

Statistical Methods

Identification of Differentially Expressed Genes by Weighted Gene Analysis

Michael Bittner
mbittner@nhgri.nih.gov

Yidong Chen
yidong@nhgri.nih.gov

Assuming that we have two groups of samples, one from BRCA1 mutation positive ($n_1 = 7$) and the other from BRCA1 negative ($n_2 = 15$). For a given two-cluster setting, a discriminative weight for each gene can be evaluated by,

$$w = d_B / (k_1 d_{w_1} + k_2 d_{w_2} + \epsilon)$$

where d_B is the center-to-center distance (between cluster Euclidean distance), d_{w_i} is the average Euclidean distance among all sample pairs, total of t_1 and t_2 sample pairs for cluster 1 and 2, respectively, and $k_1 = t_1 / (t_1 + t_2)$, and $k_2 = t_2 / (t_1 + t_2)$. ϵ is a small constant (0.01 in our study) to prevent zero denominator case. Genes may then be ranked on the basis of w . The equation for weight w is not only designed to evaluate discriminative ability for single gene, but also capable of evaluate discriminative ability for 2 or more genes together. If you do not assume the second group of samples to be a tight cluster one may drop the d_{w_2} term.