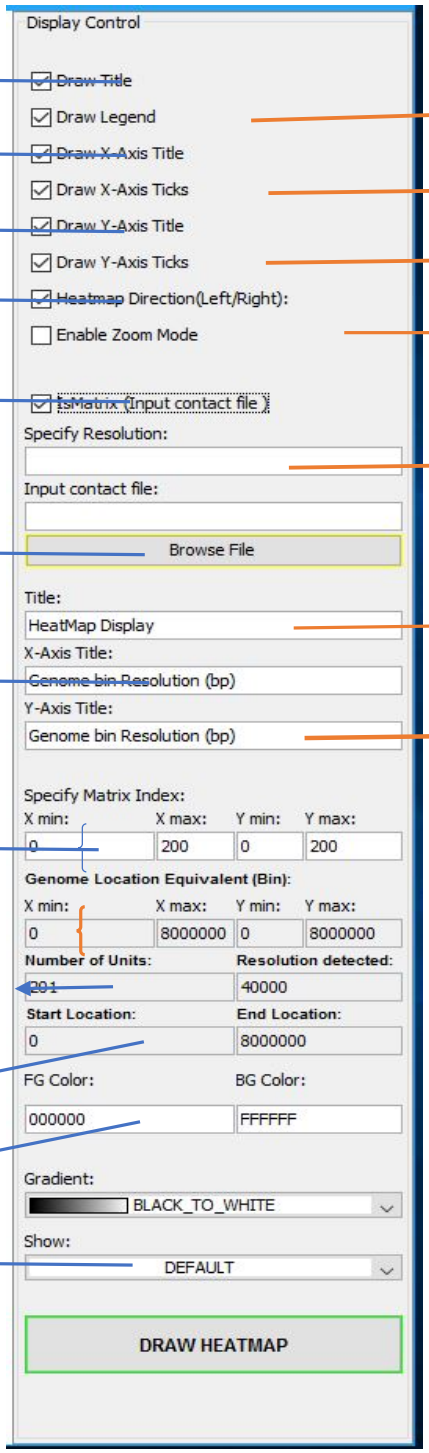
d) Graphical User Interface (GUI) Segments

The 2D Graphical User Interface (GUI) window is divided into 3 segments: LEFT, CENTER, AND RIGHT.

a. The RIGHT: Contains the Display Controls for the data heatmap
b. The CENTER: It displays the dataset in heatmap format
c. The LEFT: Contains the TAD Visualization controls


TAD Extraction:
Identify TAD
Load TAD file:
Browse & Load File
Identified TAD:
Show TAD on Heatmap
LEFT
CENTER
RIGHT

e) The RIGHT Segment - Display Controls



Show/Hide Title

Show/Hide X-Axis Label

Show/Hide Y-Axis Label

Change heatmap direction

Indicate if dataset is a Matrix

Browse and select input data

Add the X-Axis Title Label

X min: Minimum X-axis index
X max: Maximum X-axis index
Y min: Minimum Y-axis index
Y max: Maximum X-axis index

Number of Bins/Regions identified from dataset

Dataset Start Genome Location

Foreground Color

Calculate the Pearson, Spearman and Tanh of data and Visualize

Show/Hide Legend

Show/Hide X-Axis Ticks

Show/Hide Y-Axis Ticks

Enable Heatmap Zoom

Enter Data Resolution (if IsMatrix is checked)

Add the Title Label

Add the Y-Axis Title Label

Dataset Genome Location equivalent of X min, X max. Y m and Y max

Resolution detected for input d

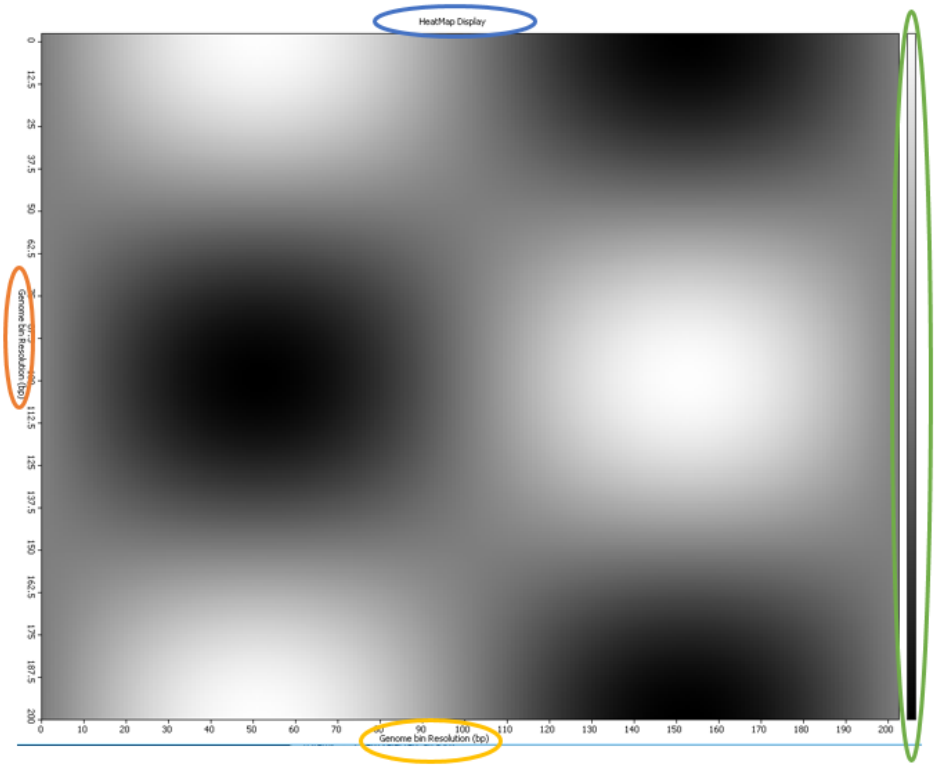
Dataset End Genome Location

Background Color

Choose Color Gradient

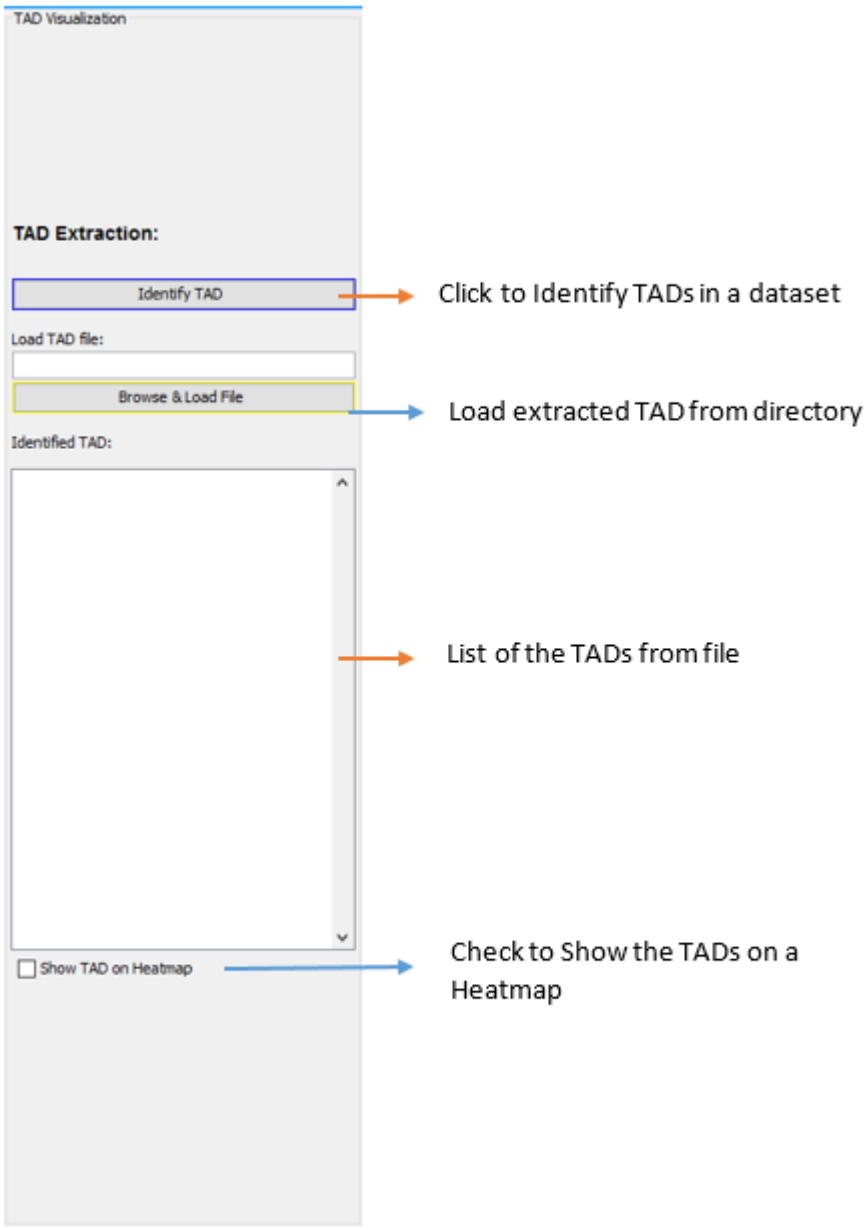
Draw the input data as Heatmap

f) The CENTER Segment – Display Area



Sphere Representation	Description
	Heatmap Title
	X-Axis Label
	Y-Axis Label
	Color Gradient. [Top means High, Bottom means Low]

g) The LEFT Segment – TAD Visualization

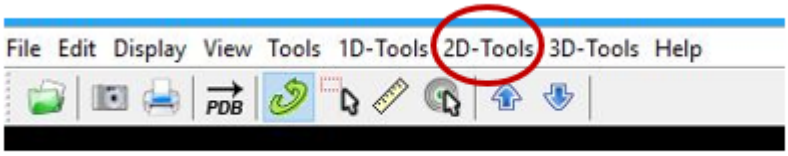

TAD Visualization
TAD Extraction:
Identify TAD
Click to Identify TADs in a dataset
Load TAD file:
Browse & Load File
Load extracted TAD from directory
Identified TAD:
List of the TADs from file
Show TAD on Heatmap
Check to Show the TADs on a Heatmap

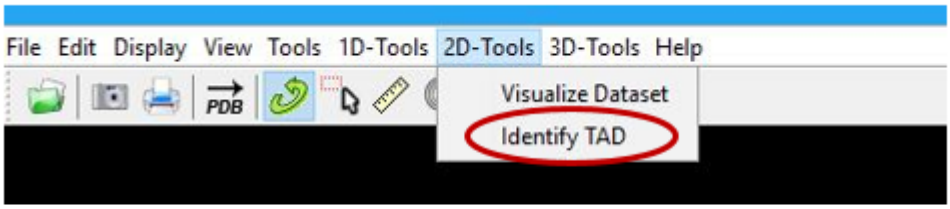
2. TAD Identification

a) Navigation to TAD Identification window

To use the 2D functions, Navigate to the 2D-Tools menu in the Menu bar.

