BIOINFORMATICS APPLICATIONS NOTE

Vol. 22 no. 23 2006, pages 2955–2957
doi:10.1093/bioinformatics/btl525

Gene expression

NMPP: a user-customized NimbleGen microarray data processing pipeline

Xiangfeng Wang[1,2,†], Hang He[3,†], Lei Li[2], Runsheng Chen[3], Xing Wang Deng[1,2,*] and Songgang Li[1,*]

[1]Center of Bioinformatics & Peking-Yale Joint Research Center of Plant Molecular Genetics and Agrobiotechnology, College of Life Sciences, Peking University, Beijing 100871, China, [2]Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, USA and [3]Bioinformatics Laboratory, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

Received on August 31, 2006; accepted on October 5, 2006
Advance Access publication October 12, 2006
Associate Editor: Alvis Brazma

ABSTRACT

Summary: NMPP package is a bundle of user-customized tools based on established algorithms and methods to process self-designed NimbleGen microarray data. It features a command-line-based integrative processing procedure that comprises five major functional components, namely the raw microarray data parsing and integrating module, the array spatial effect smoothing and visualization module, the probe-level multi-array normalization module, the gene expression intensity summarization module and the gene expression status inference module.

Availability: http://plantgenomics.biology.yale.edu/nmpp

Contact: Xiangfeng.Wang@Yale.edu

Supplementary information: The supplementary materials, figures, testing dataset, software instructions and a typical workflow of using NMPP package are accessible through the above homepage.

1 INTRODUCTION

High-density oligonucleotide microarrays produced by NimbleGen Systems using the Maskless Array Synthesis technology offer several competitive features, such as isothermal probe design, flexible probe lengths (24 to 70mer), and higher oligo capacity (390 000 per chip) and are gaining popularity among biologists to construct custom microarrays in many aspects of genomics study. Yet, there is still shortage of academic software for NimbleGen array users. Here, we report a user-customized NimbleGen Microarray data Processing Pipeline that integrates several modules and algorithms developed for dealing with high-density oligo microarray, typically the Affymetrix GeneChip arrays. NMPP is not only applicable to regular gene expression array, but also adaptable to other types of microarrays based on tiling-path design, for the purpose of raw data preprocessing.

2 FUNCTIONAL MODULES AND FEATURES

NMPP is composed of five major modules, and its workflow is designed to run in a mode of integrative processing that largely simplifies manual operation. In the initial step, the raw data from multiple microarrays are combined as a single input to the pipeline. The NMPP workflow then proceeds as follows: (1) to remove spatial effect across each array's surface; (2) to perform probe-level normalization across multiple arrays; (3) to summarize gene expression intensity from a matrix of multiple probes with replicate data by Tukey's median polish algorithm and (4) to estimate gene 'on/off' status, based on the summarized gene expression intensities and a remodeled global background distribution.

2.1 Raw data parsing and integrating module

NMPP workflow is initialized with the raw data parsing and integrating module that automatically collects probe intensities from all chips in accordance with a microarray design file and an experiment description file, and integrates them to a TAB-delimited table as the only input file to the pipeline. The subsequent modules will recognize each column as a set of data from the same array.

2.2 Array spatial effect smoothing module

The spatial effect across array surface (Fig. 1A) is the predominant within-slide experimental artifact that needs to be eliminated before any other normalization procedure. Different from spotted arrays whose spatial effect arises from print-tip bias, the spatial artifact observed on *in situ* synthesized microarray is mainly caused by uneven sample hybridization or washing and usually has gradient trend of distribution (more examples shown Supplementary material). We used a distance-weighted smoothing algorithm to resolve this problem, which is similar with the background correction procedure used in Affymetrix MAS 5.0. First, the array surface is segmented to $n \times n$ sub-squares ($n = 16$ is the default), and then, a set of $n \times n$ weight factors for each probe are computed based on the distances from its position to the centers of all sub-squares (Affymetrix, 2002) (http://www.affymetrix.com/). The following smoothing procedure uses a linear regression model:

$$\log_2[FG_i'(x,y)] = \log_2[FG_i(x,y)] + \frac{1}{\sum_{J=1}^{J} W_{i,j}(x,y)} \times \sum_{J=1}^{J} W_{i,j}(x,y)\{[\log_2(FG_{\text{baseline}}) - \log_2[M(FG_{z(j)})]]\},$$

