

Research article

Open Access

## A two-sample Bayesian t-test for microarray data

Richard J Fox\*<sup>1</sup> and Matthew W Dimmic<sup>2,3</sup>

Address: <sup>1</sup>Codexis, Inc., Redwood City, CA 94063, USA, <sup>2</sup>Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY 14853, USA and <sup>3</sup>Divergence, Inc., St. Louis, MO 63141, USA

Email: Richard J Fox\* - richard.fox@codexis.com; Matthew W Dimmic - matt.dimmic@niduscenter.com

\* Corresponding author

Published: 10 March 2006

Received: 18 October 2005

BMC Bioinformatics 2006, 7:126 doi:10.1186/1471-2105-7-126

Accepted: 10 March 2006

This article is available from: <http://www.biomedcentral.com/1471-2105/7/126>

© 2006 Fox and Dimmic; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Determining whether a gene is differentially expressed in two different samples remains an important statistical problem. Prior work in this area has featured the use of t-tests with pooled estimates of the sample variance based on similarly expressed genes. These methods do not display consistent behavior across the entire range of pooling and can be biased when the prior hyperparameters are specified heuristically.

**Results:** A two-sample Bayesian t-test is proposed for use in determining whether a gene is differentially expressed in two different samples. The test method is an extension of earlier work that made use of point estimates for the variance. The method proposed here explicitly calculates in analytic form the marginal distribution for the difference in the mean expression of two samples, obviating the need for point estimates of the variance without recourse to posterior simulation. The prior distribution involves a single hyperparameter that can be calculated in a statistically rigorous manner, making clear the connection between the prior degrees of freedom and prior variance.

**Conclusion:** The test is easy to understand and implement and application to both real and simulated data shows that the method has equal or greater power compared to the previous method and demonstrates consistent Type I error rates. The test is generally applicable outside the microarray field to any situation where prior information about the variance is available and is not limited to cases where estimates of the variance are based on many similar observations.

### Background

Determining whether a gene is differentially expressed under different conditions is an important statistical problem [1-3]. Consider, for example, a common microarray experiment: for a particular gene  $Y$ , one has an unordered set of measurements of log-expression levels  $\gamma_1^c \cdots \gamma_{n_1}^c$  in a control situation and  $\gamma_1^t \cdots \gamma_{n_2}^t$  in a treatment situation (or any non-control situation) (notation similar

to [4]). The question of interest is whether any expression

difference  $\Delta\bar{y} = \sum_{j=1}^{n_2} \gamma_j^t / n_2 - \sum_{i=1}^{n_1} \gamma_i^c / n_1 = \bar{y}_2 - \bar{y}_1$  is a signifi-

cant difference, or if it would be expected under the null hypothesis of no actual difference in expression.

Though multifactorial experimental designs are becoming increasingly popular [5-7], there continue to be experimentalists interested in analyzing two-sample designs.

There are many approaches to determining whether genes are differentially expressed in such designs and there are a number of excellent reviews of the various methods [2,3]. Many of the approaches include the use of a *t*-test [8-14], which is a common frequentist statistical approach to comparing a difference in sample means. Some have pointed out that the normality assumption used in *t*-tests may not always hold and other statistical techniques may be warranted [15,16]. Even for those measurements where normality holds the *t*-test is difficult to apply because the number of replicates  $n_1$  and  $n_2$  are often quite small (typically,  $n_1 = n_2 = 1$  to 3), leading to uncertainty in the sample variance and relatively low power. An early approach to dealing with this problem was based on the addition of a constant value to the estimate of the standard deviation [16,17]. While the approach does tend to regularize the variance estimates, it is *ad hoc* and is not expected to exhibit statistically consistent behavior (unknown or incorrect Type I error rate under the null).

More rigorous Bayesian methods have been developed that incorporate prior information about the variance but require Markov chain sampling of the posterior density or numerical integration of the cumulative distribution function [18-21]. A very popular work by Baldi and Long [4] avoids such calculations while using a statistically justified regularization technique via construction of a probabilistic Bayesian framework that applies a prior probability to parameters of interest such as the expression means and variances. An analytically tractable solution can then be obtained by taking a point estimate for the variance. The chief advantages of this Bayesian framework are twofold: it allows the use of a regularized *t*-test to determine whether a difference in expression is significant, and it provides a natural method for incorporating information from putatively related measurements. It is this basic framework that we seek to extend in the work presented here.