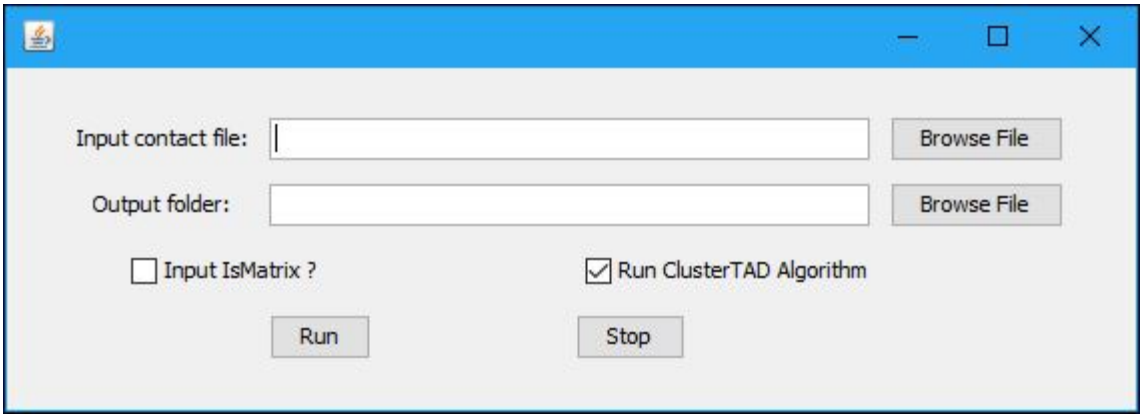
Select the Submenu, Identify TAD





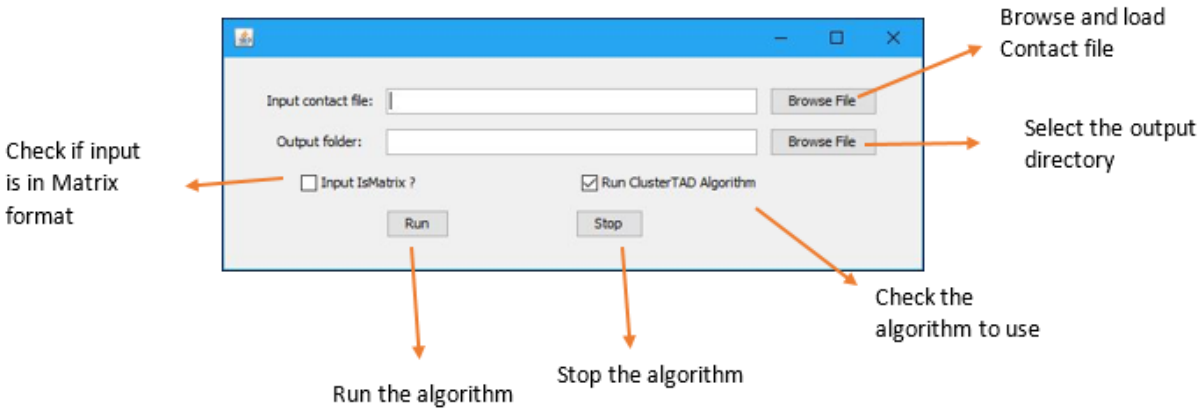
b) Graphical User Interface window

The Graphical User Interface (GUI) window below pops up once you click on the Identify TAD Sub-menu.

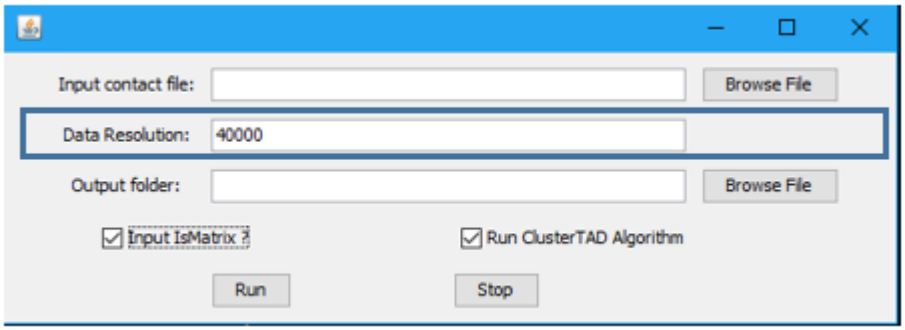


c) TAD identification window controls

- The function of each button is labelled below:



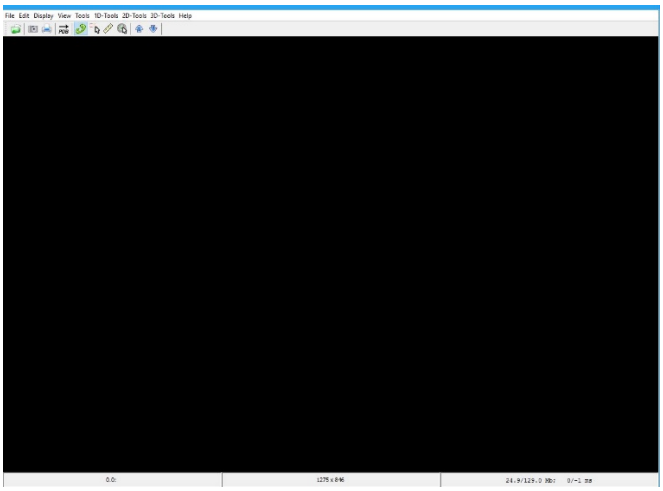
- If the Input IsMatrix? Checkbox is checked, the Data Resolution is required from the user.



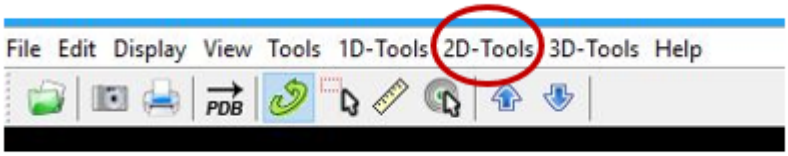
3. Demonstration

a) How to visualize a dataset in 2D Heatmap?

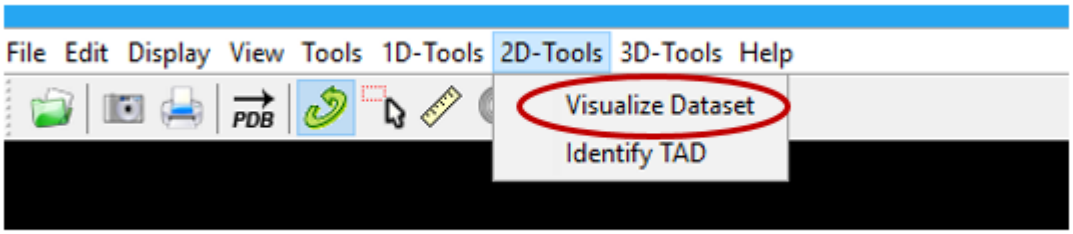
1. Goto Genome Home Screen



2. Click on the 2D-Tools Menu in the Home Screen

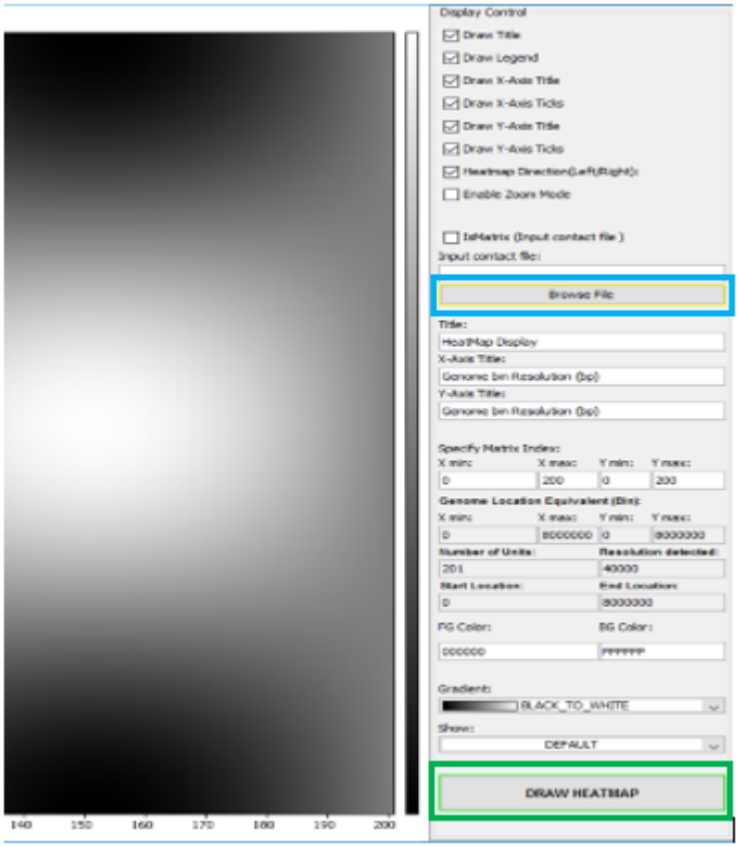


3. Click on the Visualize Dataset Sub-menu

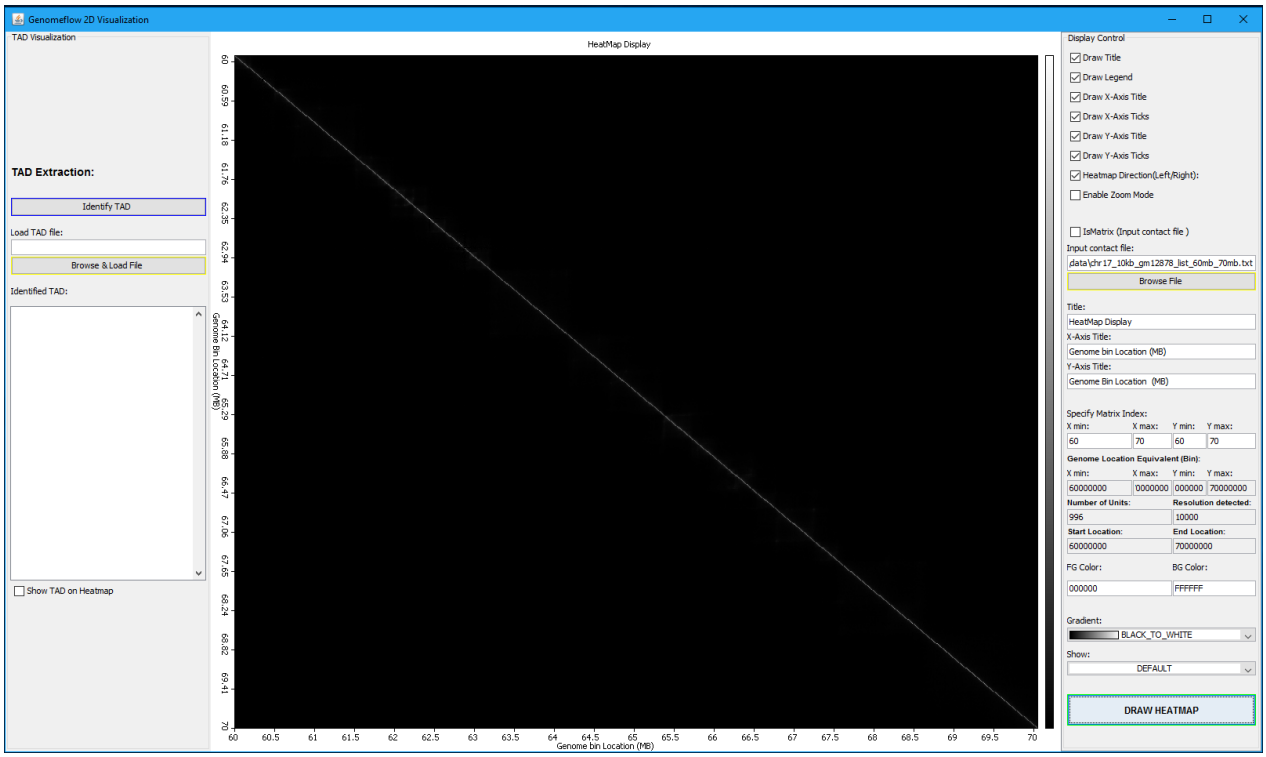


4. Click on the Browse File button in the Display Control of the RIGHT Segment

5. Click the Draw Heatmap button to visualize the dataset on a Heatmap



6. Following Steps 1 to 5 the Heatmap below is shown for the “sample_data.txt”

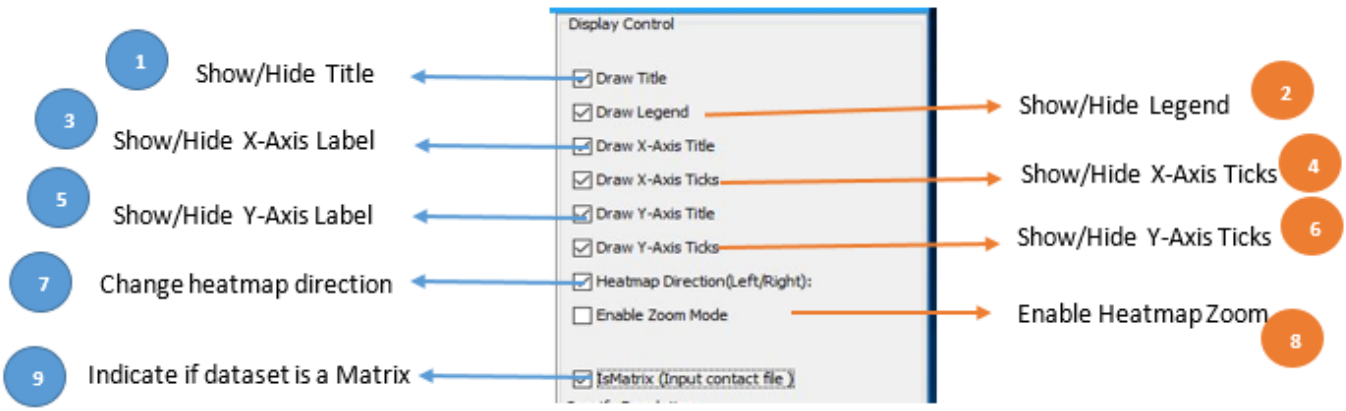


7. The default input data format is a Tuple. That is input data has 3 columns, column 1 and 2 represent the pair genomic location in contact, and column 3 represent the

interaction frequency between them. An example is shown below:

