BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 14 2009, pages 1838–1840
doi:10.1093/bioinformatics/btp320

Genome analysis

NTAP: for NimbleGen tiling array ChIP-chip data analysis

Kun He[1,2,*], Xueyong Li[3,4], Junli Zhou[3,5], Xing-Wang Deng[3], Hongyu Zhao[6] and Jingchu Luo[1,*]

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

Advance Access publication May 25, 2009

Associate Editor: John Quackenbush

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 transcriptome 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).

*To whom correspondence should be addressed.

Here, we present a NimbleGen Tiling array data Analysis Package (NTAP) designed for histone modification profiling analysis (Li *et al*., 2008) that can also be applied to other ChIP-chip data (Lee *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

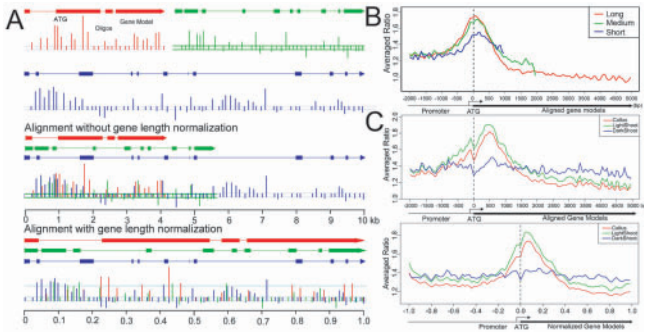

Fig. 1. Demonstration of two different methods for the alignment of gene models and reorganization of histone modification patterns. (A) Two different strategies to align genes (three genes with different lengths were used as examples). The alignment without gene length normalization overlapped all the oligos based on their absolute distance (kb) to the transcription start site while with length normalization based on the relative positions (percentile) to the transcription start site. (B) The histone modification distribution pattern between different user-defined gene sub-groups that contain various length genes in this particular case. (C) The tissue-specific histone modification distribution pattern on all genes by the two different strategies demonstrated in (A).

tiled at ~250 bp resolution may contain ~400k oligos. The much larger number of oligos on a single array makes the traditional methods for feature identification unfeasible. For tiling array data, expressed mRNA or pulled-down DNA fragments can cause the signal of a group of neighboring oligos to increase simultaneously. Therefore, our package implements the non-parametric Wilcoxon rank-sum method to compare the signal differences between the ChIP channel and the reference channel for a group of oligos using sliding windows. Under certain circumstances, however, the density of some tiling arrays may not be high enough to use the Wilcoxon method. In these cases, we utilize simple comparison linear models implemented in limma (Smyth 2004) to identify single oligos whose signal increased significantly in the ChIP channel. Then, we consider a genomic region as 'positive' if the region contains a single oligo that meets stringent user-defined criteria or the region contains a group of neighboring oligos that meet less stringent criteria.

2.4 Mapping oligos to gene models

Genome data are usually kept up-to-date by genome sequencing consortia or curation groups, who usually release their data as standard XML format files that can be parsed to easily obtain coordinates of gene models. A Perl module was implemented to retrieve records of the gene model position information on each chromosome and to determine the relative position of a specific oligo to its nearby gene model(s). Signal distribution patterns among different groups of genes can then be determined based on the stored relative position information.

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? Is there an association 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).

2.6 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

Bertone,P. *et al.* (2004) Global identification of human transcribed sequences with genome tiling arrays. *Science*, **306**, 2242-2246.

Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349-360.

Chung,H.R. *et al.* (2007) A physical model for tiling array analysis. *Bioinformatics*, **23**, i80-i86.

Gendrel,A.V. *et al.* (2005) Profiling histone modification patterns in plants using genomic tiling microarrays. *Nat. methods*, **2**, 213-218.

Hayashi,H. *et al.* (2007) High-resolution mapping of DNA methylation in human genome using oligonucleotide tiling array. *Hum. Genetics*, **120**, 701-711.

Ji,H. *et al.* (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nat. Biotech.*, **26**, 1293-1300.

Lee,J. *et al.* (2007) Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development. *Plant Cell*, **19**, 731-749.

Li,W. *et al.* (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21**(Suppl 1), i274-i282.

Li,X. *et al.* (2008) High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression. *Plant cell*, **20**, 259-276.

Smyth,G.K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**.

Stein,L.D. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

Wang,X *et al*. (2006) NMPP: a user-customized NimbleGen microarray data processing pipeline. *Bioinformatics*, **22**, 2955–2957.

Zhang,Z.D. *et al*. (2007) Tilescope: online analysis pipeline for high-density tiling microarray data. *Genobiology*, **8**, R81.