Genome analysis

NTAP: for NimbleGen tiling array ChIP-chip data analysis

Kun He1,2,*, Xueyong Li3,4, Junli Zhou3,5, Xing-Wang Deng3, Hongyu Zhao6 and Jingchu Luo1,*

1 College of Life Sciences, Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, Peking University, Beijing 100871, China, 2Department of Plant Biology, Carnegie Institution, Stanford, CA 94305, 3Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, USA, 4Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, 5Beijing Kaituo DNA Biotech Research Center CO., Ltd., 39 West Shangdi Rd, Haidian District, Beijing 100085, China and 6Department of Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, USA

Received on January 24, 2009; revised on May 10, 2009; accepted on May 12, 2009

ABSTRACT

Summary: NTAP is designed to analyze ChIP-chip data generated by the NimbleGen tiling array platform and to accomplish various pattern recognition tasks that are useful especially for epigenetic studies. The modular design of NTAP makes the data processing highly customizable. Users can either use NTAP to perform the full process of NimbleGen tiling array data analysis, or choose post-processing modules in NTAP to analyze pre-processed epigenetic data generated by other platforms. The output of NTAP can be saved in standard GFF format files and visualized in GBrowse.

Availability and Implementation: The source code of NTAP is freely available at http://ntap.cbi.pku.edu.cn/. It is implemented in Perl and R and can be used on Linux, Mac and Windows platforms.

Contact: ntap@mail.cbi.pku.edu.cn; luojc@pku.edu.cn; hekun78@gmail.com

1 INTRODUCTION

Genome-level high-density tiling arrays are becoming more accessible for genome-wide profiling studies including transcriptional binding site identification (Bertone et al., 2004), transcription factor binding site identification (Lee et al., 2007), histone modification profiling (Gendrel et al., 2005; Li et al., 2008), DNA methylation profiling (Hayashi et al., 2007) and comparative genome hybridization. Specific analysis methods and tools are required for each type of study because the strategies behind different tiling array applications vary extensively. As a result, several models have been proposed and software tools have been developed for the analysis of different types of tiling array data (Chung et al., 2007; Ji et al., 2008; Li et al., 2005; Wang et al., 2006; Zhang et al., 2007). However, there is still room to improve for data analysis of epigenetic features including histone modifications and DNA methylation. The recognition of distribution patterns of modifications at both the local (gene) and global (chromosome) levels are usually required to infer biologically meaningful conclusions (Hayashi et al., 2007; Li et al., 2008).

Here, we present a NimbleGen Tiling array data Analysis Package (NTAP) designed for histone modification profiling analysis (Li et al., 2007). The advantage of our package is its ability to generate reports for various pattern recognition questions instead of focusing only on identifying significantly enriched oligos or genomic regions.

2 FUNCTIONS AND FEATURES

NTAP was developed using the R statistical language to take advantage of the powerful statistical functions of other open source packages especially those from the Bioconductor project (http://www.bioconductor.org/). It contains five main steps for data analysis: importing, normalization, feature identification, oligos mapping and post-processing for pattern recognition.

2.1 Data importing

We implemented an R function similar to the ‘read.maimages’ function in the limma package (Smyth 2004) to import NimbleGen raw data into limma data object formats for normalization.

2.2 Data normalization

Users can apply various microarray normalization methods to the imported datasets through the limma package functions ‘normalizeBetweenArrays’ and ‘normalizeWithinArrays’. Unlike the expression profiling arrays whose log transformed ratio distributions are usually symmetric around zero, the distribution of the ChIP-chip result tends to skew to the ChIP channel. Because only the protein-bound DNA fragments will be pulled down by a specific antibody, more positive log transformed ChIP/Input ratios are expected. Thus, the rank-invariant set scheme (Buck and Lieb, 2004) was incorporated for better data normalization.

2.3 Feature identification

Tiling arrays usually contain several oligos per single gene rather than one oligo per gene. For example, the traditional whole-genome array for expression studies in Arabidopsis thaliana usually contains only 23k oligos, while a customized whole-genome tiling array...
The following questions are frequently asked in epigenomics:

1. Is there a connection between specific histone modification levels and gene sizes, or gene expression levels? To answer these questions, we implemented several R functions to align genes, to calculate the average ChIP/Input intensity ratio of the oligos within sliding windows, and to plot the final results for different groups (Fig. 1).

2.5 Post-processing functions

The following questions are frequently asked in epigenomics research. What is the modification distribution pattern relative to genes and does it vary between different organs/tissues? To answer these questions, we implemented several R functions to align genes, to calculate the average ChIP/Input intensity ratio of the oligos within sliding windows, and to plot the final results for different groups (Fig. 1).

2.6 Result visualization

Quality control is a key step to guarantee the validity of the overall data analysis. An R function was implemented to calculate the raw intensities correlation coefficient between any pair of two replicates. MA-plots of array hybridization results are also generated in order to examine the intensity ratio (M) versus averaged intensity (A) to discover possible non-linear biases that require special normalization methods. After raw data processing, all the oligos are mapped back to the most up-to-date chromosomes and the ChIP/Input ratio value of each oligo can then be plotted along the chromosome. These values can be displayed either by a program within NTAP or they can be exported in the GFF format to be displayed in the Generic Genome Browser GBrowse (Stein et al., 2002).

3 IMPLEMENTATION

Most of the functions are implemented in the R statistical language (http://www.r-project.org/) and Perl. Users can also choose any other software to pre-process their data before using our post-processing modules.

ACKNOWLEDGEMENTS

We thank Dr Kate Dreher for providing critical comments.

Funding: NSFC (grants 90408015, 863: 2006AA02Z334); China high-tech platform; Monsanto Fellowship and the China Postdoctoral Program (to K.H.).

Conflict of Interest: none declared.

REFERENCES

K.He et al.