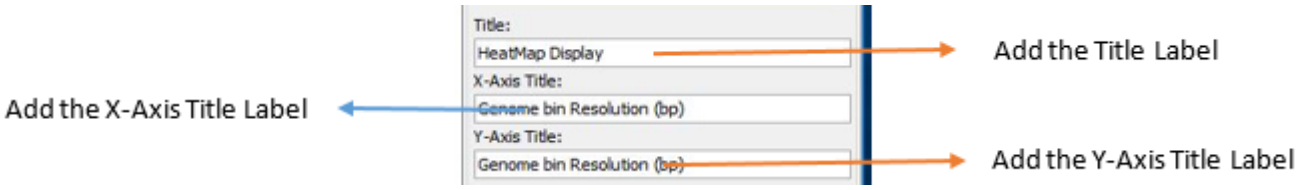
60000000	60000000	1277.2255853605477
60000000	60010000	555.3944883513798
60010000	60010000	1303.547855090625
60000000	60020000	265.8276247470504
60010000	60020000	500.7223214001157
60020000	60020000	1404.2005949454763
60000000	60030000	233.38186055746246
60010000	60030000	313.58347483503377
60020000	60030000	595.709350668511
60030000	60030000	1415.2894545838824
60000000	60040000	170.26755906956208
60010000	60040000	245.3383624354434

b) Effect of Check boxes in the Display control

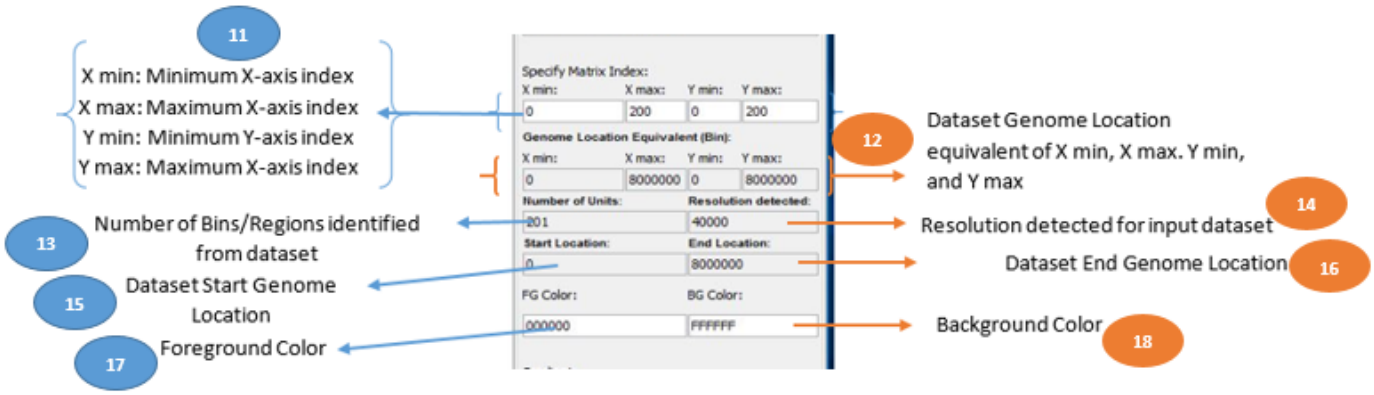


1. Click on the Draw Title check box to determine the Heatmap Title visibility. Check to Show, Uncheck to hide.
2. Click on the Draw Legend check box to determine the Heatmap Gradient visibility. Check to Show, Uncheck to hide.
3. Click on the Draw X-Axis Title check box to determine the Heatmap X-Axis Title visibility. Check to Show, Uncheck to hide.
4. Click on the Draw X-Axis Ticks check box to determine the Heatmap X-Axis Ticks visibility. Check to Show, Uncheck to hide
5. Click on the Draw Y-Axis Title check box to determine the Heatmap Y-Axis Title visibility. Check to Show, Uncheck to hide.
6. Click on the Draw Y-Axis Ticks check box to determine the Heatmap Y-Axis Ticks visibility. Check to Show, Uncheck to hide
7. Click on the Heatmap Direction (Left / Right) check box to determine the Heatmap draw direction. Check to draw from Left to Right, uncheck to draw from Right to Left.
8. Click on the Enable Zoom Mode check box to enable zoom in or out. Check to zoom, uncheck to return to default.
9. Click on the IsMatrix check box if input is in Matrix format. Check to specify input data is in Matrix form, uncheck to accept default input format.

c) Effect of Text boxes in the Display control



10. Enter the Heatmap, the X-Axis, and the Y-Axis title in the textbox as shown above.



11. Shows the minimum and maximum index for X-Axis and Y-Axis for the input dataset. User can specify a new X-Axis and Y-Axis range, and the data will be shown on the Heatmap Display.

12. Shows the equivalent of the X and Y Axis label in 11 above in Genomic Location. This conversion is done based on the Resolution of the dataset.

13. It shows the number of bins/Regions in the dataset extractable based on the input dataset resolution.

14. It detects the resolution of the dataset from the dataset.

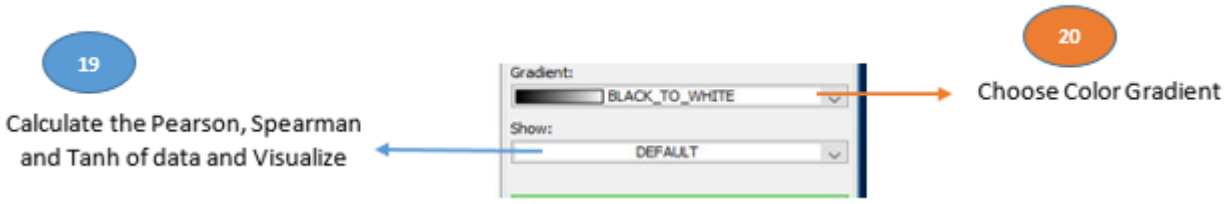
15. It shows the Genomic Start location of the input dataset.

16. It shows the End Start location of the input dataset.

17. Specify the foreground color for the heatmap display.

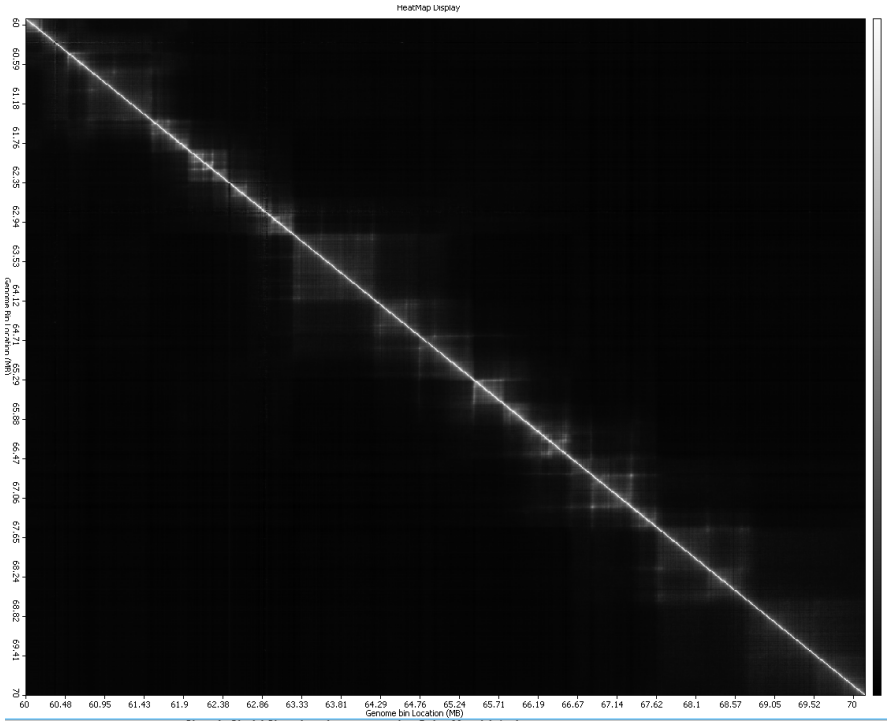
18. Specify the background color for the heatmap display.

d) Effect of Dropdown list in the Display control

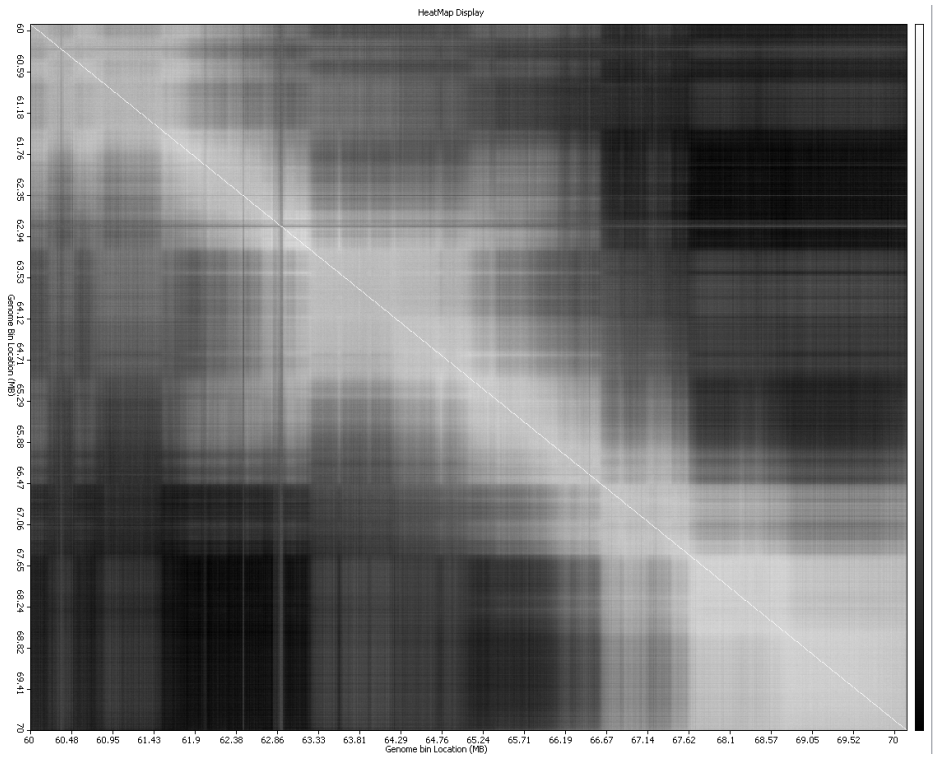


19. Show the Default, Pearson, Spearman, and Tanh equivalent of dataset. Heatmap below shows the representation of the dataset for each of them.

