M-Peak version 3.0

Release 01/2013

A model-based peak detection software for ChIP-chip data.

References:

Zheng, M., Barrera, L.O., Ren, B. and Wu, Y.N. (2007) ChIP-chip: Data, Model and Analysis. *Biometrics*, 2007, Vol 63, P.P. 787-796.

Kim, T.H., Barrera, L.O., Zheng, M., Qu, C., Singer, M.A., Richmand, T.A.,Wu, Y.N., Green, R.D., and Ren, B. (2005) A high-resolution map of active promoters in the human genome. *Nature*, 2005, Vol 436, P.P. 876-880.

M-Peak is free to download and the source code is open. The users are welcomed to contact us for free consulting.

Contact email: zmdl2000@gmail.com

****

Start using M-Peak:

This version of Mpeak was developed in Microsoft Visual C++ 2008 in Win7/WinXP environment. Currently only pc-executable version is provided and tested in Win7/WinXP operating system.

This software is a green software, i.e. after you download the software, it is directly ready for use without the need of installation. The software is just a single file plus a Readme file with instructions.

The input format for the software is .GFF format. Here is an example of .GFF format:

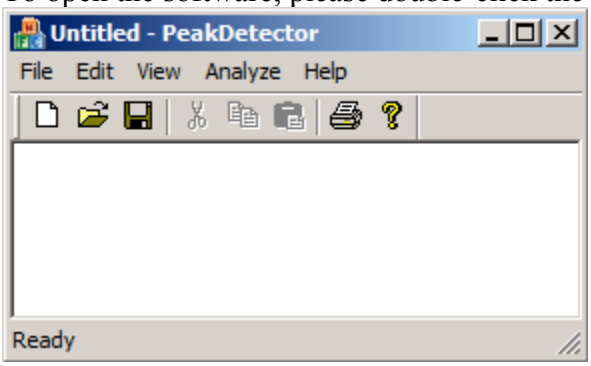
```
CHR11	NimbleScan	taf_Cy5_ENCODE/Total_Cy3_ENCODE	116190433	116190483	-0.24536755	.	.	CHR11:115994758-116494757;ENCO00P000000986;1
CHR11	NimbleScan	taf_Cy5_ENCODE/Total_Cy3_ENCODE	116190533	116190583	-0.166167469	.	.	CHR11:115994758-116494757;ENCO00P000000987;1
```

The first column is the name of the chromosome, and the third column is the feature label, and the forth column is the starting position of the probe in the sequence, and the

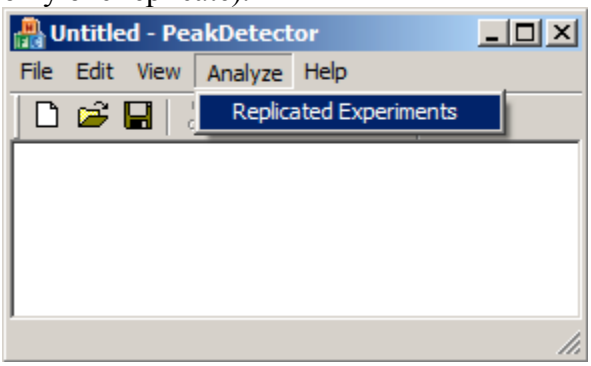
fifth column is the ending position of the probe, and the sixth column is the signal strength of that probe after pre-processing. The software only utilizes the information in these 5 columns, and the other four columns are reserved for other purposes. For more information about .GFF file, please refer to
http://genome.cse.ucsc.edu/FAQ/FAQformat#format3 and
http://www.sanger.ac.uk/Software/formats/GFF

(CAUTION: please avoid ANY space in the entries of the data. Otherwise the data may not be able to be loaded correctly.)

