Protocol

Calculation of Absolute Expression Values for DNA Microarray Data

Kazuro Shimokawa, Rimantas Kodzius, Yonehiro Matsumura, and Yoshihide Hayashizaki

This protocol was adapted from “Methods for Increasing the Utility of Microarray Data,” Chapter 6, in *DNA Microarrays* (ed. Schena). Scion Publishing Ltd., Bloxham, UK, 2007.

INTRODUCTION

In terms of cost per measurement, the use of DNA microarrays for comprehensive and quantitative expression measurements is vastly superior to other methods such as Northern blotting or quantitative reverse transcriptase polymerase chain reaction (QRT-PCR). However, the output values of DNA microarrays are not always highly reliable or accurate compared with other techniques, and the output data sometimes consist of measurements of relative expression (treated sample vs. untreated) rather than absolute expression values as desired. In effect, some measurements from some laboratories do not represent absolute expression values (such as the number of transcripts) and as such are experimentally deficient. To address the problem that some microarray data sets fail to reflect the number of mRNA molecules sufficiently in a given sample (i.e., fail to provide absolute expression levels), additional methods are required. The procedure described here provides a new method for converting microarray data to absolute expression values with the use of external data such as expressed sequence tags (ESTs) and cap analysis of gene expression (CAGE) tags.

RELATED INFORMATION

Superior microarray data produce superior results. Input the highest-quality microarray data possible, taking care to manufacture and hybridize the microarrays using the most rigorous scientific procedures. Poorly printed microarrays and low-quality samples produce inferior raw data, which will negatively affect the downstream computational processes. Superior robotics, printing technology, surface chemistry, target and probe preparation, and other molecular aspects produce superior data for analysis. See also **Calculation of Spot Reliability Evaluation Scores (SRED) for DNA Microarray Data** (this issue), which addresses the problem of absence of accurate measurements in DNA microarrays.

MATERIALS

CAUTIONS AND RECIPES: Please see Appendices for appropriate handling of materials marked with <I>, and recipes for reagents marked with <R>.

Equipment

Personal computer running Windows 2000 (or newer version) with at least an Intel Pentium IV CPU, 3.6 GHz processor, and 800 MHz front side bus

METHOD

1. Establish a bioinformatic link between external absolute expression data such as CAGE, EST, or serial analysis of gene expression (SAGE) tags and each microarray spot or element.
This can be done using genome mapping (see Discussion).
2. Given the annotations established in Step 1, calculate absolute expression values using the equations:

$$S_{\text{Array_relative}}^{(TU_x)} = \log_2 \left[\frac{S_{1\text{Array}}(TU_x)}{S_{2\text{Array}}(TU_x)} \right]$$

$$S_{1\text{Array_TPM}}^{(TU_x)} = S_{2\text{CAGE_TPM}}^{(TU_x)} \cdot 2^{S_{\text{Array_Relative}}^{(TU_x)}}$$

where $S_{1\text{Array_TPM}}^{(TU_x)}$ is the transcripts per million (t.p.m.) that corresponds to a specific TU_x in sample 1, $S_{2\text{CAGE_TPM}}^{(TU_x)}$ is the CAGE or EST expression value that corresponds to a specific TU_x in sample 2, and $S_{\text{Array_relative}}^{(TU_x)}$ is the relative expression value in each microarray spot that corresponds to a specific TU_x .

3. (Optional) Use an external data set for confirmation, especially if the absolute values will be used to evaluate many different microarrays.

DISCUSSION

Most microarray experiments utilize either single- or dual-color labeling and detection approaches. Dual-color labeling uses two probe mixtures having distinct labels (e.g., Cy3 and Cy5), allowing the measurement of expression ratios reliably by competitive hybridization. In such approaches, one probe mixture typically serves as reference and is derived, for example, from all of the transcripts represented on the microarray. However, some measurements that provide relative expression levels between two samples may not provide absolute expression values. The use of single-color labeling eliminates differences in hybridization efficiency seen in most dual-color approaches. Differences in hybridization efficiency lead to a loss of quantification and artifacts in ratiometric data. Single-color approaches also generally allow a simpler experimental set-up and represent the predominant approach used by commercial microarray providers.

Single-color labeling methods primarily allow direct measurement of absolute expression values using precalibration or exogenous controls. However, these efforts can be limited by insufficient information available on the reference data. These limitations can be partially overcome by integration with external data obtained by different experimental methods such as EST sequencing, SAGE, or the novel CAGE method. The tags produced by these methods can be used to provide absolute expression values for every sample used on a DNA microarray, with the units represented in t.p.m. An example of the calculation of absolute expression values for mouse transcripts in the RIKEN Expression Microarray Database (READ) using quantitative CAGE and EST tag data can be found in Kasukawa et al. (2004). The READ database (Bono et al. 2002) contains expression information for 50 mouse tissues, where dual-labeled relative gene expression levels are shown using the expression levels obtained from mouse embryo E17.5 mRNA as the reference sample. E17.5 mRNA is derived from whole body, mixed-sex mouse embryo tissue taken at mouse embryonic day 17.5. Using the absolute expression values of the E17.5 mRNA sample, the READ values of the mRNA samples from the 50 tissues can be converted into absolute values.