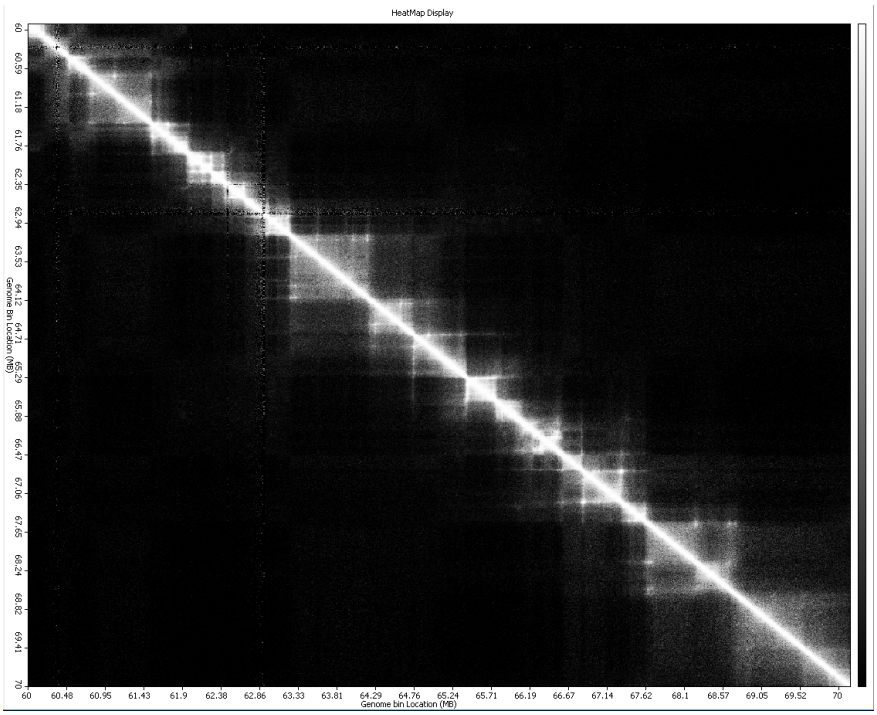
Pearson



Spearman

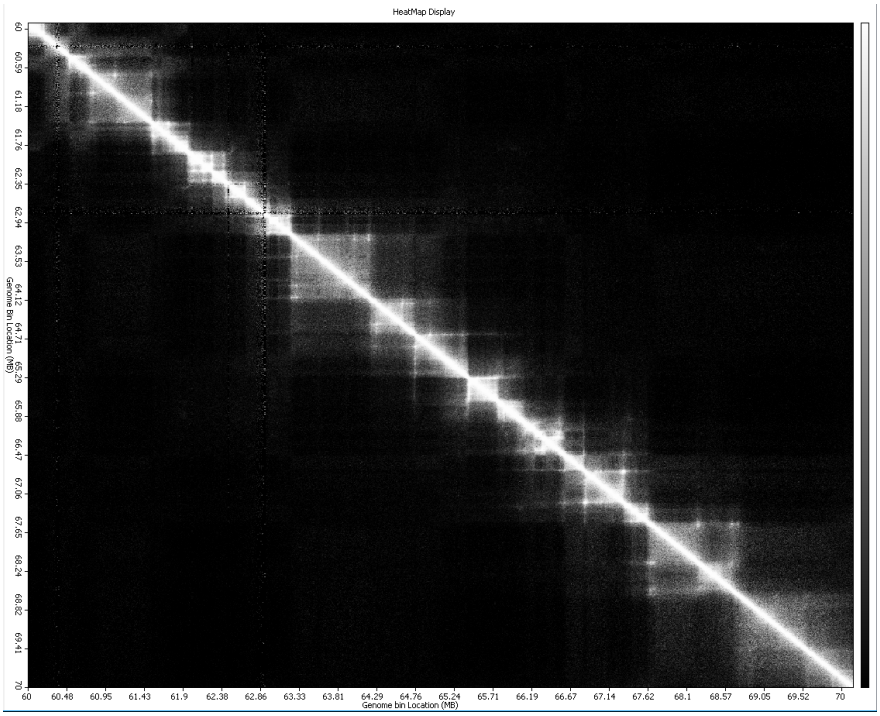

HeatMap Display
Genome bin Location (MB)
60
60.59
61.18
61.76
62.35
62.94
63.53
64.12
64.71
65.29
65.88
66.47
67.06
67.65
68.24
68.82
69.41
70
60
60.48
60.95
61.43
61.9
62.38
62.86
63.33
63.81
64.29
64.76
65.24
65.71
66.19
66.67
67.14
67.62
68.1
68.57
69.05
69.52
70
Genome bin Location (MB)

Tanh

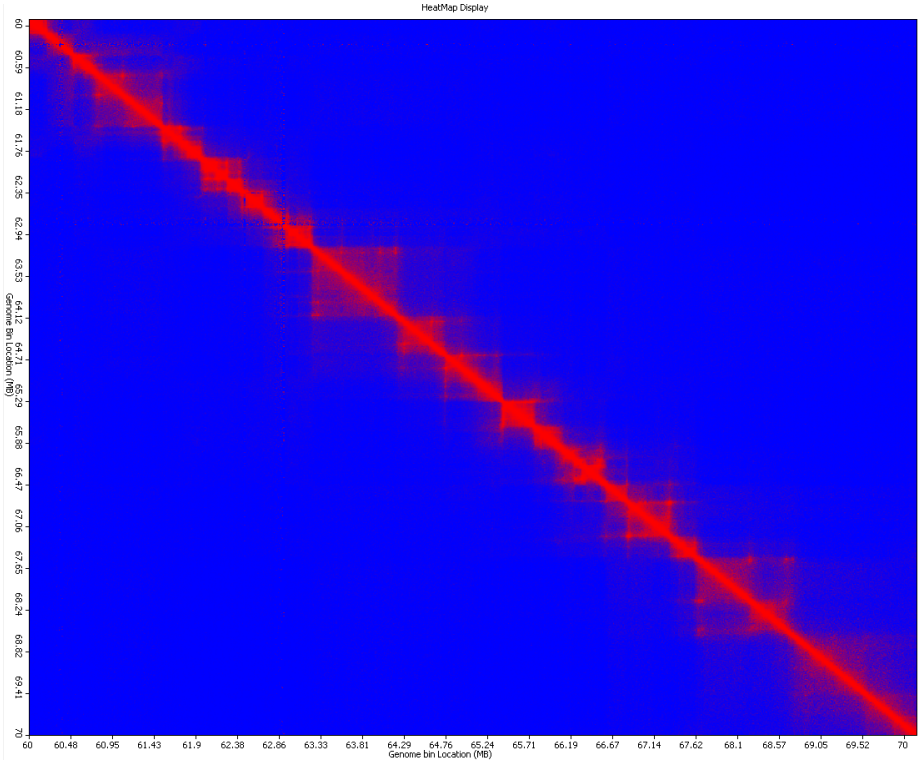


20. Choose the color gradient to represent the input dataset. The Heatmaps below shows the dataset representation for each of the gradient representation when **Tanh dataset** is used.

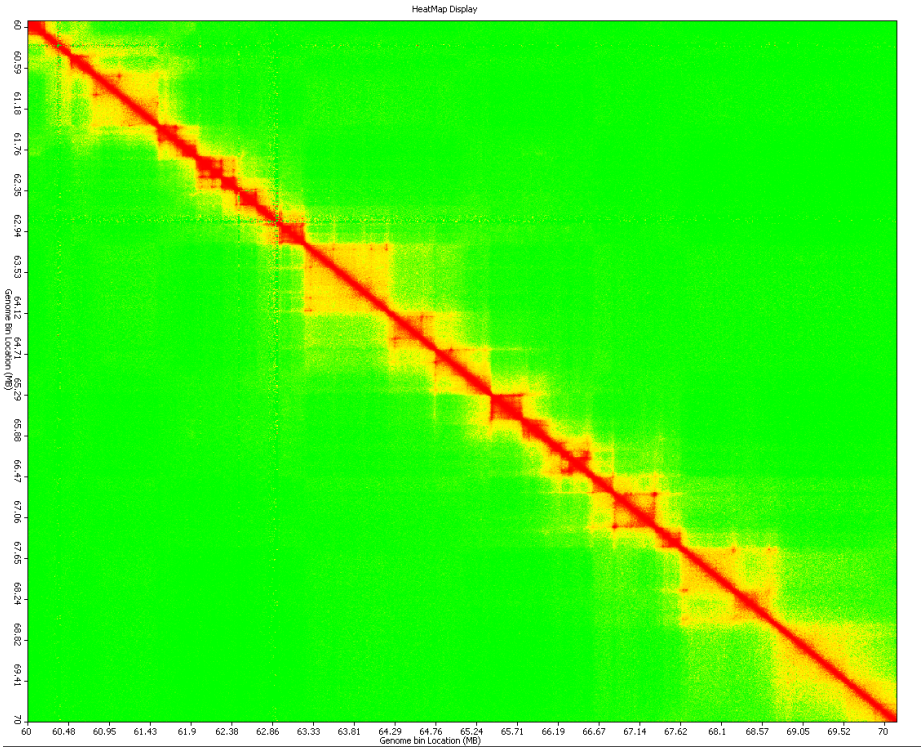
BLACK_TO_WHITE GRADIENT



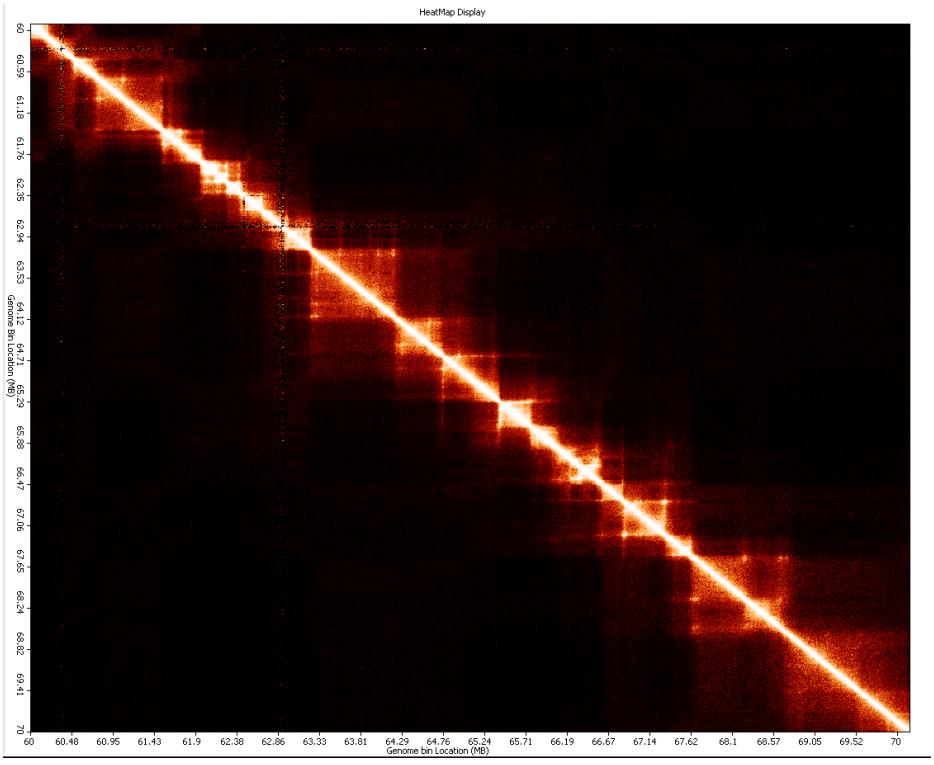
BLUE_TO_RED GRADIENT



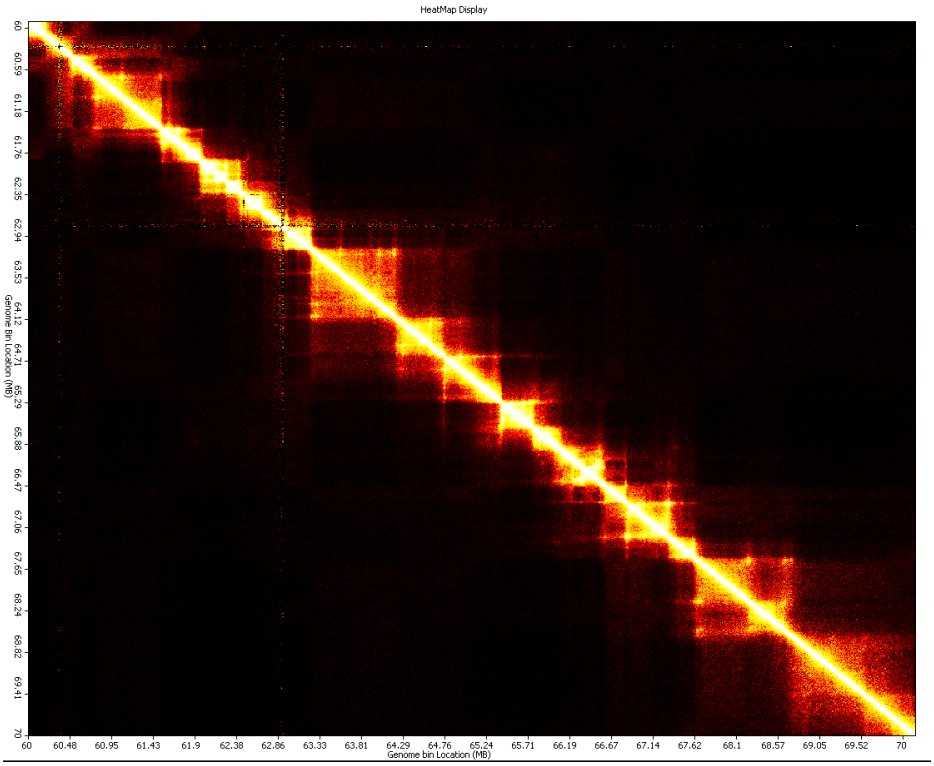
GREEN_YELLOW_ORANGE_RED


HeatMap Display
Genome bin Location (MB)
60
60.59
61.18
61.76
62.35
62.94
63.53
64.12
64.71
65.29
65.88
66.47
67.06
67.65
68.24
68.82
69.41
70
60.48
60.95
61.43
61.9
62.38
62.86
63.33
63.81
64.29
64.76
65.24
65.71
66.19
66.67
67.14
67.62
68.1
68.57
69.05
69.52