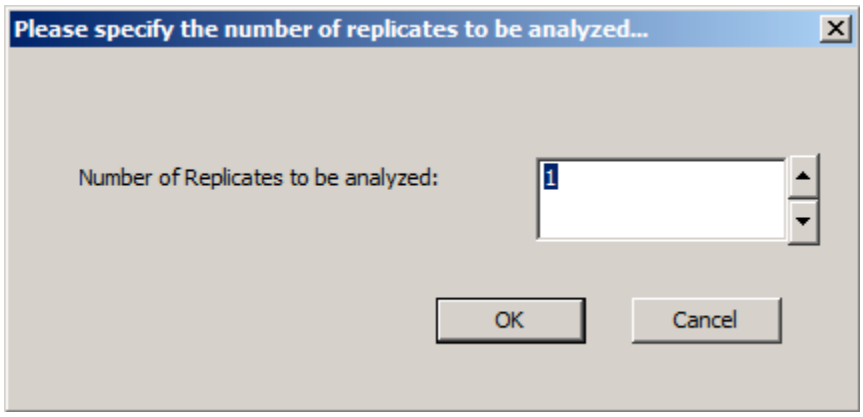
To open the software, please double-click the program to open the software interface:



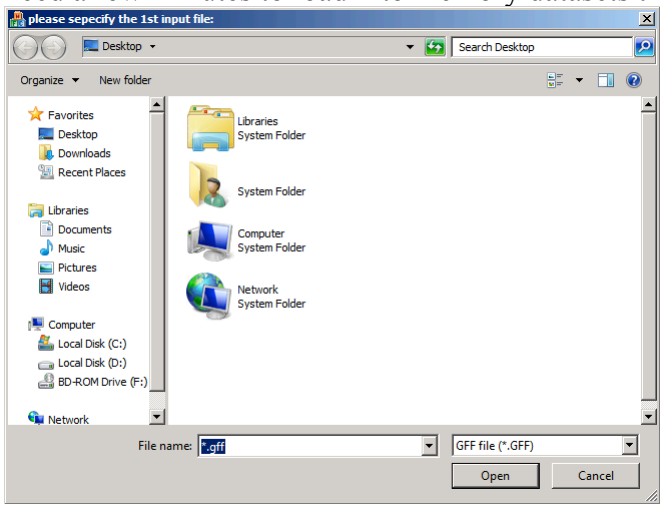
Then please click “Analyze” and then click “Replicated Experiments” (even if there is only one replicate).



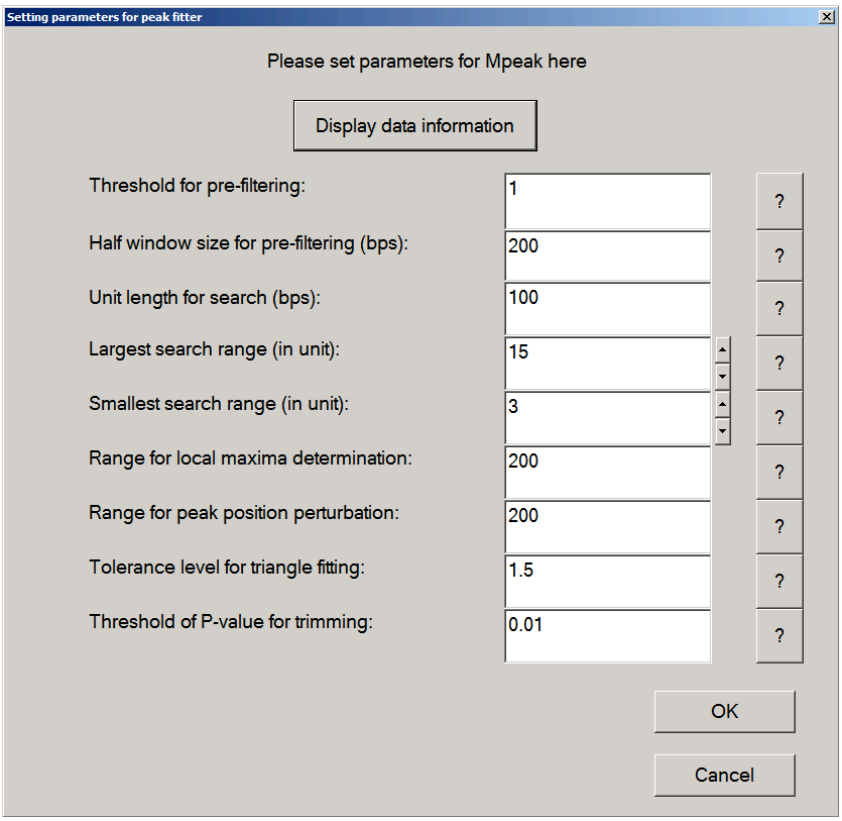
Next, please follow the on-screen instruction and input the number of replicates (non-replicated data will automatically have only one replicate.)