Both CAGE and EST data are independent of microarray data and have different data properties. CAGE and EST sequencing technologies involve the sequencing and mapping of transcripts (tags) to the genome. In order to link external EST and CAGE data to the cDNA targets used for microarray analysis, we used the FANTOM representative transcript set (RTS) based on RIKEN cDNAs and associated transcriptional unit (TU) definitions (Kasukawa et al. 2004). Briefly, EST and CAGE tag sequences are mapped to the mouse genome and then linked to unique TUs by identifying the closest TU within a 10 kb window. The cDNA microarray targets are generally based on RIKEN cDNA clones and also have an annotated TU.

The cDNA library made from E17.5 mRNA contains 49,806 5'-ESTs grouped into 7164 unique TUs by RTS. In this way, the correspondence between sequenced tags and READ clone IDs used for the microarray analysis can be established. With this preprocessing, each sequence tag and microarray spot is then annotated with a corresponding TU identifier. It is then possible to count the number of tags per TU and multiply those by the corresponding READ expression value to obtain the conversion to absolute t.p.m. values as shown in the equations used in Step 2:

$$S_{Array_relative(TU_x)} = \log_2 \left[\frac{S_{1Array}(TU_x)}{S_{2Array}(TU_x)} \right]$$

$$S_{1Array_TPM(TU_x)} = S_{2CAGE_TPM(TU_x)} \cdot 2^{S_{Array_Relative(TU_x)}}$$

In these equations, $S_{1Array_TPM(TU_x)}$ is the t.p.m. that corresponds to a specific TU_x in sample 1, $S_{2CAGE_TPM(TU_x)}$ is the CAGE or EST expression value that corresponds to a specific TU_x in sample 2, and $S_{Array_relative(TU_x)}$ is the relative expression value in each microarray spot that corresponds to a specific TU_x . In the case of absolute expression for READ (Kodzius et al. 2004), the relative expression values are obtained from the READ database. Sample 2 is always the mouse E17.5 library and sample 1 is one of the mRNA samples from the 50 tissues used in READ.

Once relative microarray expression values are converted to absolute values, it is possible to compare the converted data set directly with other externally obtained data including CAGE, EST, or SAGE data not used in the conversion procedure. This can be used to verify the conversion efficiency and accuracy. Briefly, to confirm the converted absolute expression values, publicly available expression data from SAGE and EST databases were used as a control set for direct comparison (Kodzius et al. 2004). As the number of tags contained in the libraries increases, a higher correlation can be observed between the libraries. For example, the CAGE cerebellum library has the highest number of tags (327,178) and the highest correlation of READ absolute values (0.699). Thus, the number of CAGE and EST tags used in sample 2 is important for both the accuracy of the absolute data and the detection of rare transcripts. To improve the accuracy of the absolute expression values, TUs with few tags should be ignored before applying the equations in Step 2. However, as the system may fail to detect rare transcripts because of this operation, this is a trade-off between specificity and sensitivity.

ACKNOWLEDGMENTS

We would like to thank Albin Sandelin and, for help with editing, Ann Karlsson. This work was supported (in part) by a grant from the Genome Network Project from the Ministry of Education, Culture, Sports, Science and Technology, Japan. Rimantas Kodzius was supported courtesy of an FP5 INCO2 to JAPAN fellowship from the European Union.

REFERENCES

- Bono, H., Kasukawa, T., Hayashizaki, Y., and Okazaki, Y. 2002. READ: RIKEN Expression Array Database. *Nucleic Acids Res.* **30**: 211–213.
- Kasukawa, T., Katayama, S., Kawaji, H., Suzuki, H., Hume, D.A., and Hayashizaki, Y. 2004. Construction of representative transcript and protein sets of human, mouse, and rat as a platform for their transcriptome and proteome analysis. *Genomics* **84**: 913–921.
- Kodzius, R., Matsumura, Y., Kasukawa, T., Shimokawa, K., Fukuda, S., Shiraki, T., Nakamura, M., Arakawa, T., Sasaki, D., and Kawai, J. 2004. Absolute expression values for mouse transcripts: Re-annotation of the READ expression database by the use of CAGE and EST sequence tags. *FEBS Lett.* **559**: 22–26.

Cold Spring Harbor Protocols

Calculation of Absolute Expression Values for DNA Microarray Data

Kazuro Shimokawa, Rimantas Kodzius, Yonehiro Matsumura and Yoshihide Hayashizaki

Cold Spring Harb Protoc; doi: 10.1101/pdb.prot4938

Email Alerting Service

Receive free email alerts when new articles cite this article - [click here](#).

Subject Categories

Browse articles on similar topics from *Cold Spring Harbor Protocols*.

- [Analysis of Gene Expression](#) (197 articles)
- [Analysis of Gene Expression, general](#) (129 articles)
- [Analysis of Microarrays](#) (19 articles)
- [Analysis of Microarrays, general](#) (13 articles)
- [Bioinformatics/Genomics, general](#) (192 articles)
- [Computational Biology](#) (100 articles)
- [Expression Analysis of RNA](#) (14 articles)
- [Microarrays](#) (54 articles)
- [Microarrays, general](#) (55 articles)
- [Molecular Biology, general](#) (1293 articles)

To subscribe to *Cold Spring Harbor Protocols* go to:
<http://cshprotocols.cshlp.org/subscriptions>