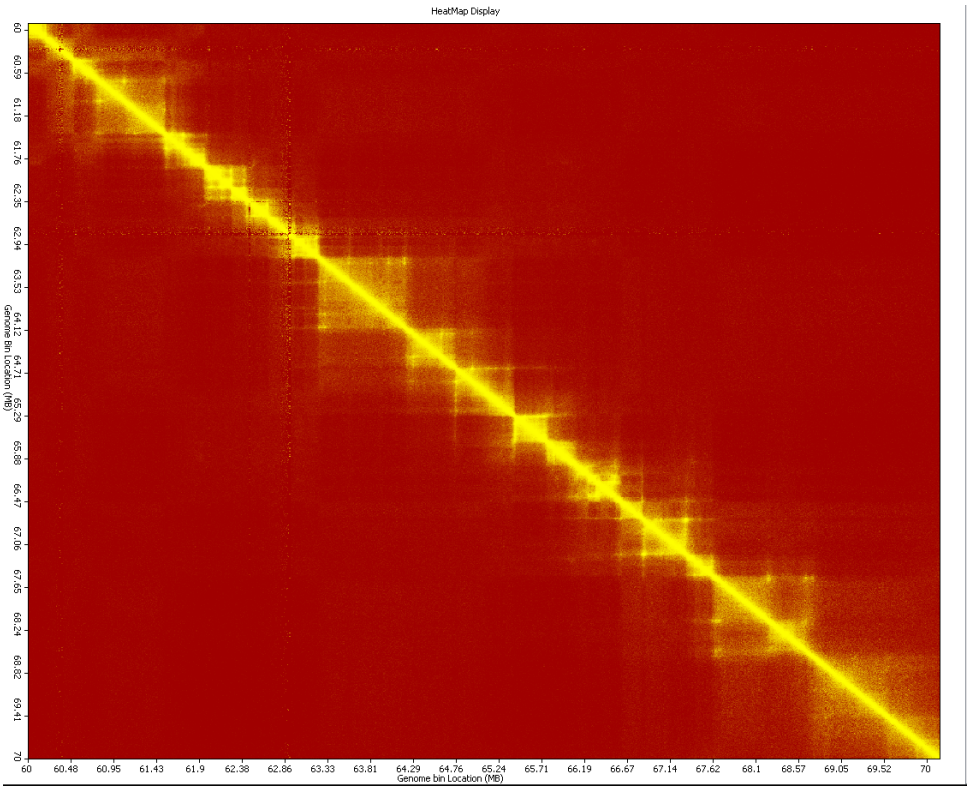
HEAT


HeatMap Display
Genome bin Location (MB)
60
60.59
61.18
61.76
62.35
62.94
63.53
64.12
64.71
65.29
65.88
66.47
67.06
67.65
68.24
68.82
69.41
70
60.48
60.95
61.43
61.9
62.38
62.86
63.33
63.81
64.29
64.76
65.24
65.71
66.19
66.67
67.14
67.62
68.1
68.57
69.05
69.52

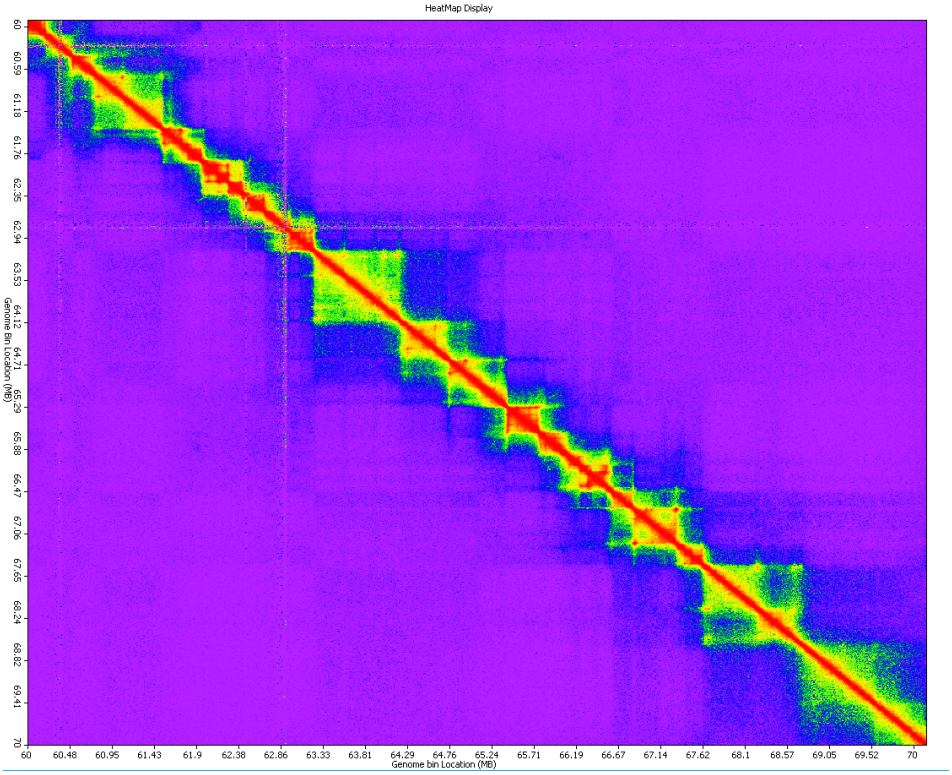
HOT



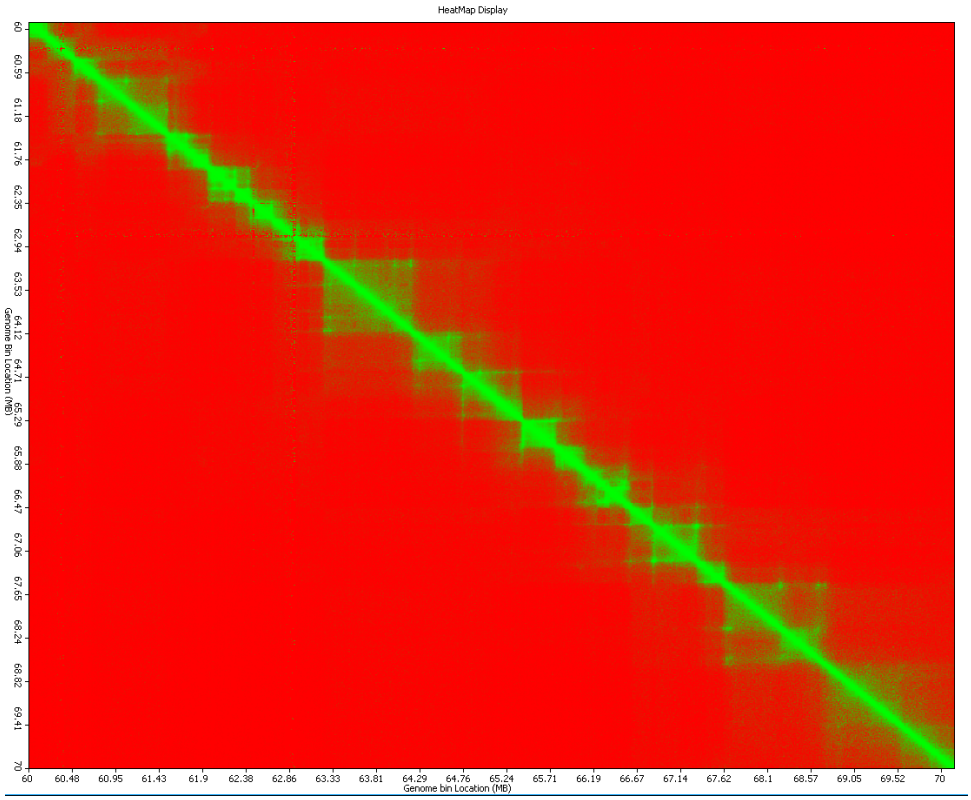
MAROON_TO_GOLD


HeatMap Display
Genome bin Location (MB)
60
60.59
61.18
61.76
62.35
62.94
63.53
64.12
64.71
65.29
65.88
66.47
67.06
67.65
68.24
68.82
69.41
70
60.48
60.95
61.43
61.9
62.38
62.86
63.33
63.81
64.29
64.76
65.24
65.71
66.19
66.67
67.14
67.62
68.1
68.57
69.05
69.52

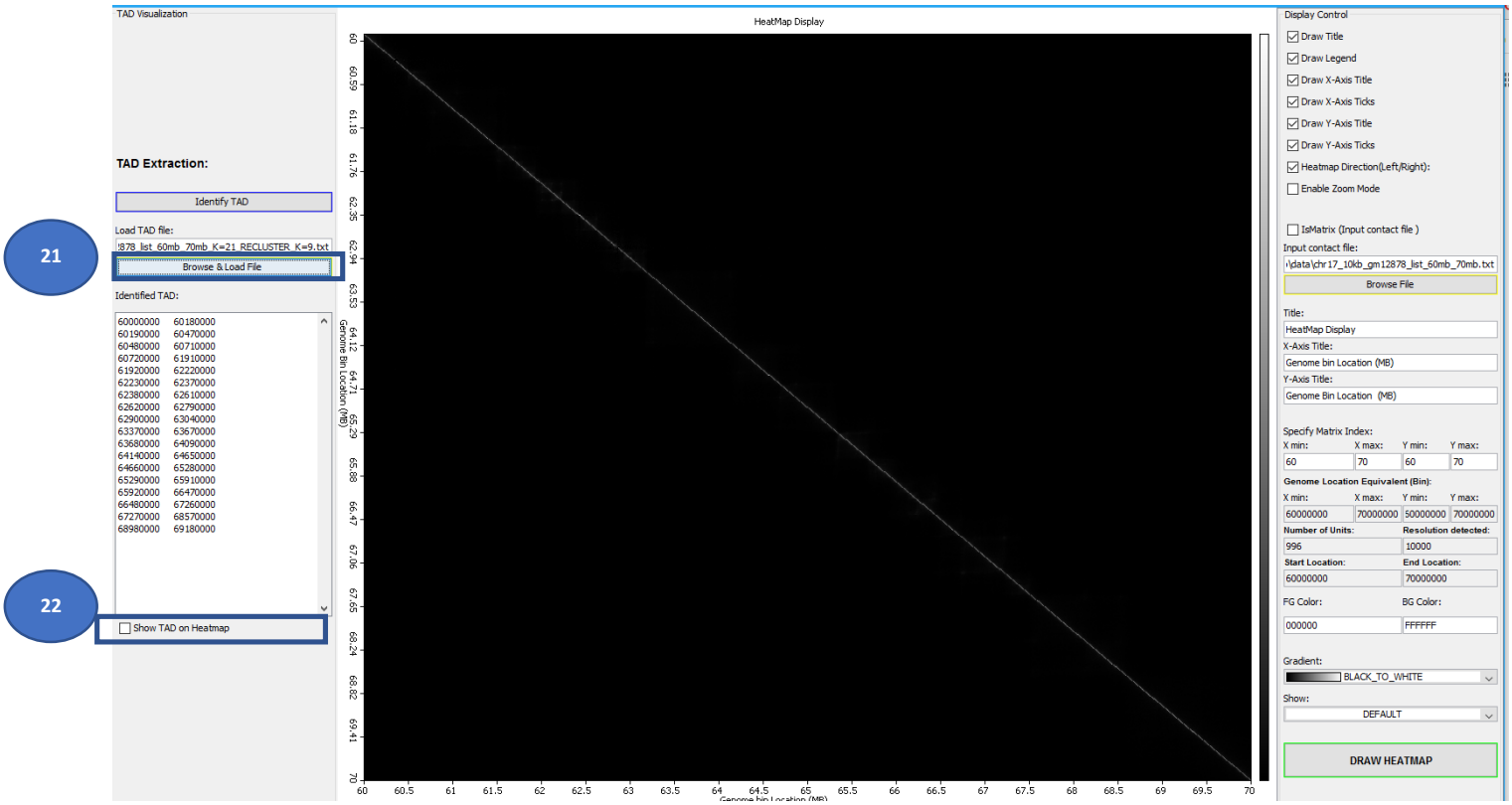
RAINBOW


HeatMap Display
Genome bin Location (MB)
60
60.59
61.18
61.76
62.35
62.94
63.53
64.12
64.71
65.29
65.88
66.47
67.06
67.65
68.24
68.82
69.41
70
60.48
60.95
61.43
61.9
62.38
62.86
63.33
63.81
64.29
64.76
65.24
65.71
66.19
66.67
67.14
67.62
68.1
68.57
69.05
69.52

