Supplementary information: Mathematical description

Suppose that we consider a set of microarray experiments that contains expression levels for N genes g_i (we call this set of genes $\check{N}$ – so N = # $\check{N}$) measured under several conditions (tumor types) and that we have already used a certain hypothesis test to calculate their respective p-values p_i (that reflect the probability that an equally good or better test statistic, quantifying the difference between the expression profiles of the different conditions, is generated if a certain null hypothesis is true). Also assume that the genes are ordered according to this p-value, so that $p_1<p_2<\ldots<p_i<\ldots<p_N$. Note, that in this text, we use the Wilcoxon rank sum test to generate the p-values (not corrected for multiple testing (e.g., Bonferroni correction)), which is a nonparametric test that examines the null hypothesis that the medians of two populations generating expression levels for a certain gene, under two different conditions, are identical (Pagano and Gauvreau, 2000; Troyanskaya et al., 2002). This test uses a test statistic that is based on the ranks of the expression levels of one gene rather than on the values themselves. Note that this test can only be used when there are only two different conditions (when there are more, the Kruskal-Wallis test could be appropriate (Dawson-Saunders and Trapp, 1994)). Also note, that in principle, every procedure (e.g., through random column permutations of the data (Tusher et al., 2001)) or hypothesis test (e.g., t-test), that generates p-values for every individual gene, is suited as long as its underlying assumptions are checked or assumed.

Now assume that the null hypothesis is actually false for N_1 genes (these genes are actually differentially expressed - we call this set of genes $\check{N}_1$, so N_1 = # $\check{N}_1$). Assume further that the null hypothesis is actually true for N_0 genes (these genes are not actually differentially expressed - we call this set of genes $\check{N}_0$, so N_0 = # $\check{N}_0$). Note that at this stage we do not yet consider if the null hypothesis is rejected or not (i.e., if the genes are declared differentially expressed or not).

In the next section we present a simple and straightforward method for calculating N_1 (and N_0) (also see for other approaches Benjamini and Hochberg ,2000; Keselman et al., 2002; Reiner et al., 2003). Note that all methods described in this paper were implemented in MATLAB but are straightforward to implement using other packages.

Calculation of N_1 and N_0

Assume that a gene g_t with associated p-value p_t can be found with t defined as follows:

$$t = \min_j \left\{ g_j \in \breve{N}_0 \mid \forall g_i \in \breve{N}_1 : p_i \leq p_j \right\}. \qquad (1)$$

The assumption of the existence of such a gene g_t comes down to the fact that one supposes that the largest p-value in the data set belongs to $\breve{N}_0$ (genes belonging $\breve{N}_1$ will, in general, have relatively small p-values since they are not generated under the null hypothesis).

Now choose any gene g_k with $p_k \geq p_t$ (by definition g_k belongs to $\breve{N}_0$, since all genes belonging to $\breve{N}_1$ have p-values smaller than p_t). Note that, since the genes were ordered according to their p-value, k is the number of genes belonging to $\breve{N}$ with a p-value equal to or smaller than p_k. Since $\breve{N} = \breve{N}_1 \cup \breve{N}_0$, we can write the following set of equations:

$$\begin{cases} k = \#\left\{ g_i \in \breve{N}_1 \mid p_i \leq p_k \right\} + \#\left\{ g_i \in \breve{N}_0 \mid p_i \leq p_k \right\} & (2) \\ N = N_1 + N_0. & (3) \end{cases}$$

Since, by definition, all genes belonging to $\breve{N}_1$ have p-values smaller than p_k, the first term of Equation (2) equals N_1. To calculate the second term, we assume that the test statistics of the gene expression profiles of $\breve{N}_0$ (that are generated by the distribution of the null hypothesis) are independent (all genes, that exhibit coexpression that can change the test statistic, are assumed to belong to $\breve{N}_1$). Under this condition and by definition, the probability, that a gene from $\breve{N}_0$ has an equally good or better test statistic than g_k (i.e., has a p-value equal to or smaller than p_k), equals p_k. This means that the expected value (mean of the binomial distribution) of the second term in Equation (2) equals $p_k.N_0$ and that we can approximate Equation (2) as follows:

$$k \approx N_1 + p_k.N_0. \qquad (4)$$

Deriving N_1 from the set of Equations (3) and (4) gives:

$$N_1 \approx \frac{k - p_k.N}{1 - p_k}. \qquad (5)$$

Note that, for a given data set, N_1 is constant. Now define V_i, for *every* gene g_i, as follows:

$$V_i = \frac{i - p_i.N}{1 - p_i}. \qquad (6)$$

According to Equation (5) and for $p_i \geq p_t$, V_i is constant and equals N_1. Moreover, it is easy to prove that $V_i < N_1$ when $p_i < p_t$ and that V_i goes to zero when p_i gets smaller.

Using this information, we can present an easy method to derive N_1 (and N_0 through Equation (3)): Calculate V_i for every gene g_i and plot these values in a graph (e.g., i on the X-axis and V_i on the Y-axis). If this graph reaches a constant level at a certain gene, this gives us respectively N_1 and g_t. In practice, after reaching the constant level, the graph will slightly vary around a mean value (because of the approximation we used to derive Equation (4)). So for the calculation of N_1, it is better to take the mean of V_i in a certain interval [r,s] where $r > t$ and $s << N$, if possible (if $i \approx N$, $p_i \approx 1$ and the denominator in Equation (6) gets very small and the formula for V_i becomes ill conditioned).

References

Benjamini Y and Hochberg Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat*., **25**, 60-83.

Dawson-Saunders B and Trapp R.G. (1994) *Basic & Clinical Biostatistics, 2nd edition*. Appleton & Lange, Connecticut.

Keselman HJ, Cribbie R and Holland B. (2002) Controlling the rate of Type I error over a large set of statistical tests. *Br. J. Math. Stat. Psychol.* , **55**, 27-39.

Pagano M and Gauvreau K. (2000) *Principles of Biostatistics, 2nd edition*. Duxbury Press.

Reiner A, Yekutieli D and Benjamini Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368-375..

Troyanskaya OG, Garber ME, Brown PO, Botstein D and Altman RB. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-1461.