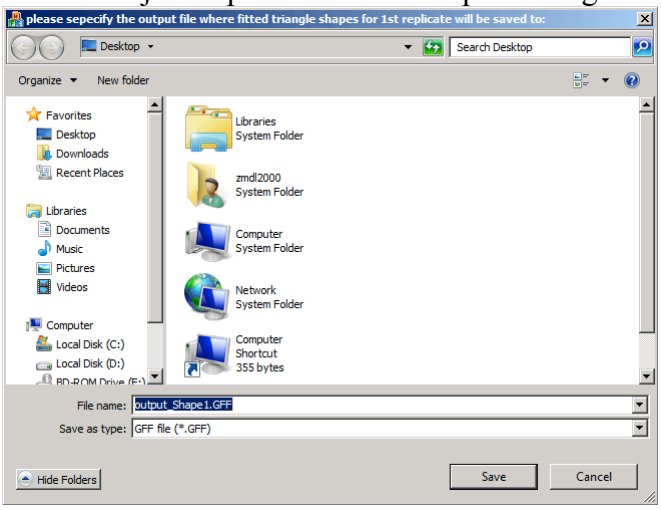
In the next dialog(s), please choose the target .GFF file(s) one-by-one. (The program may need a few minutes to load into memory datasets that are extremely large.)



After the data has been loaded, a dialog will appear and parameters for model fitting are set here. Details for setting these parameters are provided in the Appendix.



Finally, please follow the on-screen instruction to specify the output files (one GFF file for each input replicate to save the fitted shaped, one GFF file to save to peak detection results using all replicates and one text file to save additional information (primary the raw and adjusted p value after multiple-testing correction) on the detected peaks.



For more information, please contact the author Dr. Ming Zheng (zmdl2000@gmail.com).

The output files (except the p value result file) are still in .GFF format. The software itself cannot visualize the output. Currently, the SignalMap software developed by

NimbleGen Inc. is used for this purpose. For more information about SignalMap, please contact NimbleGen Inc. at www.nimblegen.org.

Appendix: Setting Parameters for Mpeak

The users are recommended to use the default values, but Mpeak does provide the users with the flexibility to change the values of the parameters. Mpeak can be divided into 3 steps during the peak detection process and the parameters involved in each step are described accordingly:

A.1 Pre-filtering to identify candidate peaks. First, probes with large differential hybridization signals are identified as local maxima. They are screened in a defined pre-filtering value within half windows adjacent on both sides of such local maxima. Those that meet the criteria will be considered as potential peaks. Parameters critical for initial peak selection are:

- Range for Local Maximum Determination (in base pair) – If the differential hybridization intensity of a probe is larger than any other probes within this range on either side, the subject probe will be considered as a potential peak. The default value is 200 bp.