RED_TO_GREEN


HeatMap Display
Genome bin Location (MB)
60
60.59
61.18
61.76
62.35
62.94
63.53
64.12
64.71
65.29
65.88
66.47
67.06
67.65
68.24
68.82
69.41
70
Genome bin Location (MB)
60
60.48
60.95
61.43
61.9
62.38
62.86
63.33
63.81
64.29
64.76
65.24
65.71
66.19
66.67
67.14
67.62
68.1
68.57
69.05
69.52
70

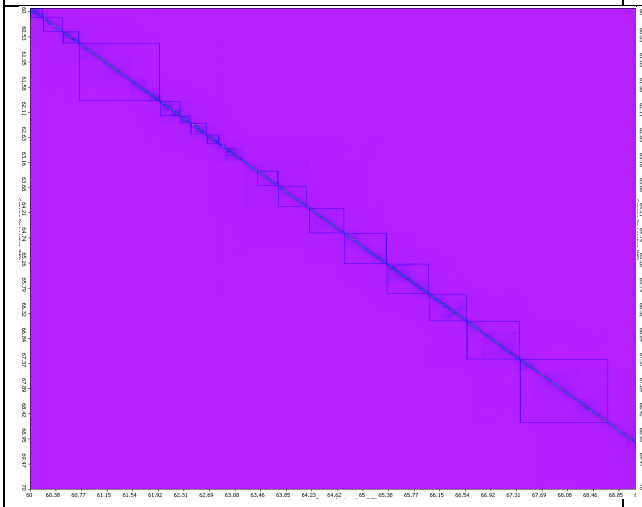
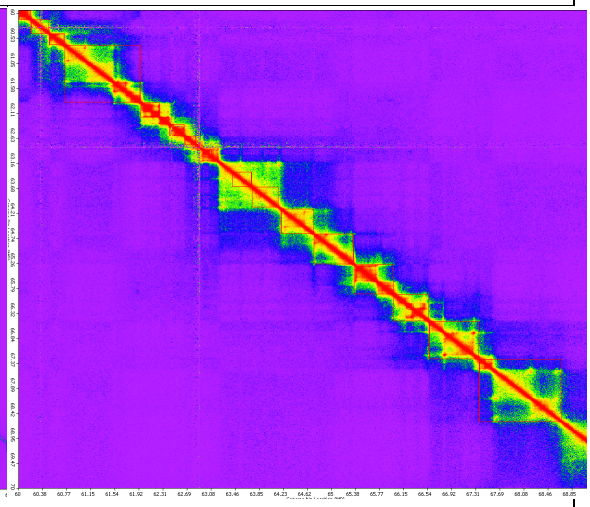
e) How to show TAD on the Heatmap?



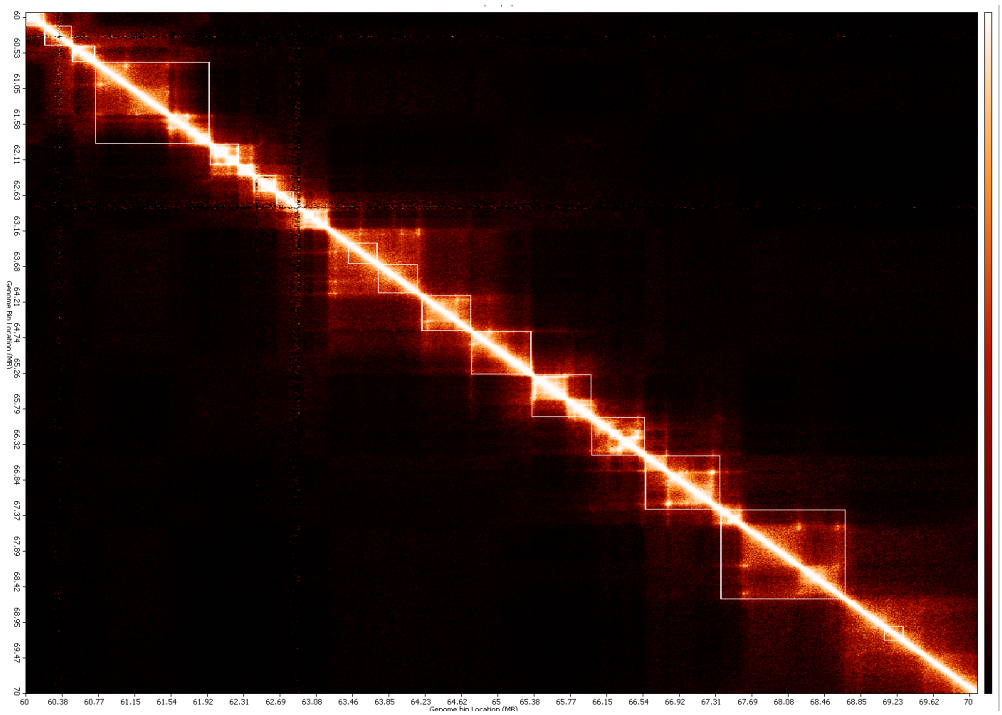
21. Click on the Browse and Load File button to load previously identified TAD

22. Check the Show TAD on Heatmap Check box to visualize the extracted TAD in the display window

23. The heat maps below shows the result obtainable when user clicks the Show TAD on Heatmap button.

Gradient = RAINBOW	Gradient = RAINBOW, Show = Tanh
	

Gradient = HEAT, Show = Tanh



24. The **squares** at the diagonal, are the TAD identified for the dataset.

4. Convert mapped Hi-C reads to hic format file

a) **Purpose**

To create a binary hic format file containing contact matrices at different resolutions and normalized by different methods from a text file describing mapped Hi-C reads

b) **Input file format**

Five formats are acceptable: short format, short format with score, medium format, long format and 4DN DCIC format. A sample file is executable/sample_data/GSM1551688_HIC143_merged_nodups.zip (unzip it before use)

- **Short format**

A whitespace separated file that contains, on each line

```
<str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment

If not using the restriction site file option, frag will be ignored, but please see above note on dummy values. readname and strand are also not currently stored within *.hic* files.

- **Short with score format**

This format is useful for reading in already processed files, e.g. those that have been already binned and/or normalized; this format can be easily used in conjunction with the -r flag to create a *.hic*file that contains a single resolution.

A whitespace separated file that contains, on each line

```
<str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2> <score>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment
- score = the score imputed to this read

  If not using the restriction site file option, frag will be ignored, but please see above note on dummy values. readname and strand are also not currently stored within *.hic* files.

- **Medium format**

A whitespace separated file that contains, on each line

```
<readname> <str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2> <mapq1> <mapq2>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment
- mapq = mapping quality score

If not using the restriction site file option, frag will be ignored, but please see above note on dummy values. If not using mapping quality filter, mapq will be ignored. readname and strand are also not currently stored within .hic files.

- **Long format**

The long format is used by Juicer and takes in directly the *merged_nodups.txt* file.

A whitespace separated file that contains, on each line

```
<str1> <chr1> <pos1> <frag1> <str2> <chr2> <pos2> <frag2> <mapq1> <cigar1> <sequence1> <mapq2> <cigar2> <sequence2> <readname1> <readname2>
```

- str = strand (0 for forward, anything else for reverse)
- chr = chromosome (must be a chromosome in the genome)
- pos = position
- frag = restriction site fragment
- mapq = mapping quality score
- cigar = cigar string as reported by aligner
- sequence = DNA sequence

If not using the restriction site file option, frag will be ignored, but please see above note on dummy values. If not using mapping quality filter, mapq will be ignored. readname, strand, cigar, and sequence are also not currently stored within *.hic* files.

- **4DN DCIC format**

A file that follows the 4DN DCIC format specification (the 4DN DCIC format specification).
See the link for more information. Briefly, there should be a header with the first seven columns reserved:

```
## pairs format v1.0

#columns: readID chr1 position1 chr2 position2 strand1 strand2
```

If the columns line contains (in any field after field 7) both frag1 and frag2, those will also be read in; otherwise Pre will set frag1=0 and frag2=1, so that no reads are discarded. Other fields are ignored.

c) **Output**

A binary .hic file containing contact matrices

d) **Running**

Access the function from the menu toolbar: *2D-Functions/Convert to HiC*

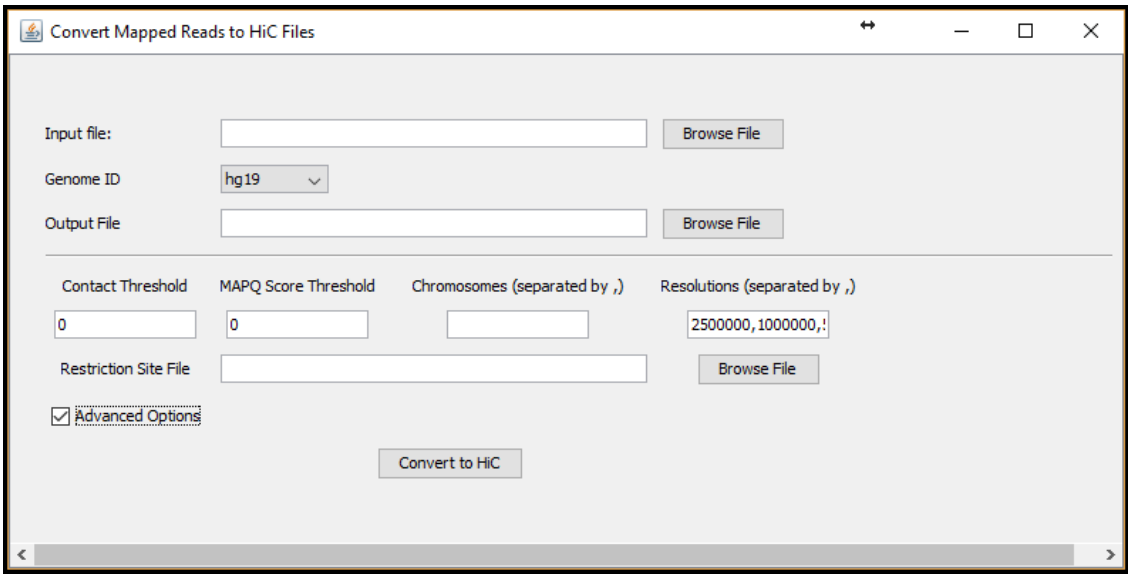


Figure 1 Convert to HiC function

Field	Description	Default
Input file	A text file describes mapped Hi-C reads (format described above)	NA
Genome ID	Version genome of Hi-C data	hg19
Output File	A name of the output hic format file	NA
Contact Threshold	Number of interaction threshold for contacts to be used in creating contact matrices.	0
MAPQ Score Threshold	Mapping quality score threshold for reads to be considered in creating contact matrices.	0
Chromosomes	Chromosomes for which their contact matrices to be created. When left blank, all chromosomes will be considered. Chromosomes must be separated by a comma (,).	All (when left blank)
Resolutions	List of resolutions of contact matrices to be created. Resolutions are separated by a comma (,)	2500000, 1000000, 500000, 250000, 100000, 50000, 25000,10000,5000
Restriction Site File	Each line starts with a chromosome number followed by positions of restriction sites on that chromosome, in numeric	blank

	order, and ending with the size of the chromosome. When provided, 8 additional fragment-delimited resolutions are added: 500f, 250f, 100f, 50f, 20f, 5f, 2f, 1f	

5. Extract contact matrices from a hic format

a) **Purpose**

To extract a contact matrix from a hic format into a sparse matrix format in a text file

b) **Input**

A local path to a hic format or an online link to a hic format. A link to a hic file: https://www.encodeproject.org/files/ENCFF219YOB/@@download/ENCFF219YOB.hic

c) **Output**

A contact matrix in sparse matrix format (each line represents a contact by three numbers separated by whitespaces: position1 postion2 interaction_frequency)

d) **Running**

Access the function from the menu toolbar: 2D-Functions/Extract HiC

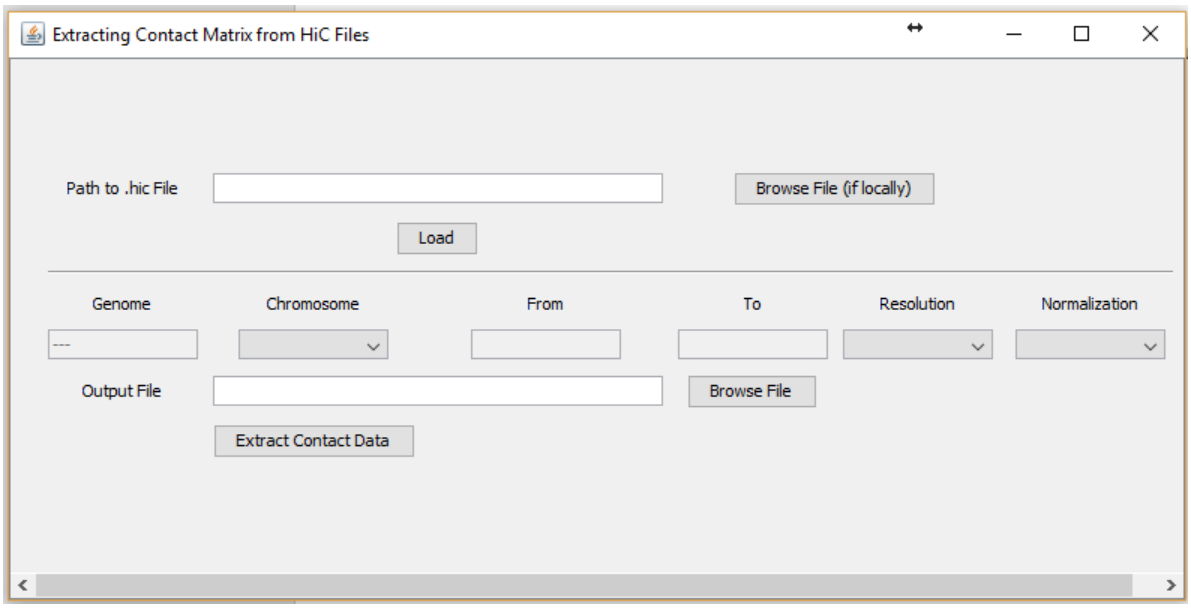


Figure 2 Extract Contact Matrices from a hic file

Field / Button	**Description**	**Default**
Path to .hic File	An online link or local path to a hic format file	NA
Load	Click this button to fetch information from the header of the hic file.	NA
Genome	Genome version of the hic file	NA
Chromosomes	List of resolutions of contact matrices in the hic file	NA
From	Start of a fragment (to extract its contact matrix). When From and To are left blank, the whole chromosome is considered.	Blank
To	End of a fragment (to extract its contact matrix). When From and To are left blank, the whole chromosome is considered.	Blank
Resolution	List of resolutions of contact matrices in the hic file	NA
Normalization	List of normalization methods used to normalize contact matrices	NA
Extract Contac Data	Click this button to initiate extracting contact data	NA

6. Normalize HiC contact matrices

a) **Purpose**

To normalize contact matrices in sparse matrix format.

b) **Input**

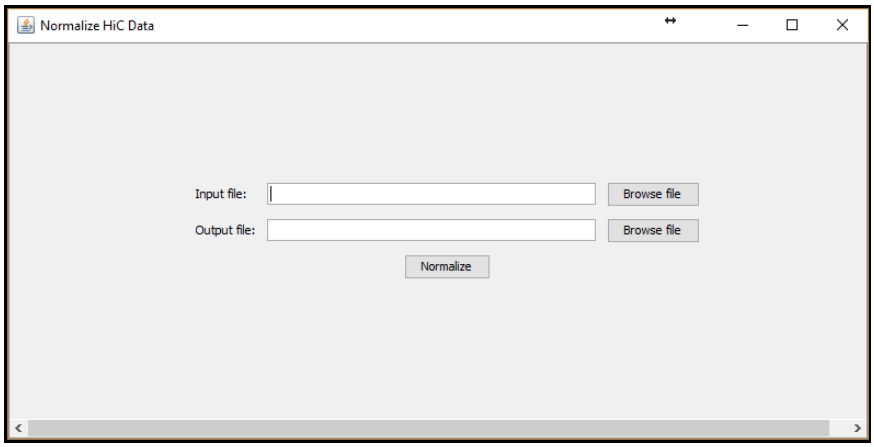
A contact matrix in sparse matrix format (each line represents a contact by three numbers separated by whitespaces: position1 position2 interaction_frequency)

c) **Output**

A normalized contact matrix in sparse matrix format. The matrix is normalized by the Iterative Correction and Eigenvector decomposition (ICE) method

d) **Running**:

Access the function from the menu toolbar: 2D-Functions/Normalized HiC Data



7. 3D model reconstruction by LorDG

a) **Purpose**

To build 3D chromosomes and genome models

b) **Input**

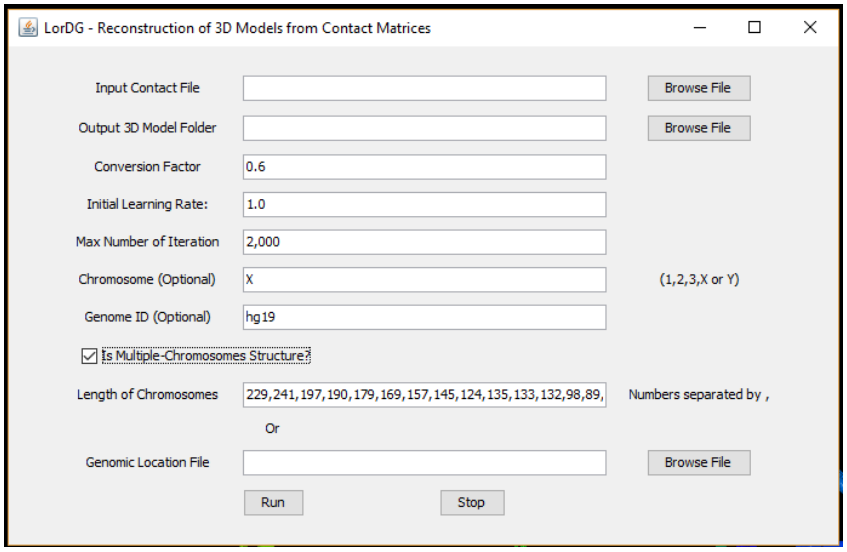
A contact matrix in sparse matrix format

c) **Output**

3D models in .gss format file

d) **Running**

Access the function from the menu toolbar: 3D-Functions/LorDG-3D Modeller



Field /Button	Description	Default
Conversion Factor	α in the formula $d_{ij} = \frac{1}{IF_{ij}^{\alpha}}$, where IF_{ij} is interaction frequency between i and j. When the field is left blank, the program will search for the best value in the range [0.1-3.0] with a step size of 0.1. Users can also specify a range to search by put 2 numbers separated by a hyphen (e.g. 0.5-1.0). During the searching, the right-top corner of the main screen displays information about the current value being tested.	1.0
Initial Learning Rate	Initial learning rate of the optimization. Higher learning rate can speed up the reconstruction process but can cause the process to fail as well	1.0
Max number of Iteration	Maximum number of iterations for the optimization	1000

Chromosome	Chromosome name of the contact matrix in the input. If the input contains contact matrix of the whole genome, leave this field blank.	X
Genome ID	Genome version of the contact matrix in the input.	hg19
Is Multiple-Chromosomes Structure?	if the input contains both inter-and intra-chromosomal contacts data, this checkbox should be checked.	unchecked
Length of Chromosomes	This field contains a list of lengths of chromosomes in increasing order of chromosome names and separated by commas, if "Is Multiple-Chromosomes Structure" is checked. Please note that these lengths should not contain omitted regions (e.g. centromeres) in the input of chromosomes.	
