Supporting Text

Methodological Details of Estimation of False Discovery Rate (FDR) and False Negative Rate (FNR): Data Model. The measurements of the expression of an individual gene can be represented as $d = s + n$, where d is the measured data value, s is the signal value, and n is the noise value, as simplified from refs. 1 and 2. Signal represents the real biological gene expression level; noise is random fluctuation due to factors other than actual gene expression.

Fig. 2 shows a single variable case of measured data obtained from one microarray. The same relationship extends to instances including multiple variables, for example, multiple microarray slides used to evaluate expression of an individual gene. Probability distribution is a function in a multidimensional space. The measurement for the expression of an individual gene can be represented as $\underline{d} = \underline{s} + \underline{n}$, where the underlined notation indicates that the variable is a vector (a variable with multiple data points), $\underline{d}$ is the measured value, $\underline{s}$ is the signal value, $\underline{n}$ is the noise value, and that each of these values consist of multiple data points.

FDR and FNR. FDR and FNR have been defined in *Methods*. An example is given in Fig. 2. Thr is the threshold above which a gene is considered to show overexpression, and below which a gene is considered to show no change in expression. If Thr = 0.6, a gene having signal $s = 0$ and noise $n = 0.7$ will be determined to be a false positive, whereas a gene having signal $s = 1.0$ and noise = –0.4 will be falsely considered negative.

Different definitions of FDRs and FNRs have been proposed (3). As the number of true positives and true negatives commonly is unknown in microarray studies, we have followed Tusher (4) and Storey and Tibshirani (5) and have defined the FDR as

FDR = Number of false positives / Total number of genes in the data set satisfying the specified condition.

FDR also equals 1 less than the positive predictive value, where the positive predictive value is TP (true positives) / (TP + FP (false positives)). FDR is thus different from false positive rate (FPR), where FPR = FP / (FP + TN (true negatives)) (3). Similarly, FNR is defined in our analysis as:

FNR = Number of false negatives / Total number of genes that should satisfy a condition = Number of false negatives / (Total number of genes satisfying a condition + Number of false negatives),

whereas the conventional definition has been FNR = FN / (FN + TP) (3).

For example, if 80 genes that satisfy the specified conditions are selected in a data set, and FDR = 0.05 and FNR = 0.20, these results mean that an average of four genes (5%) out of these 80 genes satisfy the condition as a result of random fluctuation in the data, and 20% of genes having an expression level that should satisfy the specified conditions were falsely removed.

Bootstrapping and the Estimation of FDR. Bootstrapping (4, 6, 7) provides a means for estimating the random fluctuation distribution based on the data itself. To apply bootstrapping, many measurements are required, and these data values are randomly resampled to generate a bootstrapping data set. In a microarray context, assuming there are *n* microarrays and *g* genes in the data set, the measured expression values for each gene can be represented as

$d_i = (d_{i1}, d_{i2},\ldots d_{in})$, i = 1,…, *g*,

where d_i is the multivariate vector for gene i, d_{i1} is the measured expression value for gene i in microarray slide 1, d_{i2} is the measured expression value for gene i in microarray slide 2, etc.

In the bootstrapping data set, the observed values are permuted with each other or replaced with values in which the numerical sign has been randomly flipped of the original values:

$d_i^B = (d_{i1}^B, d_{i2}^B, \ldots d_{in}^B)$, i = 1,…, g,

where d_i^B is the multivariate vector for gene i in this bootstrapping data set, d_{i1}^B is the bootstrapping sampled value for gene i in microarray slide 1, d_{i2}^B is the bootstrapping sampled value for gene i in microarray slide 2, and so on. d_{i1}^B can be either d_{i1} or $(-1) \times d_{i1}$, each with probability 0.5 (4).

To find genes having significant change in expression level, the null hypothesis is that the gene expression shows no change because the random bootstrapping simulates the randomness.

The probability that a gene in the bootstrapping data set satisfies specified conditions provides an estimate of the probability that a gene with no change in expression satisfies the condition due to random fluctuation (6, 7). Bootstrapping was applied to various algorithms of GABRIEL (8).

