Manual for BroadPeak

1. Preparation:

In order to run BroadPeak, you need to:

1) Download the compressed folder “BroadPeak.tar.gz”;
2) Decompress the folder. There are three files within the created folder: A) BroadPeak, B) unsupervised_estimation.R and C) Manual.
3) **Make sure these files are always kept in the same folder.**
4) Make sure R program is already installed on your computer.
5) **Check the shebang line of the BroadPeak file and correct it by the path of *env* of your computer.**
6) **Add the directory of the folder BroadPeak into the PATH.**

2. Running BroadPeak:

The command line for BroadPeak is:

$ BroadPeak -i [input bedGraph file] -m [identifier for output files] -b [bin size] -g [genome size] -t [type of parameter estimation] -r [BED file for supervised parameter estimation] -R [the directory of R]

The detailed explanations of the parameters can be found in Section 4.

One example of running BroadPeak is:
$ BroadPeak –i ./H3K36me3.bed –m H3K36me3 –t unsupervised

This command takes the sorted bedGraph format file of H3K36me3 ChIP-seq data as the input and use the unsupervised method for parameter estimation. A folder named as “H3K36me3” will be created (specified by -m) and all outputs will be in this folder. Other parameters use the default values.

3. File Format:

The input file of the sorted ChIP-seq read-mapping profile in the genome needs to be in bedGraph format. The four tab-delimited columns are chromosome, start, stop and tag count.

Note:
1) a BED format file of reads will not work. You need to first scan the genomic read-mapping to create the bedGraph file of ChIP-seq profiles (*i.e.* tag counts for each small genomic bin).
2) The size of each record (the bin) should be equal (*e.g.* 200bp, if you scan the genome by dividing it to 200bp non-overlapping bins).
3) The locations should be already sorted for each chromosome.

If you want to do supervised parameter estimation (*i.e.* -t supervised), you also need to provide (using -r) a BED format file of the genomic regions that are believed to be enriched with broad peaks (*e.g.* some highly expressed genes for broad-peak calling of H3K36me3).

The output file will be a BED format file of the genomic locations of broad peaks. In the output folder, there will be a folder named as **_broad_peak_**, the final BED output file is located in this folder.

4. Parameters:

-i: The bedGraph format input file (with the correct path) of the sorted ChIP-seq read-mapping profile in the genome.

-m: The identifier used to name the output folder, *e.g.* use H3K27me3 to name the output folder for broad-peak calling of H3K27me3.

-b: The size of bin, default value 200 (bp). It should be consistent with the bedGraph format input file (*i.e.* equal to the size of the bins in the input file).

-g: The size of the genome under consideration, default value 3107677273 (bp) for the human genome (hg18).

-t: The type of parameter estimation. It can be either "supervised" or "unsupervised".

-r: If **-t** is set as "supervised", a BED format file (with the correct path) of genomic regions used for supervised parameter estimation need to be given.

-R: If R program is not in the PATH, you can use -R to specify the directory of R program.

-h,-help: Display brief explanations of parameters.