Run	To start the reconstruction process. The main screen displays how models are being formed from initially random models. The information about the reconstruction is displayed in the top-right corner of the main screen. The conversion factor is being used to build model and the current value of the objective function (higher is better). After the reconstruction is finished, the score of the model is displayed in the top-right corner of the main screen (the lower the value is, the better the model is).	NA
Stop	During the reconstruction, if this button is pressed, the program will stop and output the currently best structure. If the program is searching for the best conversion factor, it will stop the searching and use the best-found conversion factor to build models.	NA

8. Chromatin loop identification

a) **Purpose**

To identify chromatin loop in 3D models

b) **Input**

A 3D model to visualize

c) **Output**

A list of chromatin loops in a bed format file (optional) and highlighted in the 3D model

d) **Running**

Access the function from the menu toolbar: 3D-Functions/Loop Detection

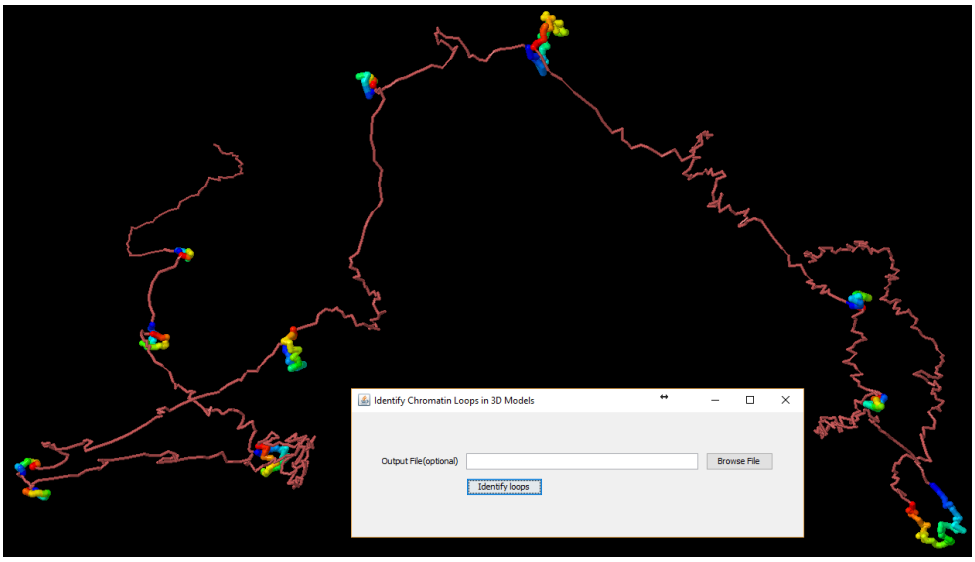


Figure 3 Chromatin loops

The function identifies chromatin loops and highlights them in the 3D model. The loops can also be outputted into a bed format file specified in the Output File field. The top-right corner of the main screen displays the number of chromatin loops identified.

Loops are colored in spectrum (from blue to red). To highlight loops better, color the model by a single color (right-click on the main screen, choose color/structure/chain)

9. Model annotation

a) **Purpose**

To annotate 3D models with genomic elements

b) **Input**

A 3D model (e.g. in executable/sample_data/models) and genomic elements in bed format files (e.g. in executable/track_files)

c) **Output**

3D model is annotated with data from bed format files

d) **Running**

Access the function from the menu toolbar: 3D-Functions/Model Annotation

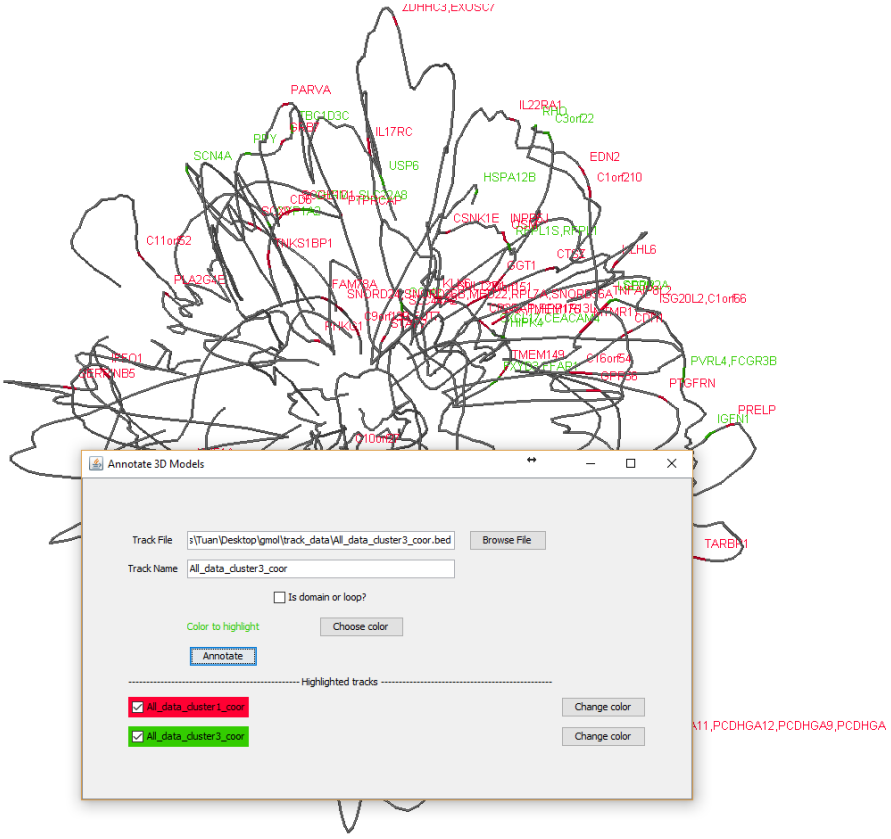


Figure 4 Function to annotate 3D models

To better highlight track data, change the color of the model to a sing color (right-click on the main screen, Color/Structure/Reset). The background can be changed to white to (Color/Background/White)

Field / Button	Description	Default
Track file	A file in bed format (see executable/track_files for example) to annotate the model	NA
Track name	A unique name associated with the above input file	Name of track file

Is domain or loop?	Indicate if the track file contains domains or loops. Adjacent domains/loops will be colored in red/blue alternatively.	Unchecked
Choose color	To pick a color to label annotation and points overlapped by genomic elements in the track file.	Random
Change color	To change color of the corresponding track	NA
	Checking corresponding track names will display or hidden the content of tracks.	

To get the genomic coordinate of a point, left-click or mouse-over to the point as shown in **Figure 5.**

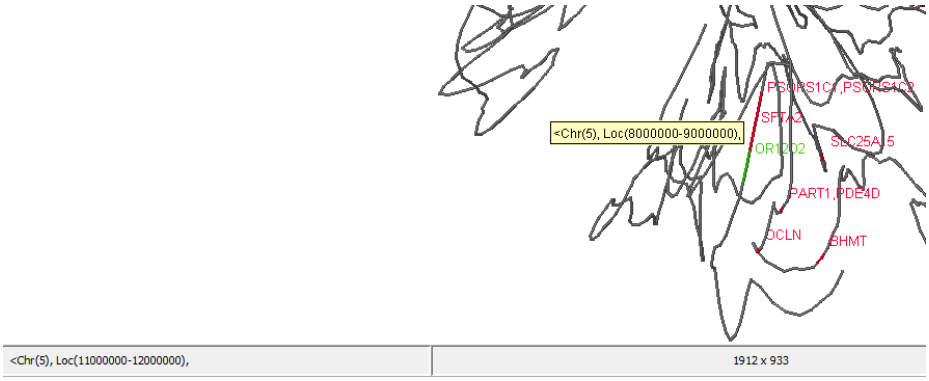


Figure 5 Coordinate of a point in the model

10. Gene expression data visualization (a special case of model annotation)

a) **Purpose**

To display gene expression level along a 3D model

b) **Input**

- A 3D model in GSS format (e.g. in executable/sample_data/models/ chr11_10kb_gm12878_list_60mb_70mb_1514493462531.gss) to visualize,
- A gene expression data file in GCT format (http://software.broadinstitute.org/cancer/software/genepattern/file-formats-guide#GCT), an example file is executable/sample_data/gene_expression/allaml.dataset.gct .
- And a text file to specify genomic coordinates of probes/genes in the GCT format file (each line consists of 4 elements separated by space or tab, e.g.:

probe_or_gene_name chr_number start end). A sample is executable/sample_data/gene_expression/probe_coordinates.txt

These 3 following files are prepared for demo: executable/sample_data/models/ chr11_10kb_gm12878_list_60mb_70mb_1514493462531.gss, executable/sample_data/gene_expression/allaml.dataset.gct and executable/sample_data/gene_expression/ probe_coordinates.txt.

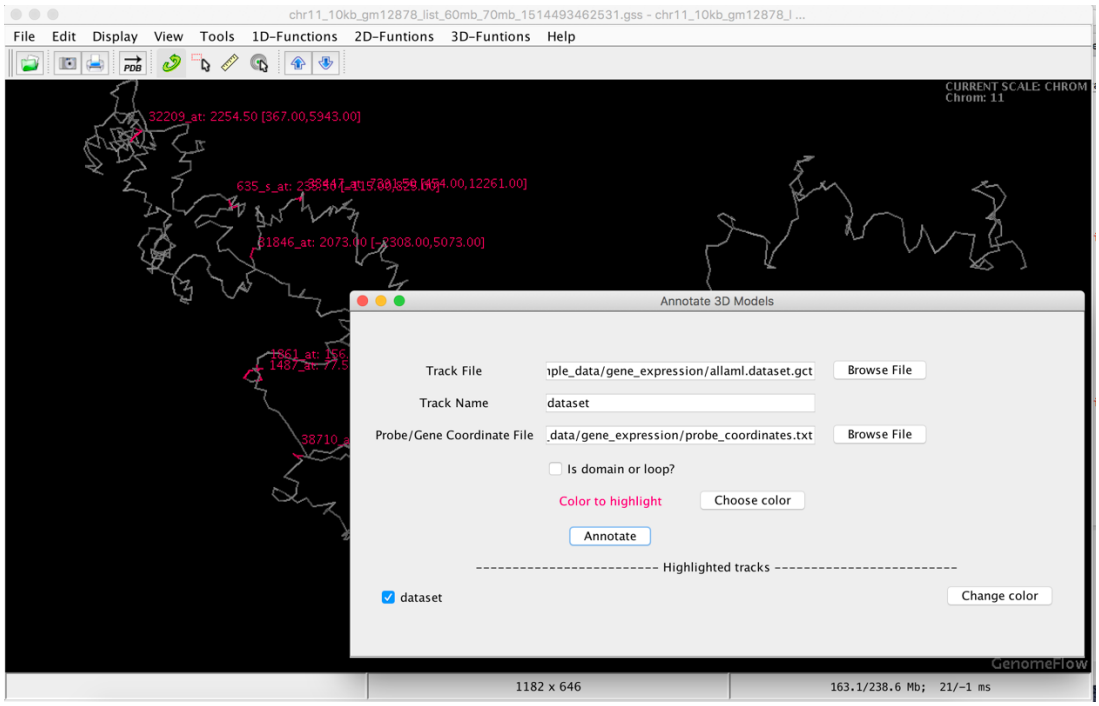
c) **Output**

Expression levels of genes/probes are annotated in the 3D model. Usually, the GCT file contains several samples and therefore, the median value (across all samples) together with minimum and maximum values (in brackets) are displayed next to probe/gene names.

If the 3D model and the gene expression data file have no overlap, no annotation will be added to the 3D model.

d) **Running**

Access the function from the menu toolbar: 3D-Functions/Model Annotation. A GCT file must be filled in the “Track File” field.



11. Comparing 2 models

a) **Purpose**

To superimpose and compare two 3D-models in GSS format.

b) **Input**

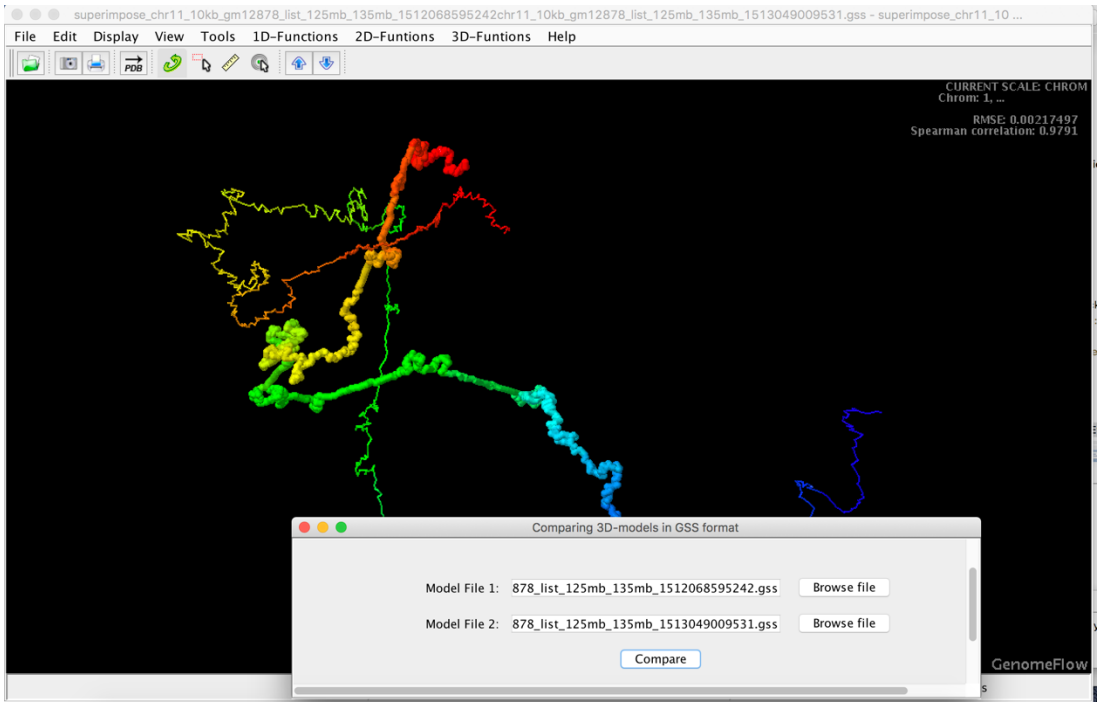
Two chromosome models in GSS format.

c) **Output**

The two models are scaled, superimposed and visualized. Spearman's correlation and RMSE between pairwise distances of the two models are calculated.

d) **Running**

Access the function from the menu toolbar: 3D-Functions/Compare Models


superimpose_chr11_10kb_gm12878_list_125mb_135mb_1512068595242chr11_10kb_gm12878_list_125mb_135mb_1513049009531.gss - superimpose_chr11_10 ...
File
Edit
Display
View
Tools
1D-Functions
2D-Funtions
3D-Funtions
Help
CURRENT SCALE: CHROM
Chrom: 1, ...
RMSE: 0.00217497
Spearman correlation: 0.9791
Comparing 3D-models in GSS format
Model File 1:
878_list_125mb_135mb_1512068595242.gss
Browse file
Model File 2:
878_list_125mb_135mb_1513049009531.gss
Browse file
Compare
GenomeFlow