

Efficient data processing method for large-scale cDNA microarray analysis

Koji Kadota^{1 2} **Yasushi Okazaki**¹ **Hidemasa Bono**¹
kadota@bi.a.u-tokyo.ac.jp okazaki@rtc.riken.go.jp bono@rtc.riken.go.jp
Rika Miki¹ **Yoshihide Hayashizaki**¹ **Kentaro Shimizu**²
rika@rtc.riken.go.jp yoshihide@rtc.riken.go.jp shimizu@bi.a.u-tokyo.ac.jp

¹ Laboratory for Genome Exploration Research Group, RIKEN Genomic Sciences Center, RIKE(GSC) and Genome Science Laboratory, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan.

² Department of Biotechnology, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan.

cDNA microarray analyses enable us not only to monitor massive gene expression but also to practically diagnose disease [1]. But there are few studies for measuring their reproducibility. Since the experimental procedure consists of multi-steps, there are several factors that can affect the reproducibility of the final results [2], it is necessary to omit genes with certain problem for clarity. Ascent of correlation coefficient (**R**) gives a good index on the reproducibility of the duplicated experiment at the same condition. However, it may also be given descent in terms of the total number of genes (**N**) that survived from a certain filtering program.

Thus, we have developed an efficient data processing method for cDNA microarray and applied for our massive data. The procedure of this filtering program consists of three steps: (1) omit the spots which have flags (flags are built manually when the spot image does not fulfill a certain criteria), (2) eliminate spots whose signal intensity is less than $\mu_{bg} + \mathbf{x} \times \sigma_{bg}$ in both Cy-3 and Cy-5, (3) eliminate spots that are located outside the best fit line (least-mean squares) $\pm \mathbf{y} \times \sigma$. Here, abbreviations are following: μ_{bg} , average background signal intensity at the experiment; σ_{bg} , standard deviation of the background signal intensity; \mathbf{x} and \mathbf{y} , parameter. These two parameters are automatically chosen so that it should have higher reproducibility on each pair.

To evaluate the efficiency of our filtering program, the data from tissue expression profiling using RIKEN full-length mouse 19K set were used for the analysis. Hierarchical complete linkage clustering of expression profiles showed appropriate results. The germ layer such as endoderm, mesoderm, or ectoderm could be adequately classified when using our filtering program, concluding the importance of adequate data pre-processing for the microarray data analysis.

References

- [1] Alizadeh, A. A., Eisen, M. B., Davis, R. E., *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.
- [2] Schuchhardt, J., Beule, D., Malik, A., *et al.* Normalization strategies for cDNA microarrays. *Nuc. Acids Res.*, 28, E47–E47, 2000.