- Threshold for Pre-filtering – Local maxima with intensity lower than this threshold are filtered (excluded from subsequent analysis). This is the most important parameter, with a predetermined P-value limit (below). It defines the stringency of the Mpeak peak fitting function. Setting a low number may include tiny peaks, which might be false-positives. Setting a high number may exclude true peaks in the results. The default value is mean+1*SD of the raw data. The "Display Data Information" function will provide preliminary statistics for the raw data as references. User should explore various values for pre-filtering setting, such as mean+1*SD, mean+1.5*SD, or increasing numerals, e.g. 1, 1.5, 2, 2.5, etc. for evaluation, and biological considerations of the results, so obtained.

- Half Window Size For Pre-filtering (in base pair) – This is the size of a half window, in which the intensities of the other probes of a local maximum will be analyzed with reference to the pre-filtering value, defined above. If no signal from the neighborhood of a probe defined by this half window is equal or higher than the pre-filtering value, the probe is considered as background noise and excluded from further consideration. We suggest using a number around half of the typical size of a DNA fragment. The default value is 200 bp.

A.2 Model fitting to fit triangular model to each candidate local maximum. For each candidate local maximum, the triangle model will try to fit a triangular shape with the local maximum at its peak. The range of the triangle model can vary, and the optimal range is selected on the goodness-of-fit by the standard deviation of the residuals from the regression model. Parameters relevant for this step are:

- Unit Length for Search (in base pair) – This number specifies the density of the search grid. If the unit length is defined as N, then the search range will be separated into N, 2N, 3N, etc.. Only triangles whose size is an integer multiple of N will be considered. A small number could slightly increase the precision of the fitted triangle and the peak position, but it will increase the computational burden. The default value is 100-bp.

- Largest Search Range (in unit) – User specifies the largest number of unit length for search, defined above. For the largest search range, it is recommended that 2-4 times the average length of DNA fragment should be used. For an average length of ~500 base pairs and 3 times such length, the largest search range will be 1,500 bp, or 15 x 100-bp unit. The default value is 15 units.

- Smallest Search Range (in unit) – User specifies the smallest number of unit length for search, defined above. This value should be chosen such that the average number of probes within the range specified by this number (i.e. the current probe position ± the Smallest Search Range * Unit Length for Search) should be reasonably large for the regression analysis to have a decent degree-of-freedom. Default value is 3, i.e. 300 bp, which roughly includes 1+3+3=7 probes in the Smallest Search Range when the spacing between probes is 100 base pair.

- Range for Peak Position Perturbation (in base pair) – For a given local maximum, the program searches a small neighborhood of this local maximum and marks the site with the best triangle model fit (measured by the SD of the residuals in the regression analysis) within the neighborhood, i.e. best fitting as a peak(s). This parameter specifies the radius of this neighborhood. A number between 200 to 500-bp is recommended. The default value is 200.

A.3 Post-Processing of Peaks. Smaller peaks detected in the screening could be either artifacts due to a nearby large peak or false positives due to background noise. Therefore, a post-processing statistical step is performed to exclude such artifacts or false-positive peaks. The parameters important for this post-processing step are:

- Tolerance Value of Triangle Fitting – This is used to determine the inhibition of a smaller peak by a nearby larger peak. A (smaller) peak, which is within the range of the fitted triangle of a nearby (larger) peak, is regarded as "explained by the nearby peak", if the fitted value of this probe (smaller peak) in the best fitted triangle of this probe (smaller peak) is smaller than that of the larger peak plus some amount. The amount is defined as the residual standard deviation of the fitted triangle of the nearby (larger) peak times the tolerance level, specified here. The larger/smaller tolerance value will lead to more/less number of small peaks being filtered (excluded). A value of 1 to 1.5 (i.e. 1 to 1.5 standard deviation) is generally acceptable.

- Threshold of P-value for trimming – Non-significant peaks with raw P-value larger than this threshold will be filtered (excluded). A threshold P-value of 0.01 is commonly accepted as a threshold for trimming.

Brief description of each of these parameters can be viewed by clicking the small button with a “?” sign on the right hand side of each parameter entry row.