Bootstrapping FDR = Average number of genes satisfying the rule in the bootstrapping data set / Number of genes satisfying the rule in the original data set.

However, the null hypothesis that genes in the data set show no change in expression at all is not fully valid, because there may be genes that have changes in expression in real data sets. The assumption that the simulated data set generated by bootstrapping contains completely random data are also not exactly true, because some correlation between data points could still persist after the random sampling, for example, there is a $2 \times 1/2^5 = 1/16$ chance that data values in five microarray slides are all nonflipped or flipped at the same time. Thus, the bootstrapping value provides only an estimate.

Estimation of FNR. FNR is much harder to estimate than FDR because microarrays are used for novel discovery, and there is no previously known distribution of signal value (*s*). To overcome this problem, we used the expression of the proband gene (a gene known to satisfy specified conditions) to estimate signal value. The noise value was estimated using bootstrapping again, that is,

$d_i^E = (d_{p1} + d_{i1}^B, d_{p2} + d_{i2}^B, \ldots d_{pj} + d_{in}^B)$, i = 1,…, *g*,

where d_{pj} is the expression value of proband in slide *j*, and d_i^B is generated in the same way through random sign flipping of the original data as in the FDR estimation.

d_i^E constitutes a simulated signal perturbed by noise. If random fluctuation (d_i^B) is small, this d_i^E should satisfy the conditions just like the original proband. However, if the random fluctuation is large, this d_i^E would not satisfy these conditions. The probability that d_i^E does not satisfy the specified conditions provides an estimate of the probability that a gene having the level of expression described in signal fails to satisfy the conditions because of random fluctuation of values in the data set. This procedure is repeated for a large number of times, such as 100 times.

Bootstrapping FNR = Percentage of simulated signal perturbed by noise (d_i^E) does not satisfy the condition.

The proband in FNR for GABRIEL proband-based rules can be used as the proband in FNR estimation. For rules without a predefined proband, such as *t* score pattern-based rules, we used the expression of genes that satisfy the rule as “anonymous” probands to estimate the signal value. For example, the expression level of the genes with *t* score > 1.0 provides an estimate of the average expression of genes with overexpression. The procedure can then be carried out for a different gene that satisfies the rule.

Bootstrapping FNR with proband *k* = Percentage of simulated signal perturbed by noise ($d^E = d_k + d^B$) does not satisfy the rule,

where $k = 1,\ldots, G$, with G = number of genes satisfying the rule, and d_k is the expression of the kth gene satisfying the rule.

The overall FNR is the average of each FNR:

Bootstrapping FNR = $\sum_k$ FNR with proband k / G,

where $k = 1,\ldots,G$.

While this manuscript was in preparation, a different nonparametric method for estimating "miss rate" based on permutation of data recently has been reported by Taylor *et al.* (9). "Miss rate" estimates the likelihood of missing a gene in local areas of threshold boundaries.

1. Draghici, S., Kulaeva, O., Hoff, B., Petrov, A., Shams, S. & Tainsky, M. A. (2003) *Bioinformatics* **19,** 1348–1359.

2. Kerr, M. K. & Churchill, G. A. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 8961–8965.

3. Shortliffe, E. H., Fagan, L. M., Wiederhold, G. & Perreault, L. E. (2000) *Medical Informatics: Computer Applications in Health Care and Biomedicine* (Springer, New York).

4. Tusher, V. G., Tibshirani, R. & Chu, G. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 5116–5121.

5. Storey, J. D. & Tibshirani, R. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 9440–9445.

6. Efron, B. (1979) *Ann. Stat.* **7,** 1–26.

7. Efron, B. & Tibshirani, R. (1994) *An Introduction to the Bootstrap* (Chapman & Hall, New York).

8. Pan, K.-H., Lih, C.-J. & Cohen, S. N. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 2118–2123.

9. Taylor, J., Tibshirani, R. & Efron, N. (2005) *Biostatistics* **6,** 111–117.