SUPPLEMENTARY INFORMATION

A ChIP-Seq Data Analysis Pipeline Based on Bioconductor Packages

Seung-Jin Park[1,2], Jong-Hwan Kim[1,2], Byung-Ha Yoon[1,2], Seon-Young Kim[1,2]*

[1]Personalized Genomic Medicine Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Korea,
[2]Department of Functional Genomics, University of Science and Technology (UST), Daejeon 34113, Korea

http://www.genominfo.org/src/sm/gni-15-11-s001.pdf

Supplementary Methods

Calling of broad peaks (step 6 in broad peak script)

Broad peaks can be identified with the MOSAiCS-HMM model [1], using the 'mosaicsFitHMM' and 'mosaicsPeakHMM' methods in the 'mosaics' package. We use ChIP-Seq data of H3K27me3 in GSM733696. At first, we proceed to make a bin with fragment length 200 and bin size 200, like the process of calling sharp peaks on control and H3K27me3 chip. Next, read the bin file and use the 'mosaicsFit' function in the MOSAiCS model and MOSAiCS-HMM model to call the broad peak. Then, 'mosaicsPeakHMM' calls peaks based on the MOSAiCS-HMM model fit. 'mosaicsPeakHMM' allows two approaches—theViterbi algorithm and posterior decoding (default)—to call peaks. 'mosaicsPeakHMM' provides its output as a 'MosaicsPeak' class object, which 'mosaicsPeak' also generates. In particular, a post-processing step is necessarily required for calling broad peaks. Post-processing consists of two steps. First, find the peak summit for the broad peak, and adjust the peak boundaries to filter out potentially false positive peaks. For starters, use the 'extractReads' function to read the peak file and each bam file. Find the summit with the 'findSummit' function. The next step is to adjust the boundary of the peak to get better results and remove false positive peaks. Particularly, in the problem of adjusting the first peak boundary, post-processing is a very effective step for calling the broad peak of histone modifications. The 'adjustBoundary' method computes the read count of the ChIP sample against the control and then trims the boundary for high-quality results. The 'filterPeak' method removes potentially false positive peaks, given the peak summit region and peak length.

Reference

1. Chung D, Zhang Q, Keleş S. MOSAiCS-HMM: a model-based approach for detecting regions of histone modifications from ChIP-Seq data. In: *Statistical Analysis of Next Generation Sequencing Data* (Datta S, Nettleton D, eds.). New York: Springer, 2014. pp. 277-295.