[*]To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

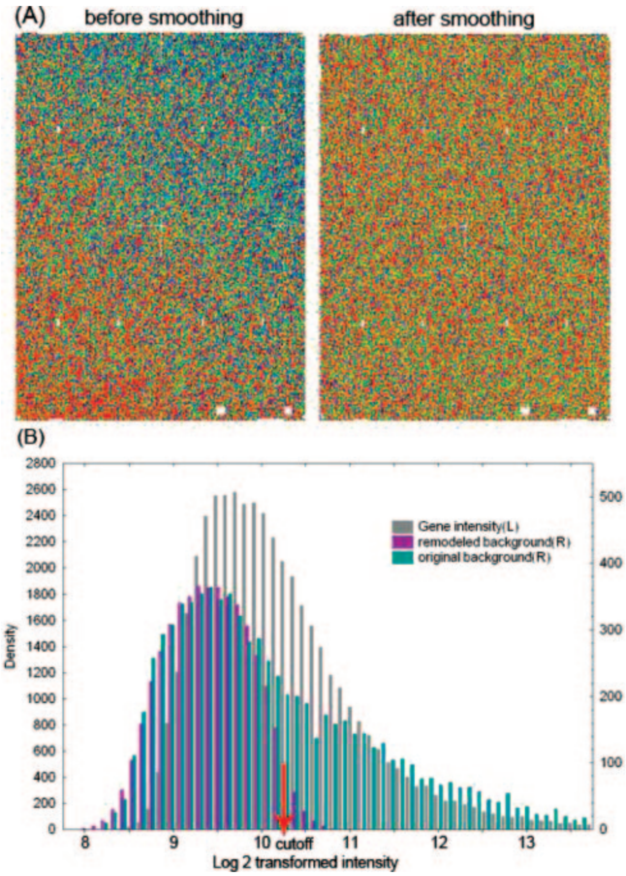

Fig. 1. (**A**) Visualization of probe signal distribution across array surface before and after array spatial effect smoothing procedure done. (**B**) Gene expression status inference method. Blue histogram is the original background distribution of negative control oligos; the purple histogram is a normal background distribution after remodeling and the red arrow indicates the cutoff at the 95% confidence level. Gray histogram is the distribution of summarized gene expression intensity.

where FG'_i and FG_i is the smoothed and original foreground intensity of oligo i, respectively. $M(FG_{z(j)})$is the median foreground intensity of zone j. FG_{baseline} is the baseline intensity and $W_{i,j}$ is the distance-weighted factor of oligo i to zone j. x and y is the coordinate of oligo i on the processed array.

2.3 Multiple arrays normalization module

NMPP uses probe-level quantile normalization to remove between-slide variations, which has been proven to perform the best on single-channel Affymetrix GeneChip data (Bolstad *et al*., 2003; Irizarry *et al*., 2003a). NMPP provides optional one-step or two-step normalization procedure based on user's choice. In the two-step procedure, NMPP first performs quantile normalization among technical replicate arrays, and then perform scaling normalization across the various tissue types or biological samples. In our online materials, we demonstrate that using the two-step normalization is more appropriate than the one-step procedure under some circumstances.

2.4 Gene expression intensity summarization module

NMPP provides a module of summarizing a gene's expression intensity from a probe set, for the need of custom multi-probe design (typically between 5 and 20 probes per gene) of NimbleGen microarray. The algorithm we used in this module is similar with RMA (Robust Multi-Array Average) algorithm (Irizarry *et al*., 2003b), which is based on a linear additive model assuming that the observed fluorescence intensity O_{ij} of oligo i on array j, is the sum of real expression value E_i on array i, plus probe affinity effect a_j of oligo j, and plus experimental errors ε_{ij}. To estimate the gene's expression value E_i, Tukey's median polish procedure (Tukey, 1977) is performed to calculate a grand median from a matrix of $n \times m$ data points, where n is the probe number of the target gene and m is the number of replicate experiments.

2.5 Gene expression status inference module

A typical gene expression microarray usually contains mainly positive features that interrogate genes of interest with a few of 'spiked-in' probes as a control for normalization and differential expression detection. To better characterize the impact of background noise, Irizarry *et al*. (2003a) have developed a mathematics model to compute a common mean background level for each array as a way of background correction for Affymetrix GeneChip. Motivated by this idea, we proposed a more direct way to construct a global background noise distribution by including a large number of control probes, taking advantage of the high capacity of the NimbleGen arrays. We recommend that the control probes have similar thermal prosperities as the interrogating probes and are enriched with probes reflect transcriptional background (details described in Supplementary materials). Based on a common assumption that the background noise is normally distributed while the foreground signal is exponentially distributed (Irizarry *et al*., 2003) NMPP first makes a transformation of the original distribution of the control oligos to exclude the false negative control who exhibit high hybridization intensities. By this means, a normal distribution of background noise is constructed and then a threshold at 95% confidence level is calculated for each given biological sample to determine a gene's transcription activity is 'on' or 'off' (Fig. 1B). The details of the remodeling and the threshold determination algorithms have been well described in our previous publication (Li *et al*., 2006) and on the NMPP homepage as well.

3 IMPLEMENTATION

To meet the demand of high throughput calculation of dozens or even hundreds of high-density oligo microarrays, NMPP was designed to have a command-line-based interface, and support the batch processing of all chips on computation server installed with either Windows or Linux system. All the functional components in the NMPP package were implemented as Perl modules and need to be installed in Perl library before properly running. Visualization of array surface spatial effect needs SVG package for Perl library, and Adobe SVG viewer to display the generated images. We also provide two compiled versions of NMPP package that can be executed independently from Perl interpreter. The source code of the NMPP modules for the purpose of self-modification in accordance with user's custom designed microarray on the NimbleGen platform is available on the NMPP homepage.

ACKNOWLEDGEMENTS

Development of NMPP package was supported in part by a grant from the NSF Plant Genome Program (DBI-0421675) to X.W.D.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2002) *Statistical Algorithms Description Document*. Technical report, Affymetrix, Santa Clara, CA.

Bolstad,B. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Irizarry,R.A. *et al*. (2003a) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Irizarry,R.A. *et al*. (2003b) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Res*., **31**, e15.

Li,L. *et al*. (2006) Genome-wide transcription analyses in rice using tiling microarrays. *Nat. Genet*., **38**, 124–129.

Tukey,J. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA.