

Detailed Description of Methods implemented in HDBStat!

1. Data Analysis

The data analysis is carried out for two group comparisons. In a two groups comparison expression level measurements for a set of genes are compared for two groups of microarray chips, which will be referred to as group 1 and group 2. Analyses can be carried out for multiple two group comparisons. In a two group comparison we distinguish between independent observations and paired observations. Per two group comparison the following five types of statistical analysis are carried out for your data:

1. Data preprocessing – image processing, normalization, etc.
2. Quality Control
3. Descriptive statistics
4. Empirical Bayes estimation
5. Hypotheses testing
6. Multiple comparison adjustments to p-values resulting from hypothesis testing
7. Mix-o-matic method

1.1 Data Preprocessing

With data preprocessing, the data is modified prior to the statistical analysis in order to remove non-biological variation and to obtain a better fit of the underlying statistical model to the data. In HDBStat!, data preprocessing is an optional step, and if selected, normalization is carried out first and transformation is applied to the normalized data. These steps are described below in more details.

1.1.1 Image processing

HDBStat! does not process images. If the data is in *.cel or *.tiff format, Bioconductor package is used to process these images.

1.1.2 Normalization

Normalization is a procedure intended to remove variability among chips that is unrelated to treatment conditions of interest. Two types of normalization methods are available: 1) chip-mean normalization, in which each observation is divided by the mean of the chip, and 2) quantile-quantile normalization, which is a ranking procedure in which each observation on the chip is ranked by fluorescence and then converted to the value of a deviation that would be expected from the standard normal distribution based on the observation's rank. Quantile-quantile normalization results in data from each chip with a mean of zero and standard deviation of 1.0.

1.1.3 Transformation

Transformation is a process of applying a mathematical function to every observation in a data set in order to satisfy assumptions of the statistical models used for analysis. We currently offer three scales of logarithmic transformation, base-2, base-e and base-10.

1.2 Quality Control Specifications

The following quality control output is generated:

1. chip summary statistics
2. chip mean by standard deviation scatter plot
3. chip correlation matrix
4. deleted residuals
5. Standard outlier detected

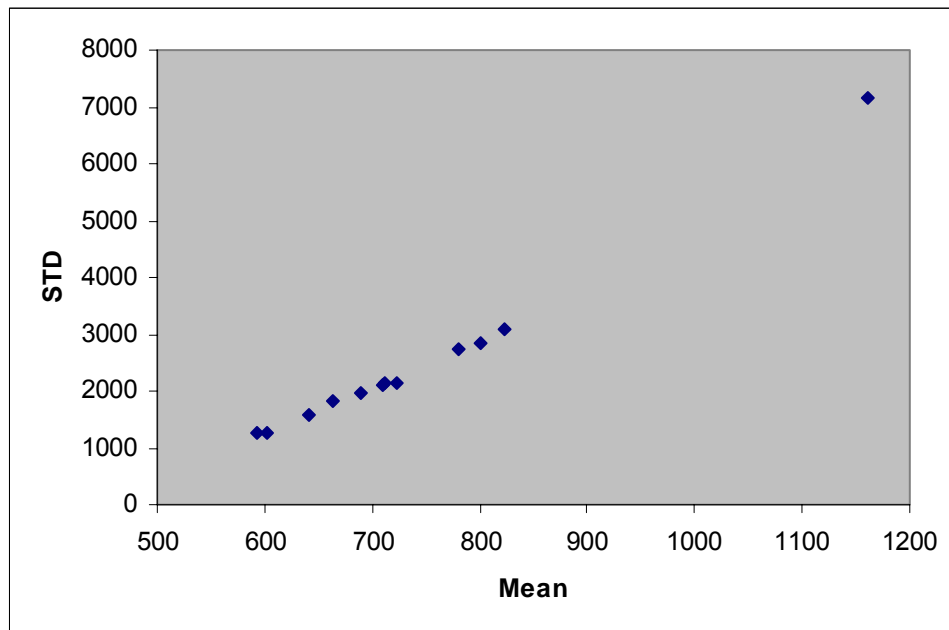
1.2.1 Chip summary statistics

For both raw and preprocessed data, the following summary statistics is calculated for each chip:

1. mean
2. standard deviation
3. number of observations
4. minimum value
5. maximum value
6. Skewness
7. Kurtosis

1.2.2 Chip mean expression by standard deviation scatter plot

This plot graphs for each chip its mean expression on the standard deviation, where mean and standard deviation are of all the probe sets (signal, RMA output, etc) entered in the gene expression level file.



In general we believe that the data should be a tight line between the lower left and the upper right as is illustrated in the graph above. This is the graph that should be seen when slightly (x-axis) different amounts of processed RNA are run on an array. Deviations from this trend may be indication of different RNA labeling efficiency in different portions of the array. Also if the values

on the x-axis exhibit a large range, some chips may have a higher proportion of the features at saturation and/or below detected than others.

1.2.3 Pearson's correlation matrix for chips

Two different tables are generated:

1. correlation matrix for non-normalized chips
2. correlation matrix for normalized and/or transformed chips

Example data

X	Y	Z _X	Z _Y	Z _X Z _Y
12	33	-1.07	-0.61	0.65
15	31	-0.07	-1.38	0.97
19	35	-0.20	0.15	-0.03
25	37	0.55	.92	0.51
32	37	1.42	.92	1.31
SUM =				3.40

Example table

	Pre1	Post1	Pre2	Post2	Pre3	Post3	Pre4	Post4	Pre5	Post5	Pre6	Post6
Pre1	1.000	0.797	0.918	0.843	0.780	0.864	0.897	0.923	0.876	0.898	0.759	0.906
Post1	0.797	1.000	0.708	0.962	0.494	0.936	0.909	0.851	0.921	0.910	0.950	0.898
Pre2	0.918	0.708	1.000	0.756	0.867	0.790	0.832	0.859	0.796	0.841	0.661	0.841
Post2	0.843	0.962	0.756	1.000	0.546	0.973	0.947	0.911	0.958	0.953	0.938	0.923
Pre3	0.780	0.494	0.867	0.546	1.000	0.583	0.629	0.714	0.583	0.645	0.438	0.668
Post3	0.864	0.936	0.790	0.973	0.583	1.000	0.965	0.935	0.956	0.967	0.914	0.932
Pre4	0.897	0.909	0.832	0.947	0.629	0.965	1.000	0.954	0.955	0.975	0.891	0.934
Post4	0.923	0.851	0.859	0.911	0.714	0.935	0.954	1.000	0.926	0.957	0.808	0.942
Pre5	0.876	0.921	0.796	0.958	0.583	0.956	0.955	0.926	1.000	0.951	0.899	0.936
Post5	0.898	0.910	0.841	0.953	0.645	0.967	0.975	0.957	0.951	1.000	0.877	0.941
Pre6	0.759	0.950	0.661	0.938	0.438	0.914	0.891	0.808	0.899	0.877	1.000	0.832
Post6	0.906	0.898	0.841	0.923	0.668	0.932	0.934	0.942	0.936	0.941	0.832	1.000

A Pearson correlation matrix should be produced among all chips before and after normalization. The Pearson correlation coefficient between two chips, 'A' and 'B', is obtained by computing the simple correlation coefficient between all the observations on chip 'A' and the observations on chip 'B'. The correlation coefficient is computed as:

$$r = \frac{Cov(x, y)}{S_x S_y}$$

where: S_x = standard deviation of values on chip 'A'

S_y = standard deviation of values on chip 'B'

$Cov(x, y)$ = Covariance between observations on chip 'A' and observations on chip 'B'

The covariance is computed as:
$$\text{Cov}(x, y) = \frac{\sum_{i=1}^n x_i y_i - (n * \bar{x} * \bar{y})}{n - 1}$$

The complete correlation matrix (a n x n table, where n=number of chips) should be produced from the observations of all genes on the chips using raw data and normalized data.

The following link graphically illustrates what a correlation from -1 to 1 means.
<http://noppa5.pc.helsinki.fi/koe/corr/cor7.html>

We don't yet have a good rule-of-thumb for what within-group and between-group correlations 'should be'. We believe that the key should not be a particular correlation (but the higher the better) but rather homogeneity with the rest of the data.

1.2.3 Deleted Residuals

Deleted Residuals (DR) is a methods we have adapted to test if a chip is homogeneous with other chips within its group. In general the deleted residuals (described below) should follow a t-distribution with n-2 degrees of freedom, where n is the total number of *chips* in the group. HDBStat! software illustrates this graphically. In general the red histogram should mimic the known t-distribution. Deviations of the histogram from the graph are evidence of non-homogeneity among chips. In general, one can decide if a chip is homogeneous or not in 3 ways. 1) subjectively make a decision based upon the graphs if a chip deviates significantly from expected t-distribution (see figure below) 2) Using p-values from a Kolmogorov-Smirnoff test comparing the observed and expected data 3) Empirical cut-offs based upon public data.

Quality Control analysis is based on computation of deleted residuals. For a sample of size *n*, the deleted residual for observation *x_i* is:

$$\sqrt{\frac{(n-2)\left(\frac{n}{n-1}\right)^2 (x_i - \bar{x})^2}{css - \frac{n}{n-1}(x_i - \bar{x})^2}} = (x_i - \bar{x}) \sqrt{\frac{(n-2)\left(\frac{n}{n-1}\right)^2}{css - \frac{n}{n-1}(x_i - \bar{x})^2}}$$

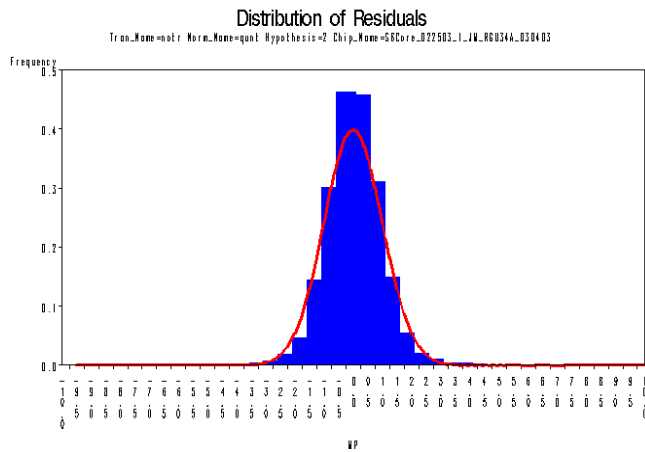
where: *n* = sample size
x_i = value of *i*th observation
 \bar{x} = sample mean
css = corrected sum of squares

For Paired data, deleted residuals are calculated for each pair instead of individual chips.

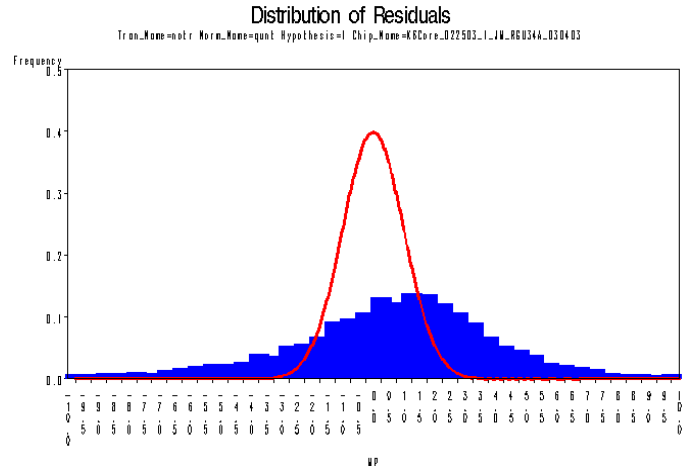
Following output is generated:

1. table of deleted residuals for each chip per group
2. table of summary statistics on deleted residuals
3. for each chip, a histogram with overlaid t-distribution
4. table of standard outliers for each chip per group

In general we would recommend this method, but the dimensionality of the data suggests that virtually all data would be significantly different from the t-distribution. In general this observation held, based upon ~2000 chips from Gene Expression Omnibus (GEO) where only 6 of the chips were not significantly different ($p < 0.05$) from the expected t-distribution. Thus we have developed empirical cut offs based upon data from GEO. Eventually the cut-offs will be integrated into the HDBStat! software, but table below gives some cut-offs for mean, kurtosis, and Kolmogorov-Smirnov d.



Homogeneous Chip



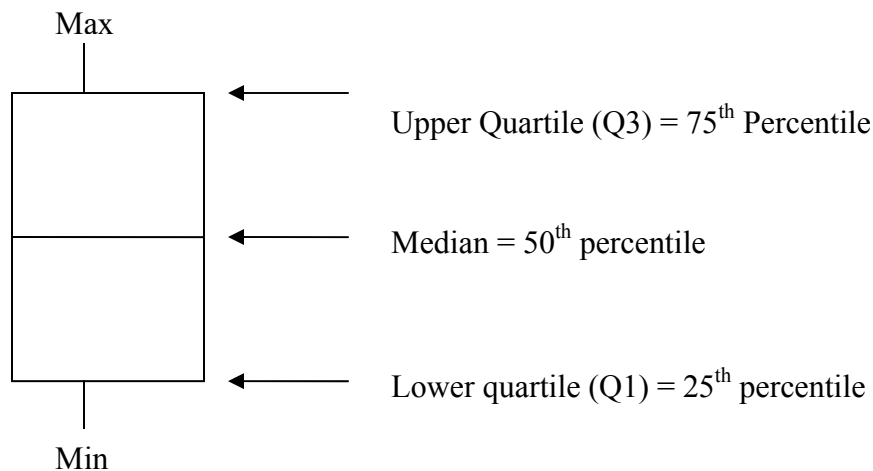
Potentially non-homogeneous chip

Table below shows Empirical cut-offs for various measures of deleted residuals. In general the smaller the k-s the closer the observed data is to the expected data. However, for the kurtosis and standard deviation values that are too small and too large are evidence on non-homogeneity. Ideally the kurtosis should be a little more than 3 and standard deviation a little more than one (the higher the degrees of freedom the closer to 3 and 1 that should be expected).

Percentile	Kurtosis	Stdev	K-S
1	-0.400	0.607	0.022
5	0.517	0.718	0.034
10	0.922	0.765	0.041
20	1.669	0.832	0.056
30	2.624	0.903	0.069
40	4.101	0.978	0.083
50	6.038	1.060	0.092
60	9.185	1.152	0.103
70	16.266	1.267	0.117
80	34.740	1.445	0.138
90	186.035	1.896	0.182
95	937.179	3.093	0.241
99	4156.699	36229.028	0.450

1.2.5 Standard Outlier Detection

Standard outliers are detected as follows using results from deleted residuals:



Interquartile Range (IQR) = $Q3 - Q1$

Vertical Axis = response variable
Horizontal axis = factor of interest

Q1 = lower quartile

Q3 = upper quartile

IQR = $Q3 - Q1$ (i.e. the difference between the upper and lower quartile)

- Mild outliers: $L1 = Q1 - (1.5 * IQR)$
 $U1 = Q3 + (1.5 * IQR)$
- Extreme outliers: $L2 = Q1 - (3.0 * IQR)$
 $U2 = Q3 + (3.0 * IQR)$

1.3 Descriptive Statistics

The following two descriptive statistics are computed for each gene individually:

1. mean of expression levels over group 1
2. mean of expression levels over group 2
3. fold change

1.4 Empirical Bayes

We have an implementation of an Empirical Bayesian Analysis to obtain shrunken estimators of gene expression differences for each gene. The motivation for this approach is to provide unbiased random-effect estimates of the size of the gene expression difference for each gene.

When a large number of effects are simultaneously estimated, under certain statistical assumptions, the variation among the sizes of the effects is generally larger than expected from underlying variability among the true (unknown) effects. The source of the bias or increase in

variability among effects is error; the error in estimation of each individual effect inflates the variation among all effect estimators. The inflation of effect size estimates does not influence validity of hypothesis testing, but it can mislead us into thinking that there is more variability in gene expression differences than there really exists. For this reason, we present Empirical Bayes estimates which are scaled according to an estimate of the true variance of gene expression differences.

In addition to measuring the amount of expression for a single gene on a single case, we will wish to measure the effect of an experimental manipulation on gene expression. EB is the name for a collection of methods and approaches (Carlin & Louis, 2000) that are closely related to shrinkage estimation, hierarchical Bayesian methods, random effects models, and measurement error models (Draper et al., 1992). EB is not the same as fully Bayesian analysis. The enormous number of different genes analyzed simultaneously in microarray studies is challenging in that if one chooses genes for which evidence of up- or down-regulation looks strongest from among a large number of genes, then effect estimates, such as fold-change values, will be highly biased as we have shown in related contexts (Allison et al., in press-b). However, the large number of things being estimated presents an ideal situation for EB estimation. To our knowledge, only Efron et al (2001) have used EB in the microarray context.

The concept behind EB is that we often have two estimates of a population parameter: an estimate based on some prior expectation and a sample estimator from a particular observed data set. Because both the prior model-based estimate and the sample statistic are associated with some error, the best overall estimate is a weighted average of the two available estimates where the weights are inversely related to the individual estimates' variances (Morris, 1983). The EB estimator has smaller mean square error (MSE) than ML estimators as long as the number of parameters being estimated is > 3 (Berger, 1982). EB methods are hybrids of classical frequentist methods and fully Bayesian methods (Shoemaker et al., 1999) because the data are used to estimate all parameters but a 'prior' distribution is postulated to describe the parameters' distribution.

Our specific EB approach derives largely from Morris (1983) but is tailored for microarray analysis. Let there be k genes. For each gene there is a parameter μ_i that describes the effect of the independent variable under study (e.g. ripeness of strawberries) on gene expression levels. Many different metrics could be chosen for. We favor the standardized difference between means, i.e., the difference between treatments means divided by the pooled within-treatment standard deviation. Like 'fold-change' it is essentially scale-free but, unlike fold-change, it has well documented statistical properties (Hedges & Olkin, 1985). For the present, we will consider that comparison of gene expression levels in two treatments (e.g. ripe vs. unripe). For the i th gene, we can derive an ordinary sample estimate of μ_i denoted \bar{y}_i and an estimate of its variance S_i . We can also calculate a predicted value of μ_i denoted $\hat{\mu}_i$ via a regression equation in which we used various gene characteristics as predictor variables. These characteristics could be as simple as the mean value of \bar{y}_i for all k genes (i.e., a constant), or more complex such as information about sequence, membership in known gene families, or expression effects observed in other studies, species, etc. Finally, the values of $\hat{\mu}_i$ will also have an estimated variance, denoted \hat{S}_i . Given these quantities (formulae for which are contained in Morris, 1983), an EB estimator of μ_i can be written as: $\hat{\mu}_i + r(\bar{y}_i - \hat{\mu}_i)$ where r is the number of predictor variables (not including a mean). The EB estimate will lie

between the ordinary estimate calculated only on data from one gene in one study and a predicted effect based on all available data.

1.4.1 Computational Details

d_i = ordinary estimate of the difference in gene expression. For unpaired data, this is estimated as the difference between the two group means, i.e., $\bar{x}_1 - \bar{x}_2$. For paired data, this will be the average of the differences.

V_i = variance of the difference in gene expression, d_i (same as the term inside the square root of the denominator of the ordinary t-test). For paired data:

$$V_i = \frac{v_1}{n_1} + \frac{v_2}{n_2}, \text{ where}$$

v_i = sample variance of treatment i ($i=1,2$), and

n_i = number of observations in treatment i ($i=1,2$),

V_i = variance of the difference.

For paired data, V_i is the sample variances of the individual gene expression differences.

k = number of genes for which the variance of the difference is properly defined and greater than zero. Any genes for which the variance of the difference is undefined or zero should be dropped from analysis.

d_i , S_i , and k are the only input variables – once these have been defined and computed, EB estimates may be obtained.

Before computation begins, the user should specify two parameters:

1. the maximum number of iterations that should be allowed in computing \hat{A} (default value of 50).
2. The tolerance value for convergence, tol (default value of 1E-10).

Step 1 is to solve for the parameter \hat{A} .

1. Set initial values of:
 - a. $\hat{A} = 0$
 - b. $obj = 2tol$
2. iterate on the following equations:

$$(1) \quad w_i = \frac{1}{V_i + \hat{A}}$$

$$(2) \quad \hat{\mu} = \frac{\sum_i w_i d_i}{\sum_i w_i}$$

$$(3) \hat{A}' = \frac{\sum_i w_i \left\{ \left(\frac{k}{k-1} \right) (d_i - \hat{\mu})^2 - V_i \right\}}{\sum_i w_i}$$

(4) If $\hat{A}' > 0$ then let:

- a. $c = 10^{(-\text{floor}(\log_{10}(\hat{A}')))}$
- b. $obj = \left(c (\hat{A}' - \hat{A})^2 \right)$
- c. $\hat{A} = \hat{A}'$

(5) If $\hat{A}' < 0$ then let $\hat{A} = 0$

(6) Stop iteration if:

- a. Number of iterations has reached user specified maximum (default value of 50)
- b. If $\hat{A} = 0$
- c. $obj < tol$

Step 2: compute EB estimates:

$$B_i = \left(\frac{k-3}{k-1} \right) \left(\frac{V_i}{V_i + \hat{A}} \right)$$

$$\theta_i = (1 - B_i) d_i + B_i \hat{\mu}$$

Step 3: (new) compute standard errors of EB estimates (s_i):

If A is not equal to zero and Number of iterations has not reached user specified maximum (default value of 50) then compute the following -

$$(1) \bar{V} = \frac{\sum_i w_i V_i}{\sum_i w_i}$$

$$(2) r_i = \frac{w_i}{\sum_i w_i}$$

$$(3) VB_i = \frac{\left(\frac{2}{k-3} \right) (\bar{V} + \hat{A}) B_i^2}{V_i + \hat{A}}$$

$$(4) s_i = \sqrt{V_i (1 - B_i (1 - r_i)) + VB_i (d_i - \hat{\mu})^2}$$

Output:

A = variance of true differences in gene expression

θ_i = EB estimate for the i^{th} gene.

1.5 Hypothesis Testing

For each gene individually p-values are computed with respect to the null-hypothesis that there is no difference in expression level between group 1 and group 2. P-values can be computed using different methods of hypothesis testing. Distinction is made between independent observations and paired observations.

For the methods explained here, let $x_1, \dots, x_m, y_1, \dots, y_n$ be the observed expression levels for a particular gene with m and n the sample sizes for group 1 and group 2.

1.5.1 Equal (Pooled) Variance t-test

This is the ordinary Student's t-test assuming equal variances and normality. The p-value of the ordinary t-test is given by: $p\text{-value} = 2P\left(t_{m+n-2} > \left| \frac{\bar{y}_n - \bar{x}_m}{s\sqrt{1/m+1/n}} \right| \right)$, where t_{m+n-2} is a stochastic variable with a student t distribution with $m+n-2$ degrees of freedom, \bar{x}_m the sample average over x_1, \dots, x_m , \bar{y}_n the sample average over y_1, \dots, y_n , $s = \sqrt{\left((m-1)s_x^2 + (n-1)s_y^2 \right) / (m+n-2)}$ the pooled standard deviation with s_x^2 the sample variance over x_1, \dots, x_m and s_y^2 the sample variance over y_1, \dots, y_n .

1.5.2 Unequal Variance (Satterthwaite) t-test

This is an approximated t-test proposed by (Welch 1947) and (Satterthwaite 1946) which can be assumed robust against heterogeneity of variances. Satterthwaite approximate t for unequal variances. This test is based on Student's t-test, but includes an adjustment for unequal sample variances between groups. The test is generally needed in the case of heterogeneous error variances AND unequal sample sizes.

The p-value of the Satterthwaite t-test is given by:

$p\text{-value} = 2P\left(t_f > \left| \frac{\bar{y}_n - \bar{x}_m}{\sqrt{\lambda_x s_x^2 + \lambda_y s_y^2}} \right| \right)$, where $\lambda_x = 1/m$, $\lambda_y = 1/n$, t_f a student-t distribution with f degrees of freedom and f given by $f = \left(\lambda_x s_x^2 + \lambda_y s_y^2 \right)^2 / \left(\lambda_x^2 s_x^4 / f_x + \lambda_y^2 s_y^4 / f_y \right)$, with $f_x = m-1$ and $f_y = n-1$.

1.5.3 Chebby Checker Methods

The Chebby checker p-value is a new method suitable for small sample sizes and non-normality distributed data, which is often being the case in microarray experiments. In fact, the Chebby checker method provides not a p-value but an upper bound for a p-value based on Chebyshev's inequality $P\left(\left| (\tau - \mu_\tau) / \sigma_\tau \right| \geq T \right) \leq 1/T^2$, where τ is a random variable representing the test-statistic with a mean and a standard deviation equal to μ_τ and σ_τ . We provide three types of Chebby Checker p-values. The first one is based on the genuine Chebyshev's inequality and the other two are based on modifications of this inequality to make this method less conservative. The second Chebby checker p-value is based on the variant proposed by (DasGupta, 2000) and obtained by multiplying the first Chebby checker p-value by 1/3, and the third Chebby checker p-value is

based on the variant obtained by (Mallows 1956) and obtained by multiplying the first Chebby checker p-values by 4/9.

