

Cluster Maps Builder

[Part of NucTools 1.0 (2016)]

User manual

Yevhen Vainshtein and Vladimir B. Teif

Updated version available at <http://generegulation.info>

Address support queries to y.vainshtein@zmbh.uni-heidelberg.de

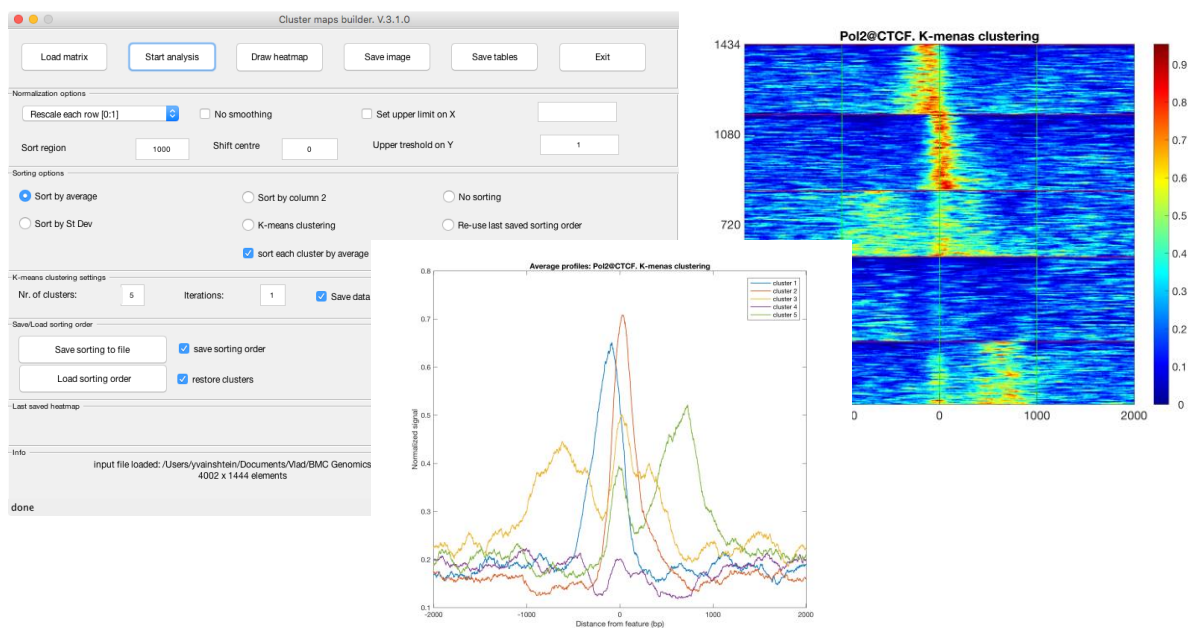


Table of Contents

Introduction	2
Package content	2
Cluster maps builder GUI	3
Buttons panel	4
Normalization options panel	5
Sorting options panel	7
K-means clustering settings panel.....	9
Save/Load sorting order panel	9
CMB graphical output	10
Known issues	11
References.....	12

Disclaimer

Cluster Maps Builder (CMB) is part of NucTools 1.0 release. At the moment it is at constant development and therefore may have some instability and bugs. If you find new bugs please contact y.vainshtein@zmbh.uni-heidelberg.de

Introduction

Cluster Maps Builder (CMB) is primarily designed to visualize nucleosome occupancy profiles of thousands of features aligned at genomic coordinate corresponding to a specific feature, like transcription factor binding site or transcription/translation initiation or termination site, using a heatmap representation. CMB includes a K-means clustering step and is able to apply the sorting/clustering order from initial matrix to a different matrix of the same size and dimensions.

CMB is written using MATLAB and is using Java-based GUI (GUIDE). At the moment the prerequisite of CMB's usage is availability of MATLAB installation. The initial development was done using MATLAB 2014b but the program was tested for compatibility with 2015a/b and 2016a. CMB has been tested for Windows and MacOS X. It has not been tested for Linux yet.

Package content

The CMB package consists of the following scripts:

Script name	Description
heatmap_builder.m	Main script of a CMB package, containing all functions evaluating interface calls and performing calculations. <i>Copyright Yevhen Vainshtein, Vladimir Teif</i>
heatmap_builder.fig	GUIDED user interface

Scripts published at MathWorks file exchange:

Script name	Description
heatmap.m	Displays a matrix as a heatmap image <i>Copyright 2014 The MathWorks, Inc.</i>
nanmean.m	Returns the sample mean of X, treating NAs as missing values <i>Copyright 1993-2004 The MathWorks, Inc</i>
smoothc.m	Smooths a 2D matrix using a cosine taper function. <i>Author: Linda Winkler</i>
progressbar.m	progressbar provides an indication of the progress of some task using graphics and text <i>Author: Steve Hoelzer</i>

statusbar.m	statusbar set/get the status-bar of Matlab desktop or a figure <i>Author: Yair M. Altman</i>
--------------------	---

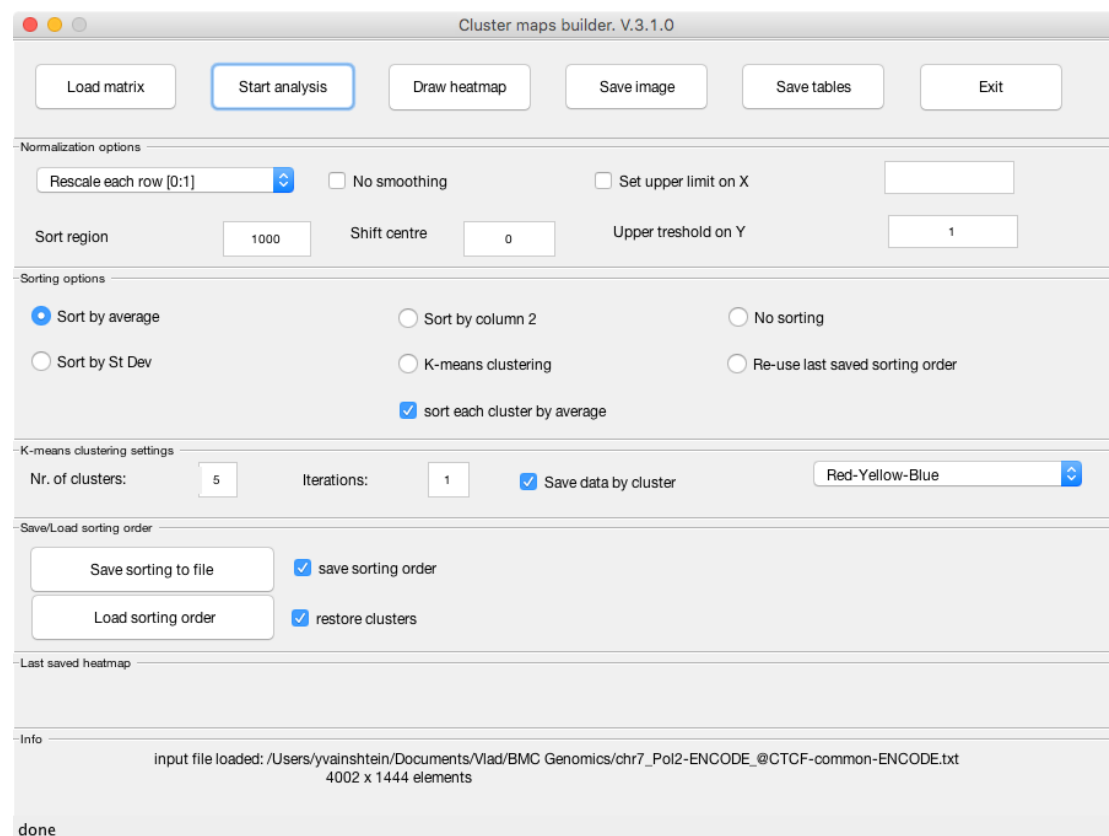
Perl scripts

Script name	Description
countlines.pl	Return Nr. of lines of the input text table
file_size.pl	Return file size in MBs

Additional files (MATLAB variable storage files):

MyBlueColormap.mat clusters_order.mat
MyRedColormap.mat sorting_order.mat

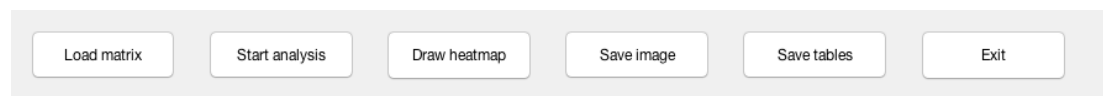
Cluster maps builder GUI



The GUI consists of 6 major panels (from top to bottom), and a status bar:

- Top buttons panel
- Normalization options panel
- Sorting options panel
- K-means clustering settings panel
- Sorting order backup panel
- Info panel

Buttons panel



“Load matrix” – opens the standard “Open File” dialog. The proper input file for the CMB application is a tab-delimited text file, containing normalized occupancy values for features aligned at defined genomic region (the output of **aggregate_profile.pl** script from NucTools package). CMB accepts any tab-delimited text file without or with a header of the following type:

		Distance to a feature start or center (bp)						
Feature ID	Sorting order (expression)	-100	-99	...	0	...	+99	+100
column 1	column 2	column 3	column 4					
ID1	-10.98	3	2.008		0.00012		0.22	0.45
ID2	0.8765	1.9018	1.022		0.001		0.00012	0
...								

Note:

Column 2 is optional. One can use it, for example, to provide gene expression values or any other arbitrary score. These values can be used to sort the data matrix accordingly for heatmap representation.

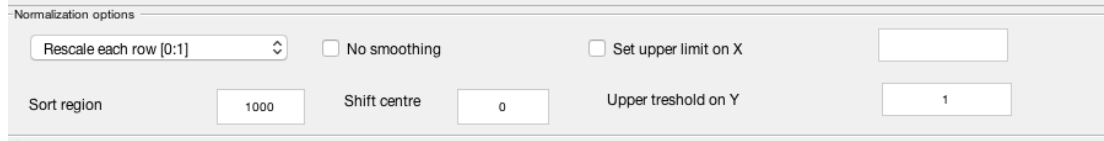
“Start analysis” – Performs normalization, rescaling, sorting or K-means clustering of the data using analysis settings defined in corresponding panels below and draws a cumulative occupancy profile using the mean of all values in each column after data normalization or rescaling (for details see below in the section “normalization options”)

“Draw heatmap” – After the analysis is finished and the data matrix is sorted according to one of the criteria defined in the “sorting options” panel, one can draw a heatmap. When prompted, specify a heat map name. It will be used as an image title as well as an image file name. After heatmap figure is generated it is saved automatically in the same folder as input matrix. In the case of a K-means clustering the aggregated profiles for each cluster will be generated and saved together with the heatmap.

“Save image” – Opens standard system “Save file as” dialog. This allows you to save the heatmap image at any location.

“Save tables” – Opens standard system “Save file as” dialog. The program will save a file containing Feature ID, mean of the occupancy in the sort region and cluster ID if applicable. As well, the original matrix with features aligned at specific genomic regions will be saved according to clusters and sorting. Such tables could be used again with the CMD to perform further analysis of selected clusters.

Normalization options panel



Normalization options

Rescale each row [0:1] ☐ No smoothing ☐ Set upper limit on X

Sort region: 1000 Shift centre: 0 Upper threshold on Y: 1

The “Normalization options” panel allows changing analysis settings related to the initial data treatment before sorting or K-means clustering.

The “**Choose normalization method**” drop-down menu contains following options:

- “**Rescale complete matrix [0:1]**” – Finds a global minimum and global maximum among all values in the matrix and assign it to 0 and 1 correspondingly. Recalculates occupancy values in the matrix as follows:

$$New.Occupancy_{xy} = \frac{Old.Occupancy_{xy} - \min(Matrix)}{\max(Matrix) - \min(Matrix)}$$

where *Matrix* is the array of genomic occupancy values provided by the user.

- “**Rescale each row [0:1]**” – Finds a minimum and maximum among all values in the each row and assign it to 0 and 1 correspondingly. Rescales all values in the row *Y* from 0 to 1:

$$New.Occupancy_{xy} = \frac{Old.Occupancy_{xy} - \min(row_y)}{\max(row_y) - \min(row_y)}$$

where *row_y* is a vector of occupancy values of a given chromatin feature.

- “**Normalize each row to a maximum**” – normalizes the occupancy value in each row by the maximum value among all values in a given row:

$$New.Occupancy_{xy} = \frac{Old.Occupancy_{xy}}{\max(row_y)}$$

- “**Normalize each row to a global maximum**” – divide the occupancy value in each row by the maximum value among all values in the whole matrix:

$$New.Occupancy_{xy} = \frac{Old.Occupancy_{xy}}{\max(Matrix)}$$

- “**Normalize each row to a leftmost value**” – divide the occupancy value in each row by the leftmost occupancy value from the sort region for each row:

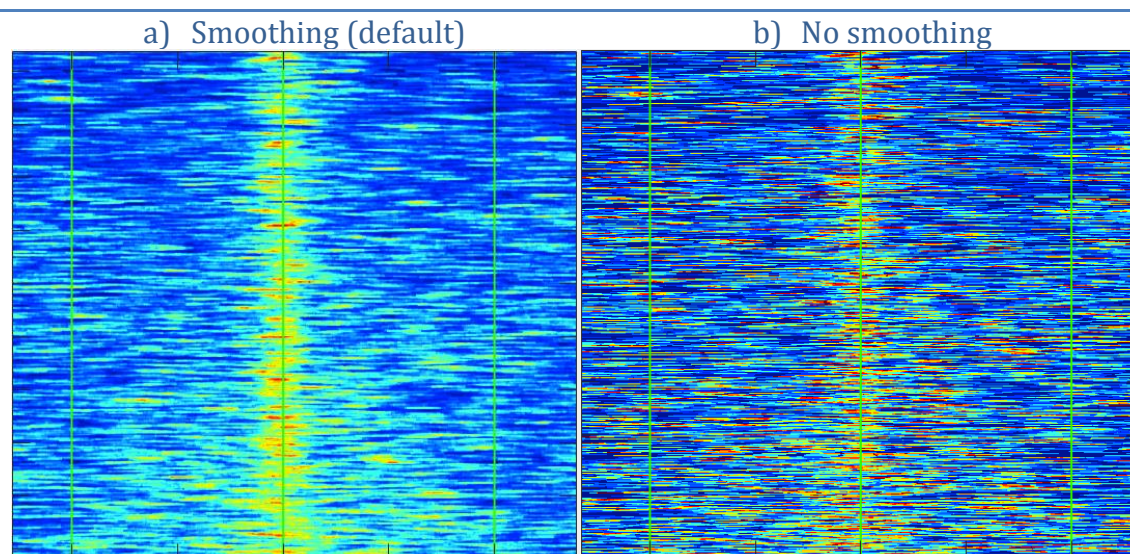
$$New.Occupancy_{xy} = \frac{Old.Occupancy_{xy}}{Old.Occupancy_{1y}}$$

- **“No normalization; Remove values above the threshold”** – reads the value from “Upper threshold on Y” and replaces all occupancy values above it with the threshold value (for example, remove outliers caused by piling-up of too many reads due to reads mapping artifacts).
- **“No normalization”** – processes data without any prior normalization.

The rest of options in the “Normalization options” panel can be divided into two categories: analysis settings and visualization settings.

Visualization settings: “no smoothing”

By default the “no smoothing” checkbox is deactivated and the 2D matrix is smoothed using a cosine taper function for better visualization:



Visualization settings: “Set upper limit on X”

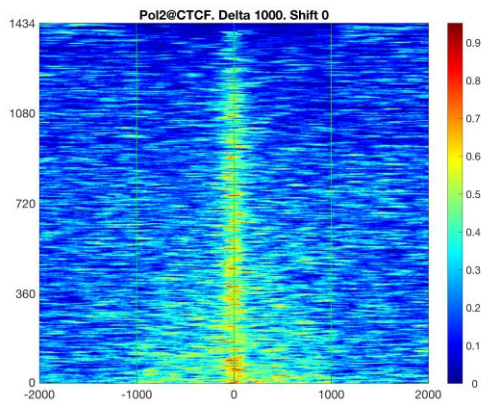
Limit the X axis when drawing average aggregated profile and per-cluster aggregated profiles. The data matrix itself is not changed and the heatmap visualization will be done for the whole data set.

Note: the X limit can be specified in the text field on the left from checkbox.

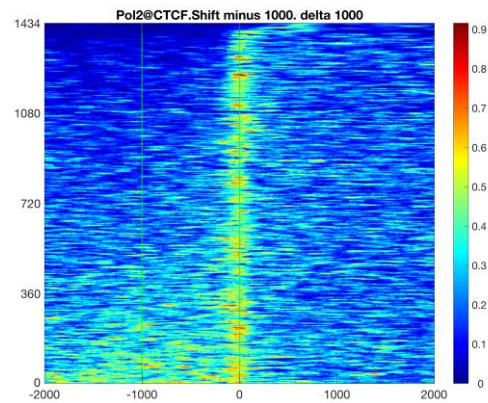
Analysis settings: “sort region” and “shift center”

These two options are a key to specify the coordinates for further analysis relative to the genomic feature. By default we assume the data is aligned and centered at the middle of the feature (e.g. TF binding site). The original data matrix could be spanning from several kbs downstream to several kbs upstream from the genomic region (limited only by the computer RAM and CPU). But we can also limit further analysis only to the region centered at the position [0+shift] and spanning from minus “sort region” to plus “sort region” to focus only on a specific data range. This data range is indicated on the heatmap with vertical lines at the center and at boundaries.

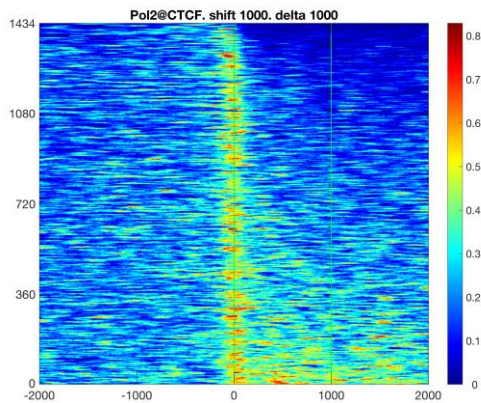
a) shift=0, sort region=1000



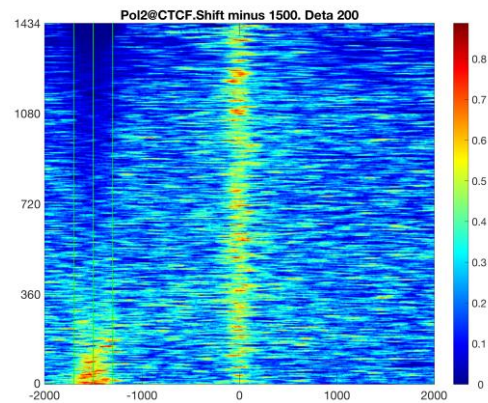
b) shift=-1000, sort region=1000



c) shift=1000, sort region=1000



d) shift=-1500, sort region=200



Sorting options panel

Sorting options

<input checked="" type="radio"/> Sort by average	<input type="radio"/> Sort by column 2	<input type="radio"/> No sorting
<input type="radio"/> Sort by St Dev	<input type="radio"/> K-means clustering	<input type="radio"/> Re-use last saved sorting order
<input checked="" type="checkbox"/> sort each cluster by average		

The “Sorting options” panel allows choosing the way the genomic regions shown on the heatmap will be sorted after normalization.

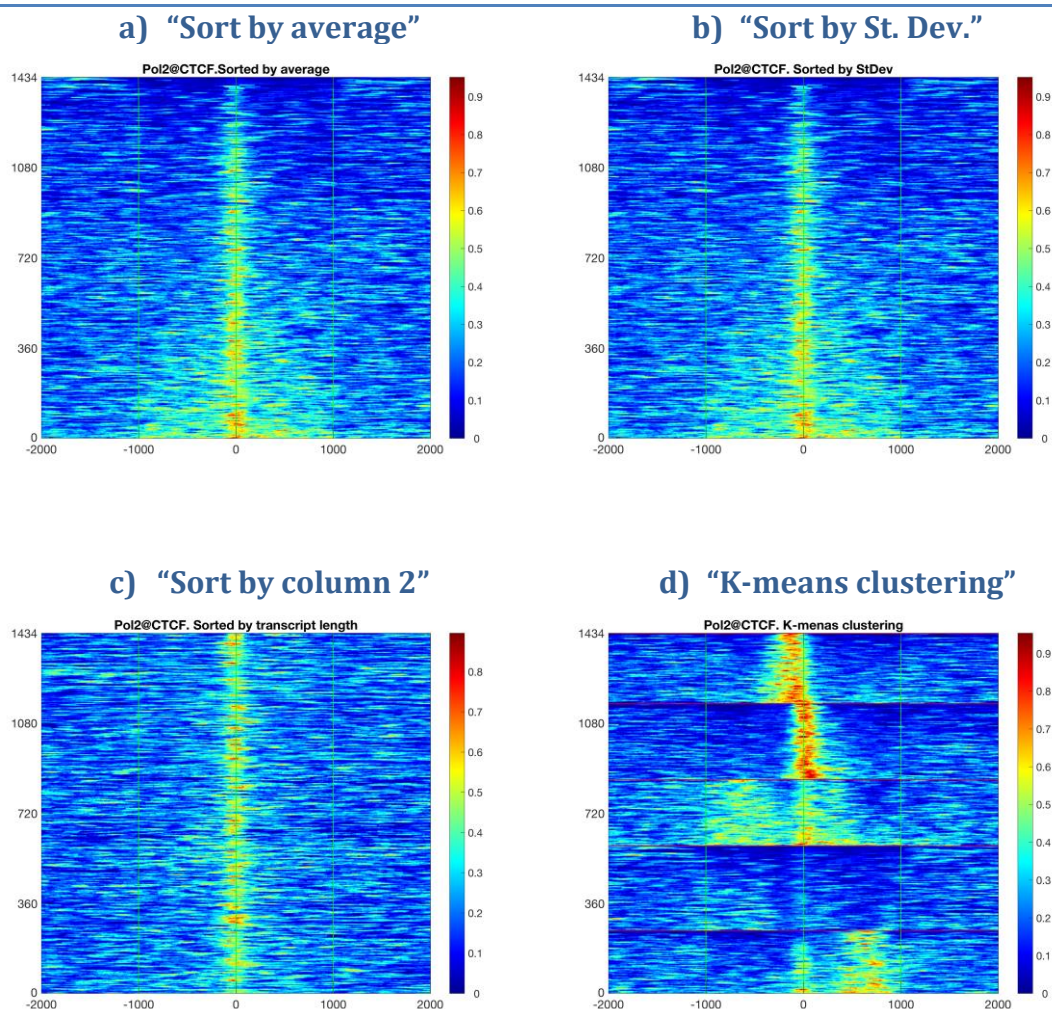
“Sort by average” – calculates the mean value in the “sort region” for each row Y and sort the matrix accordingly

“Sort by St. Dev.” – calculates the standard deviation (St. Dev.) value in the “sort region” for each row Y and sort the matrix accordingly

“Sort by column 2” – Uses the column 2 of original input table for sorting. By default, the “Aggregate_profile.pl” script from the NucTools package output the occupancy matrix leaded by the transcripts length.

“No sorting” – Do not change sorting of the original matrix

“K-means clustering” – perform K-mean clustering of the normalized/scaled data with the settings provided in the “K-means clustering settings” section.



Note: “Sort by average” and “Sort by ST. Dev.” ((a) and (b)) options very often produce similar results with nucleosomes positioning data aligned at TF binding site, because the variability in the data for each feature directly connected to the number of reads. “Sort by column 2” in panel (c) in this particular example is similar to the “no sorting” option (a), because the original matrix does not carry a transcript length information.

“Re-use last saved sorting order” – this sorting option allows applying clustering/sorting order achieved for one dataset to another dataset of the same size. This option is extremely useful when working with data series, for example studying changes in nucleosome patterns in cells of healthy/diseased/treated

patient, or differences in cell lines. Every time one presses the “Start analysis” button a new sorting order is created. Before the application is closed one can re-use this sorting with the same matrix or apply to another matrix.

K-means clustering settings panel

“Nr. of clusters” – specifies the number of expected clusters. The k-means clustering aims to partition all observations into k clusters in which each observation belongs to the cluster with the nearest mean.

“Iterations” – specifies the number of iterations for k-means algorithm. Increasing the number of iteration will produce more reproducible clusters.

Notes:

- *K-means algorithm is the simplest unsupervised learning algorithm and therefore always produces slightly different results because each time it starts from a random assignment of clusters and further optimization for each data point. Nevertheless, most of the time the core of each identified clusters will be the same if algorithm converges.*
- *K-means clustering on the big data sets can be very time consuming. Minimizing the “sort region” allow to decrease significantly clustering performance.*
- *The Nr. of clusters parameter gives only approximate number. If the K-means algorithm can’t converge with specified number of clusters, they will use lower number of clusters*

Save/Load sorting order panel

In order to preserve the sorting order for next analysis runs, one can save it to the file and load it back, pressing corresponding buttons.

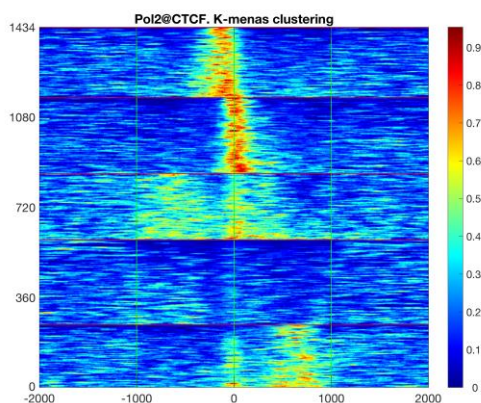
The checkbox **“save sorting order”** should always be selected in order to save the current analysis run.

The “restore clusters” checkbox instructs the program to restore not only sorting, but as the cluster order. If the saved sorting order was derived from the un-clustered dataset, please disable this checkbox to avoid error messages.

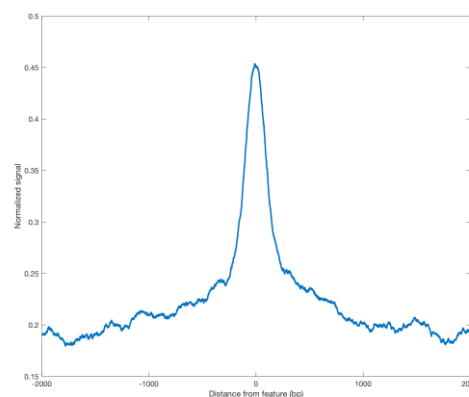
CMB graphical output

To illustrate the graphical output of CMB tool we are using the nucleosome density data around bound Pol2 in ESCs from a low-MNase MNase-seq dataset (Teif et al, 2014) around more than 100,000 sites of Pol2 enrichment in ESCs determined from CHIP-seq (mouse ENCODE). On the heat map, each horizontal line represents an individual genomic region containing Pol2 peak. A typical CMB output consists of heatmap itself, average aggregated profile plot and aggregated plots for individual clusters:

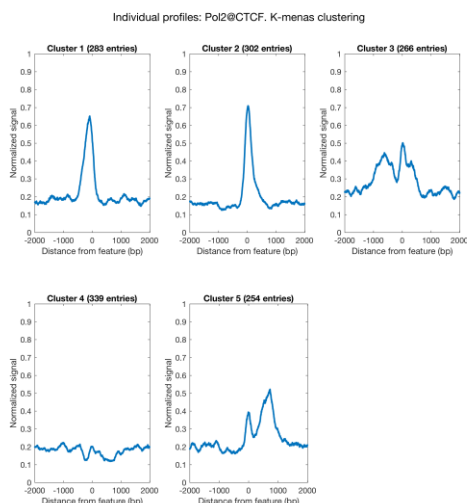
a) Heatmap



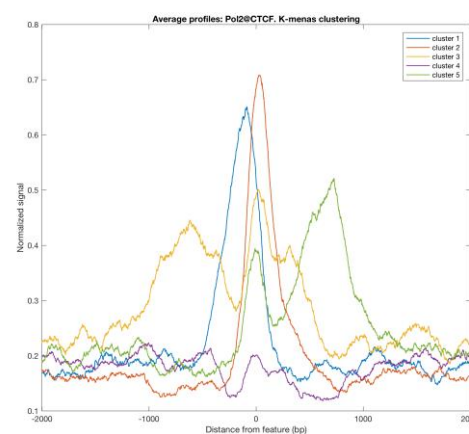
b) Aggregated average profile



c) Individual profiles (per cluster)



d) All profiles (per cluster)



Known issues

Symptoms:

GUI starts normally; the data is loaded, but after pressing the “Start analysis” button appears an error message.

Reason:

Closing the CMB application by pressing window (x) button instead of “Exit” button sometimes causing the problem upon next start.

The default CMB settings file “settings.mat” is corrupted.

Solution 1:

After GUI appears, reactivate all options – double click all checkboxes, re-enter all numeric fields. After that close application with “Exit” button.

After restarting a CMB tool everything should work fine.

Solution 2:

Close application, locate a “settings.mat” file in the CMB script directory and remove it. Restart the application.

Symptoms:

When loading the data table the error message “Wrong matrix file! Please use tab-delimited text tables with as minimum 2 columns and rows” pops-up.

Reason:

If you are loading 2D tab-delimited table with many columns but still see such message, the most probable reason is the wrong line endings in the text file (there is a operating system-specific difference in line ending of a text file)

Solution:

- Mac/Linux users: run a Perl one-liner replacing line endings in the Terminal session:

```
perl -pi -e 's/\r\r\n/\n/g' your_table.txt
```

```
perl -pi -e 's/\r\r\n/\n/g' your_table.txt
```

The command will replace original files with one with correct endings

- Windows users: load a table to Excel, remove last column and save the data using “Save As”->“Save as a Tab Delimited Text (*.txt)”

Citations: NucTools package is described in our recent manuscript [1] and has been used previously in publications [2-5].

References

1. Vainshtein Y, Rippe K, Teif VB: **NucTools: analysis of chromatin feature occupancy profiles from high-throughput sequencing data**. *Submitted* 2016.
2. Teif VB, Vainshtein Y, Caudron-Herger M, Mallm JP, Marth C, Höfer T, Rippe K: **Genome-wide nucleosome positioning during embryonic stem cell development**. *Nat Struct Mol Biol* 2012, **19**(11):1185-1192.
3. Teif VB, Erdel F, Beshnova DA, Vainshtein Y, Mallm JP, Rippe K: **Taking into account nucleosomes for predicting gene expression**. *Methods* 2013, **62**(1):26-38.
4. Teif VB, Beshnova DA, Vainshtein Y, Marth C, Mallm JP, Höfer T, Rippe K: **Nucleosome repositioning links DNA (de)methylation and differential CTCF binding during stem cell development**. *Genome Res* 2014, **24**(8):1285-1295.
5. Beshnova DA, Cherstvy AG, Vainshtein Y, Teif VB: **Regulation of the nucleosome repeat length in vivo by the DNA sequence, protein concentrations and long-range interactions**. *PLoS Comput Biol* 2014, **10**(7):e1003698.