Let cc(1), cc(2) and cc(3) the first, second and third version of the implemented Chebby checker method. The formula's for the p-values of these methods are given below.

Let $t = |\tau/\sigma_x|$, with τ the ordinary t-test statistic given by $\tau = (\bar{y}_n - \bar{x}_m)/(s\sqrt{1/m+1/n})$, $s = \sqrt{((m-1)s_x^2 + (n-1)s_y^2)/(m+n-2)}$ the pooled standard deviation and $\sigma_\tau = \sqrt{(m+n)/(m+n-2)}$ the standard deviation of τ under normality. Then the cc (1) p-value $p_{(1)}$ is given by $p_{(1)} = 1/T^2$

cc(2) p-value:

The cc(2) p-value $p_{(2)}$ is given by $p_{(2)} = (1/3)*(1/T^2)$.

The cc(3) p-value $p_{(3)}$ is given by $p_{(3)} = (4/9)*(1/T^2)$.

1.5.4 Bootstrap t-tests

A Bootstrap t-test is a resampling technique useful for data that do not adhere well to known statistical distributions, as is often the case with microarrays. In this test, large numbers of 'pseudo datasets' are made by randomly sampling observations from the real data. The t-statistics estimated in the pseudo datasets are compared to the actual t-statistic to determine the ratio of times that the pseudo data sets generate results more extreme than we actually observed. The proportion of times that more extreme results are obtained at random is an indication of whether or not it is likely that we could have obtained the observed result by chance alone.

Bootstrap p-values are computed for each specified combinations of genes and comparisons. The computations described here are for a particular gene and comparison. Distinction is made for the following four cases:

1. Exact bootstrap for independent observations
2. Random bootstrap for independent observations
3. Exact bootstrap for paired observations
4. Random bootstrap for paired observations

In case of independent observations the bootstrap p-value is computed from the observations $x_1, \dots, x_m; y_1, \dots, y_n$ where x_1, \dots, x_m the observed expression levels of a particular gene for are group 1 and y_1, \dots, y_n are the observed expression levels of this gene for group 2. These two groups of observations are assumed to be statistically independent. In case of paired observations the bootstrap p-value is computed from the paired observations $(x_1, y_1), \dots, (x_n, y_n)$. In exact bootstrap methods all possible bootstrap re-samples are generated which can be generated by drawing with replacement from the observed data. In random bootstrap these bootstrap re-samples are generated by randomly drawing with replacement. It can be seen that a p-value generated with a random bootstrap method is an approximation of a p-value generated with the corresponding exact bootstrap method. From a statistical point of view exact bootstrap methods are superior to random bootstrap methods, but for larger sample sizes exact bootstrap are computationally very intensive.

For all four cases of bootstrap methods, bootstrap p-values are computed in the following four steps:

1. Computation of observed test-statistic \hat{t} from the observed data
2. Generation of the bootstrap re-samples from the observed data
3. Computation of the bootstrap test-statistics from the bootstrap re-samples
4. Computation of the p-values from the bootstrap test-statistic

1.5.5 Interpretation of p-values

A p-value gives information about how likely the observed data is under the null-hypothesis, but gives no information how likely the truth of the null-hypothesis is given the observed data. Except for the Chebby Checker method, the p-value for a particular gene is the probability that if we repeat exactly the same experiment under the assumption that this gene is not differentially expressed (null-hypothesis is true) we will obtain a difference in gene expression which is the same as or higher than the difference we have observed. For the Chebby Checker method the p-value here is given by an upper bound for this probability. The observed difference here is expressed as the t-value.

1.6 Multiple Comparison Adjustments

An important statistical problem in microarray analysis is the large number of statistical tests made. Because of the large number of hypotheses tested (one for every gene), it is very likely that a large number of type I errors will be committed, i.e., many genes will be declared differentially expressed when in fact they are not. As an example, if 10,000 genes are analyzed and none of the genes are differentially expressed, if we use a p-value of 0.05 as a threshold for declaring genes differentially expressed, we would expect to declare 500 genes as differentially expressed. To try to control the total number of gene that are falsely declared as significant, p-values of standard statistical tests are adjusted to control the total number of false rejections. Three such corrections are available in MDAC analyses, Bonferroni, Sidak, and False Discovery Rate.

1.6.1 Bonferroni Correction

The Bonferroni correction (Bland & Altman, 1995) conservatively controls the Family-Wise Error Rate (FWER), which refers to the total number of type I errors (genes incorrectly declared differentially expressed) among a family of hypothesis tests regardless of the pattern of dependence or independence among the multiple tests. A Bonferroni p-value is the upper limit on the probability of obtaining a difference in gene expression as large as or larger than that observed for a particular gene by random chance at least once among all observed gene expression differences. For example, if a difference of 2.1 units is observed with a Bonferroni p-value of 0.05, this means that if none of the genes were in fact differentially expressed, there would be a 5% chance of obtaining one difference of 2.1 units or more among all genes analyzed.

1.6.2 Sidak Correction

The Sidak (1967) correction is similar to the Bonferroni correction, but is a little less conservative (i.e., will identify more genes). It conservatively controls the FWER under independence of tests or the covariance structure among the p-values fits into a broad class of 'positive' structures.

1.6.3 False Discovery Rate

The False discovery rate has a different interpretation than the p-value described thus far. The false discovery rate (Benjamini and Hochberg, 1995) is the expected proportion of hypotheses that were rejected that were rejected falsely. In a microarray experiment, the false discovery rate is the expected proportion of genes declared differentially expressed that was in fact not differentially expressed. This value differs from a p-value in that it is conditioned on the number of hypotheses that were rejected, whereas p-value (probabilities of type I errors) are conditioned on those null hypotheses that are in fact true. In this way, the false discovery rate has more information because it is a reflection of both reality and the data

1.6.4 False Discovery Rate with Regression Dependence

Based upon the FDR methods of Benjamini and Yekutieli (2001). The previous method assumed that the p-values were independent. This FDR assumes that there is a positive dependence between the test statistics.

1.7 Mix-o-matic Method

With the mix-o-matic we refer to the mixture model approach proposed in (Allison et al 2002). The mix-o-matic provide estimates of the number and proportion of truly differentially expressed genes and the true positive rate, false positive rate and false negative rate, by fitting a mixture model to a set of p-values. This set of p-values results from testing the null-hypothesis that there is no differential expression between group 1 and group 2 for each gene individually. We can carry out the mix-o-matic for p-values resulting from the pooled variance t-test, Satterthwaite t-test and the bootstrap t-test.

1.7.1 General Posterior Probabilities

Presented here are three types of posterior probability called the posterior True Positive (TP), posterior True Negative (TN) and posterior False Negative (FN). These three types of posterior probabilities are graphically presented as three separate graphics as function of the threshold value of the p-value on the basis of which the corresponding gene is declared significant differentially expressed. In each of the three graphics the threshold of the p-value is displayed on the x-axis and the corresponding posterior probability on the y-axis. The three graphics and their interpretation are described in more detail below.

1. True Positive (TP): For any value x on the x-axis, the corresponding value on the y-axis is an estimate of the proportion of genes which are truly differentially expressed among the genes which are declared significant differentially expressed according to $p - \text{value} \leq x$.
2. False Positive (FP): For any value x on the x-axis, the corresponding value on the y-axis is an estimate of the proportion of genes which does **not** have a true differentially expression among the genes which are declared significant differentially expressed according to $p - \text{value} \leq x$.
3. False Negative (FN): For any value x on the x-axis the corresponding value on the y-axis is an estimate of the proportion of genes which have a true differential expression among the genes which are **not** declared significant differentially expressed because their corresponding p-value is bigger than x .

1.7.2 Gene Specific Posterior Probabilities

Here we give the same three types of probabilities for each gene individually in an excel file in the columns labeled by ‘PTP’, ‘PFP’ and ‘PFN’. These three types of probabilities are defined below.

1. **PTP**: For a particular gene this is an estimate of the proportion of genes which are truly differentially expressed among the genes with an observed p-value **smaller** than the observed p-value of this particular gene;
2. **PFP**: For a particular gene this is an estimate of the proportion of genes which does **not** have a true differential expression among the genes with an observed p-value **smaller** than the observed p-value of this particular gene;
3. **PFN**: For a particular gene this is an estimate of the proportion of genes which does have a true differential expression among the genes with an observed p-value **bigger** than the observed p-value of this particular gene.

1.7.3 Description of Model

In the mix-o-matic model we assume that all genes of interested can be sub-divided into two groups denoted by H_0 and H_1 , where the group H_0 refers to all genes for which there is no differentially expression and H_1 to all genes for which there is a true differential expression. The proportion of these two groups H_0 and H_1 are represented by the two parameters λ_0 and λ_1 . Because any gene of interest can be either in group H_0 or group H_1 the two parameters λ_0 and λ_1 are restricted to satisfy $\lambda_0 + \lambda_1 = 1$. Assuming that the statistical test from which the p-values are computed is valid, the p-values of genes in group H_0 have a uniform distribution between 0 and 1. This means that for genes in group H_0 the corresponding p-values have equal probabilities of being within any equally-sized intervals between zero and one. For example, a p-value has the same chance ($p=0.1$) of being between 0.9 and 1 as being between 0 and 0.1. The distribution of p-value for genes in group H_1 is expected to be more concentrated close to zero. We have modeled the distribution of p-values in group H_1 by a beta distribution with parameters r_1 and s_1 .

In reality we don’t know which genes belong to the group H_0 and which genes belong to group H_1 . But in our statistical model we assume that any gene has a probability of λ_0 to belong to group H_0 and a probability of λ_1 to belong to group H_1 . This imply that we can model the p-values with a mixture distribution where with probability λ_0 a p-value has a uniform distribution between 0 and 1 (group H_0) and with probability λ_1 this p-values has a beta distribution with parameters r_1 and s_1 . The implicit assumption being made here is that each truly differentially expressed genes has the same statistical power to be declared differentially expressed. By means of maximum likelihood estimation (MLE) we then obtain estimates $\hat{\lambda}_0$, $\hat{\lambda}_1$, \hat{r}_1 and \hat{s}_1 of our statistical model. If k is the total number of genes being investigated, than the proportion of truly differentially expressed genes is estimated by $1 - \hat{\lambda}_0$ and the number of truly differentially expressed genes is estimated by $(1 - \hat{\lambda}_0)k$. The posterior probabilities are also a function of the estimated parameters as is worked out in a separate subsection concerning computational details.

1.7.4 Underlying Assumptions of Model

Independence of p-values

The k p-values on the basis of which the parameters of the mix-o-matic model are estimated are assumed statically independent.

Statistical validity of p-values

The statistical test by which the p-values are computed is assumed statistical valid, which means that under the null-hypothesis that the corresponding gene does not have a differential expression, the p-value has a uniform distribution between 0 and 1. Truly differentially expressed genes have the same statistical power to be declared differentially expressed. Because for all truly differentially expressed gene the corresponding p-value is modeled with the same probability distribution, we have to assume that each of these truly differentially expressed genes has the same statistical power to be detected as differentially expressed. We can roughly that we assume that each truly differentially expressed gene has the same degree of differential expression.

1.7.5 Computational Details

Here we derive formulas for the general and gene-specific posterior probabilities TP, FP and FN described earlier. Here for we will derive formula's for the following three quantities.

TP(x) = the expected proportion of genes with a true differential expression among the genes with a p-value $\leq x$

FP(x) = the expected proportion of genes which does not have a true differential expression among the genes with a p-value $\leq x$

FN(x) = the expected proportion of genes with a true differential expression among the genes with a p-value $> x$

For a randomly selected gene let H_0 and H_1 the following events

H_0 = this gene does not have a differential expression

H_1 = this gene is truly differentially expressed

According to our mix-o-matic model $P(H_0) = \lambda_0$ and $P(H_1) = \lambda_1 = 1 - \lambda_0$. In the mix-o-matic model it is further assumed that if a gene does not have a true differential expression than the p-value has a uniform distribution between 0 and 1 and if this gene is truly differentially expressed than the p-value is assumed to have a beta distribution with parameters r_1 and s_1 . This we can express by the probability statements $P(\text{p-value} \leq x | H_0) = x$ and

$P(\text{p-value} \leq x | H_1) = \text{pbeta}(x/r_1, s_1)$, where $\text{pbeta}(x/r_1, s_1)$ is the cumulative distribution function of a random variable with a beta distribution with parameters r_1 and s_1 . We are now ready to derive formulas for the quantities TP(x), FP(x) and FN(x).

$$\begin{aligned}
\text{TP}(x) &= P(H_1 | \text{p-value} \leq x) \\
&= \frac{P(H_0 \cap \{\text{p-value} \leq x\})}{P(\text{p-value} \leq x)} \\
&= \frac{1 - \frac{P(\text{p-value} \leq x | H_0) * P(H_0)}{P(\text{p-value} \leq x | H_0) * P(H_0) + P(\text{p-value} \leq x | H_1) * P(H_1)}}{1} \\
&= \frac{\lambda_0 x}{\lambda_0 x + \lambda_1 \text{pbeta}(x | r_1, s_1)}
\end{aligned}$$

$$\begin{aligned}
\text{FP}(x) &= P(H_0 | \text{p-value} \leq x) \\
&= \frac{P(H_0 \cap \{\text{p-value} \leq x\})}{P(\text{p-value} \leq x)} \\
&= \frac{P(\text{p-value} \leq x | H_0) * P(H_0)}{P(\text{p-value} \leq x | H_0) * P(H_0) + P(\text{p-value} \leq x | H_1) * P(H_1)} \\
&= \frac{\lambda_0 x}{\lambda_0 x + \lambda_1 \text{pbeta}(x | r_1, s_1)}
\end{aligned}$$

$$\begin{aligned}
\text{FN}(x) &= P(H_1 | \text{p-value} > x) \\
&= \frac{1 - \frac{P(H_0 \cap \{\text{p-value} > x\})}{P(\text{p-value} > x)}}{1} \\
&= \frac{1 - \frac{P(\text{p-value} > x | H_0) * P(H_0)}{P(\text{p-value} > x | H_0) * P(H_0) + P(\text{p-value} > x | H_1) * P(H_1)}}{1} \\
&= \frac{1 - \frac{[1 - P(\text{p-value} \leq x | H_0)]P(H_0)}{[1 - P(\text{p-value} \leq x | H_0)]P(H_0) + [1 - P(\text{p-value} \leq x | H_1)]P(H_1)}}{1} \\
&= \frac{(1-x)\lambda_0}{(1-x)\lambda_0 + (1 - \text{pbeta}(x | r_1, s_1))\lambda_1}
\end{aligned}$$

Notice that $\text{TP}(x) = 1 - \text{FP}(x)$. The quantities $\text{TP}(x)$, $\text{FP}(x)$ and $\text{FN}(x)$ are computed by plugging in Maximum Likelihood Estimates (MLE) $\hat{\lambda}_0$, $\hat{\lambda}_1$, \hat{r}_1 and \hat{s}_1 of the unknown parameters λ_0 , λ_1 , r_1 and s_1 .

Posterior True Positive (PTP)

$\text{PTP}(x)$ = The proportion of genes with a true differential expression among the genes which are declared interesting via p-value $\leq x$

$$= P(\bar{H}_0 | \text{p-value} \leq x) = 1 - \frac{P(H_0 \cap \text{p-value} \leq x)}{P(\text{p-value} \leq x)}$$

Let H_j , be the event that the v number of p-values follow a $\text{beta}(r_j, s_j)$ distribution, then

$$P(H_0 \cap \text{p-value} \leq x) = P(\text{p-value} \leq x | H_0) * P(H_0) = \lambda_0 * x \text{ and}$$

$$\begin{aligned} P(\text{p-value} \leq x) &= P(\text{p-value} \leq x \cap H_0) + \sum_{j=1}^v P(\text{p-value} \leq x \cap H_j) \\ &= P(\text{p-value} \leq x | H_0) * P(H_0) + \sum_{j=1}^v P(\text{p-value} \leq x | H_j) * P(H_j) \\ &= \lambda_0 * x + \sum_{j=1}^v \lambda_j * \text{pbeta}(x | r_j, s_j) \end{aligned}$$

so that,

$$\text{PTP}(x) = 1 - \frac{\lambda_0 * x}{\lambda_0 * x + \sum_{j=1}^v \lambda_j * \text{pbeta}(x | r_j, s_j)} = \frac{\sum_{j=1}^v \lambda_j * \text{pbeta}(x | r_j, s_j)}{\lambda_0 * x + \sum_{j=1}^v \lambda_j * \text{pbeta}(x | r_j, s_j)}$$

Posterior False Positive (PFP)

PFP(x) = The proportion of genes with no differential expression among the genes which are declared interesting via $\text{p-value} \leq x$

$$\begin{aligned} &= P(H_0 | \text{p-value} \leq x) = 1 - P(\bar{H}_0 | \text{p-value} \leq x) = 1 - \text{TP}(x) = \\ &= \frac{\lambda_0 * x}{\lambda_0 * x + \sum_{j=1}^v \lambda_j * \text{pbeta}(x | r_j, s_j)} \end{aligned}$$

Posterior False Negative (PFN)

PFN(x) = The proportion of genes with a true differential expression among the genes which are declared as **not** interesting via $\text{p-value} > x$.

$$\begin{aligned} &= P(\bar{H}_0 | \text{p-value} > x) = 1 - \frac{P(H_0 \cap \text{p-value} > x)}{P(\text{p-value} > x)} \\ P(H_0 \cap \text{p-value} > x) &= P(\text{p-value} > x | H_0) * P(H_0) = \lambda_0 * (1 - x) \\ P(\text{p-value} > x) &= P(\text{p-value} > x \cap H_0) + \sum_{j=1}^v P(\text{p-value} > x \cap H_j) \\ &= P(\text{p-value} > x | H_0) * P(H_0) + \sum_{j=1}^v P(\text{p-value} > x | H_j) * P(H_j) \\ &= \lambda_0 * (1 - x) + \sum_{j=1}^v \lambda_j * (1 - \text{pbeta}(x | r_j, s_j)) \end{aligned}$$

$$PFN(x) = 1 - \frac{\lambda_0 * (1-x)}{\lambda_0 * (1-x) + \sum_{j=1}^v \lambda_j * (1-pbeta(x|r_j, s_j))} = \frac{\sum_{j=1}^v \lambda_j * (1-pbeta(x|r_j, s_j))}{\lambda_0 * (1-x) + \sum_{j=1}^v \lambda_j * (1-pbeta(x|r_j, s_j))}$$

The mix-o-matic method is applied to the ordinary t-test p-values for the original as well as the log-transformed data. The mix-o-matic, however, is not applied to the Chebby Checker p-values, because these p-values are not expected to be uniformly distributed under H0. For both sets of p-values separately the following results are presented:

1. Graphs are presented in which the observed histogram of p-values is compared to the distribution of three fitted mixture models with respectively one uniform and one beta component, one uniform and two beta components and one uniform and three beta components. By means of visual inspection the most appropriate model is chosen;
2. The parameters representing the proportions of the uniform and beta components of the models found in 1 are presented in a table. An estimate of the proportion and number of genes for which there is a real difference in expression between the two groups A and B. This proportion and number are estimated by $1 - \hat{\lambda}_0$ and $(1 - \hat{\lambda}_0)N$, where $\hat{\lambda}_0$ is the MLE estimate of λ_0 in the model chosen in 1 and N is the total number of genes;
3. The posterior probabilities corresponding to the mixture model chosen in 1 in a graph. In this graph, the x-axis represent a threshold value x such that all genes with a p-value $\leq x$ are chosen as the ‘interesting’ genes worthwhile for follow-up study. The value of y-axis of this graph corresponding to this threshold x is equal to an estimate of the proportion of genes for which there is a true differential expression among the genes declared as ‘interesting’ according to this threshold x . This estimate is based on the fitted mixture model chosen in 1. If for instance an x-axis value of 0.05 correspond to an y-axis value of 0.8, then the fraction of genes for which there is a true differential expression among the genes for which the p-value ≤ 0.05 is equal to 0.8;

The corresponding posterior probabilities in 3 are also presented in the excel data set in the column with sub – header “posterior probability”. For a particular gene, this probability is an estimate of a proportion of genes with a true differential expression among the genes with a corresponding p-value smaller than or equal to the p-value of this gene.

A graph is presented in which the observed histogram of the p-values is compared to the distribution of the fitted mix-o-matic model. This graph can be used to visually inspect if the mix-o-matic model adequately fit to the probability distribution of the p-values. From this graph we can also get a visual impression if there is evidence in the microarray data with respect to differentially expressed genes. The more concentrated the p-values are distributed around values very close to zero, the more we expect that there is evidence in the data.

2. Result Files

List of files (In alphabetical order)

- Chip correlation.xls – pre-normalization chip to chip Pearson's correlation
- Chip statistics.xls – simple description statistics of pre-normalization chips
- Deleted residual plot #.png – deleted residual plot of each chip in each hypothesis it is tested in. Line 2 of the image is the chip for which deleted residuals have been generated. Line 3 is the rest of the chips in the group.
- Deleted residuals #.csv – calculated deleted residuals for each hypothesis
- Deleted residuals standard outliers.xls – deleted residual outliers
- Deleted residuals statistics.xls - simple description statistics of deleted residuals (each hypothesis is a different worksheet in the file)
- Mix-o-matic histogram.png – graph showing distribution of p-values for each hypothesis.
- Mix-o-matic posterior PDF.png – graph showing distribution of PTP, PTN, PFN. and EDR at each p-value cut-off from 0 to 1.
- mix-o-matic table.xls – mix-o-matic parameters lambda0, R and S
- preprocessed data.xls – post-normalization data
- probe results \$ #x&.cvs - results table where \$ is the variable tested, # is name of group 1 and & is name of group 2. See below for description of columns.
- Scatter plot raw.png – scatter plot of raw chip mean by standard deviation
- Scatter plot.png- scatter plot of normalized chip mean by standard deviation
- Summary report.html – HTML document describing details of data analysis and some results.

2.1 Result Files

Summary of report – The summary report is an HTML file that describes some of the steps that were taken in analysis. It is composed of very parts

- a. Data description – variables that were studied and the chips in each group of the hypotheses
- b. Data preprocessing – list of normalization and transformation methods used
- c. Data analysis – the statistical tests and multiple testing adjustments used
- d. Chip level statistics on raw data- self descriptive
- e. Chip level statistics on processed data- self descriptive
- f. Summary of deleted residuals by chip – self descriptive- this is where the empirical assessment of chip quality will be addressed.
- g. Mix-o-matic parameter estimates – lambda 0 is estimate of the number of genes that are not different between groups. r and s are beta shape parameters. If the restricted and

unrestricted values are similar the fit is usually good otherwise there may be some issues with the data.

- h. Empirical Bayes estimate parameter – values from EB analysis.

References

1. Allison, D. B., Gadbury G. L., Moonseong, H., Fernandez, J. R., Lee, C., Prolla, T. A., Weindruch, R. (2002). A mixture model approach for the analysis of microarray gene expression data. *Computational Statistics & Data Analysis*, 39, 1-20.
2. Beasley, T. M., Page, G. P., Brand, J. P. L., Gadbury, G. L., Mountz, J. D., & Allison, D. B. (2004). Chebyshev's inequality for non-parametric testing with small N and a in microarray research. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 53, 95-108.
3. Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.*, 57, 289-300.
4. Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat*, 29 (4): 1165-1188.
5. Bland, J. M. and Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ*, 310 (6973), 170.
6. Cui, X., Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4), 210.
7. DasGupta, A. (2000) Best constants in Chebychev inequalities with various applications. *Metrika*, 51, 185–200.
8. Davison, A. C. and Hinkley, D. V. (1997). Bootstrap methods and their application. Cambridge University Press, United Kingdom.
9. Edwards, J. W., Page, G. P., Gadbury, G., Heo, M., Kayo, T., Weindruch, R., Allison, D. B. (*In press*). Empirical Bayes (EB) estimation of gene specific effects in microarray research. *Functional and Integrative Genomics*.
10. Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman and Hall, New York.
11. Gadbury, G.L., Page, G. P., Edwards, J. W., Kayo, T., Prolla, T. A., Weindruch, R., Permana, P. A., Mountz, J., Allison, D. B. (*In press*). Power and Sample Size Estimation in High Dimensional Biology. *Statistical Methods in Medical Research*.
12. Hatfield, G. W., Hung, S. P., Baldi, P. (2003). Differential analysis of DNA microarray gene expression data. *Mol Microbiol*, 47(4): 871-7.
13. Hedges LV, Olkin L. Statistical Methods for Meta-analysis. 1985.

14. Morris, C. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, 78, 47-59.
15. Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L.V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244). New York: Russell Sage Foundation.
16. Satterthwaite, F.W. (1946). An Approximate Distribution of Estimates of Variance Components, *Biometrics Bulletin*, 2, 110-114
17. Sidak, Z. (1967). Rectangular confidence regions for the means of the multivariate normal distributions. *J Am Stat Assoc*, 62, 626–633.
18. Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.