As noted by Baldi and Long [4], the use of a point estimate for the variance is a compromise between a rigorous Bayesian approach and tractable calculation. As demonstrated below, this compromise can lead to bias in the specificity of the test and improper behavior when there are few prior degrees of freedom. Our hope here is to correct this and other issues (both theoretical and practical) towards the goal of improving the performance and understanding of microarray and general laboratory data analysis. To address these issues, we extend the basic framework put forth earlier, taking the model to a fully Bayesian approach by explicitly calculating the marginal posterior distribution for the difference in mean expression levels. This analytically tractable solution does not require simulation from the posterior distribution and obviates the need to use point estimates of the prior variance while

overcoming the undesirable correlation between specificity and prior degrees of freedom. Moreover, a clear and statistically rigorous connection between the prior variance and the prior degrees of freedom is established, reducing the number of hyperparameters from two to one while yielding consistent Type I error rates in the process. It should be stressed that achieving consistent Type I error rates is not just a conceptual luxury. Indeed proper determination of important quantities such as the false discovery rate are of great economical significance and rely on having correct Type I error rates [22].

**Method**

**Bayesian probability model**

The original model proposed by Baldi and Long [4] is briefly reviewed here. The likelihood of the observed data,  $y$ , given the true expression level,  $\mu$ , and variance,  $\sigma^2$ , for a single gene is assumed to follow a normal distribution:

$$p(y | \mu, \sigma^2) \propto \prod_{i=1}^n N(y_i; \mu, \sigma^2) \tag{1}$$

where  $n$  is the number of measurements in one sample. For the priors on  $\mu$  and  $\sigma^2$  the authors attempt to capture the *a priori* dependence between  $\mu$  and  $\sigma^2$  by factoring the joint prior as  $p(\mu, \sigma^2) = p(\mu | \sigma^2)p(\sigma^2)$  and taking each factor as

$$p(\mu | \sigma^2) = N(\mu; \mu_0, \sigma^2 / \lambda_0) \tag{2}$$

and

$$p(\sigma^2) = I(\sigma^2; \nu_0, \sigma_0^2) \tag{3}$$

where the prior probability of  $\mu$  follows a normal distribution and  $\sigma^2$  follows a scaled inverse gamma distribution. The hyperparameters  $\mu_0$  and  $\lambda_0$  represent the prior beliefs regarding the value of  $\mu$  and the confidence associated with that belief, while the hyperparameters  $\sigma_0^2$  and  $\nu_0$  represent the prior beliefs regarding the value of  $\sigma^2$  and the degrees of freedom or confidence associated with that belief. The authors subsequently let  $\lambda_0 \rightarrow 0$  later in their derivation, leading to a diffuse, noninformative prior on  $\mu$ :

$$\lim_{\lambda_0 \rightarrow 0} p(\mu | \sigma^2) \propto \frac{1}{\sigma} \tag{4}$$

The remaining  $1/\sigma$  factor that comes from a diffuse normal results in a prior of the form  $p(\mu, \sigma^2) \propto \sigma^{-1} I(\sigma^2; \nu_0, \sigma_0^2)$ .

Using Bayes' rule the authors obtain the following posterior distribution:

$$p(\mu, \sigma^2 | \mathbf{y}, \alpha) \propto N(\mu; \mu_n, \sigma^2 / \lambda_n) I(\sigma^2; \nu_n, \sigma_n^2) \quad (5)$$

where  $\alpha$  is a vector of hyperparameters  $(\mu_0, \lambda_0, \nu_0, \sigma_0^2)$  for the prior and

$$\mu_n = \frac{\lambda_0}{\lambda_0 + n} \mu_0 + \frac{n}{\lambda_0 + n} \bar{y} \quad (6)$$

$$\lambda_n = \lambda_0 + n \quad (7)$$

$$\nu_n = \nu_0 + n \quad (8)$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n-1) s^2 + \frac{\lambda_0 n}{\lambda_0 + n} (\bar{y} - \mu_0)^2 \quad (9)$$

The sufficient statistics are  $\bar{y} = \sum_{i=1}^n y_i / n$  and

$$s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n-1).$$

Finally, the authors calculate the posterior mean estimate or point estimate (PE) of the variance for one sample as

$$\sigma^2 = \frac{\nu_n}{\nu_n - 2} \sigma_n^2 \quad (10)$$

This estimate is then used in the hypothesis testing procedures implemented in their Cyber-T software. Though this model is developed using a Bayesian framework to make a point estimate of the variance, the hypothesis testing is carried out using a classical frequentist *t*-test.

In the standard *t*-test, the variance is still defined when there are no prior degrees of freedom, but inspection of Eq. (10) shows that when the number of prior degrees of freedom  $\nu_0 < 2 - n$ , the variance is undefined. This limiting behavior stems from the use of a point estimate for the variance. The total number of prior degrees of freedom used in the Cyber-T software is actually  $2\nu_0$  as it counts  $\nu_0$  prior degrees of freedom for each sample in the two-sample test. However, for consistency and clarity we hereafter refer to the total number of prior degrees of freedom as  $\nu_0$ . Therefore the total number of degrees of freedom used in the Cyber-T software for the two-sample hypothesis test then appears to be  $\nu_{t-test} = \nu_0 + n_1 + n_2 - 4$ , where  $n_1$  is the number of experimental replicates of the gene in the control regime and  $n_2$  is the number of replicates of the gene in the treatment regime. This equation for the total degrees of freedom also demonstrates undesired limiting

behavior when  $\nu_0 = 0$ , since positive degrees of freedom exist only when  $n_1 + n_2 > 4$ . This result contradicts what we would expect to see from a classic two-sample *t*-test with no prior estimate of the variance, i.e. a classical two-sample *t*-test has degrees of freedom  $n_1 + n_2 - 2$  and positive degrees of freedom exist when  $n_1 + n_2 > 2$ .

### Full two-sample model

#### Posterior distribution

Instead of using a point estimate for the variance to conduct a frequentist *t*-test, we can recast the problem to directly infer the posterior distribution for a two-sample case:

$$p(\mu, \Delta\mu, \sigma^2 | \mathbf{y}) \propto p(\mu, \Delta\mu, \sigma^2) p(\mathbf{y} | \mu, \Delta\mu, \sigma^2) \quad (11)$$

where  $\mu$  is the mean expression level of the gene in the control regime,  $\mu + \Delta\mu$  the mean of the gene in the alternate experimental regime, and  $\sigma^2$  the variance associated with the measurements. In contrast to the previous, point estimate model, in this formulation there is no explicit prior dependence between  $\sigma^2$  and  $\mu$ . While this is admittedly a simplification, this independence is not only a practical requirement for the full Bayesian integration, we also believe it is justified on several other grounds. First, while it has been widely observed that genes with different expression levels often have different variances [1], it is not clear how the prior used by Baldi and Long [4] to derive the point estimate of variance could capture the dependencies observed with actual microarray data. Their figures show that the logarithm of the mean expression level tends to decrease with increasing variance, yet the prior given by Eq. (2) asserts that the prior probability of the mean should be more diffuse – not necessarily smaller – with increasing variance. The observed trend between mean and variance cannot be captured by such a prior. Second, their method does include an *implicit* dependence between mean and variance by using other gene expression levels to determine the priors; indeed, this is what lends the method its power. Because the Bayesian formulation proposed here determines the prior hyperparameters in a similar fashion, it retains this implicit dependence. One may also consider a number of variance stabilizing transforms that may be applied in order to achieve constant variance across the expression spectrum [23,24]. Finally, as shown below, simulations reveal that the method still retains power even when the true variance is allowed to vary over a substantial range.

Returning to the posterior distribution, we assume that the samples from each experimental regime follow a normal distribution, each with equal variances and (possibly) different means.

$$y_i \sim N(\mu, \sigma^2) \quad (12)$$

$$y_j \sim N(\mu + \Delta\mu, \sigma^2) \quad (13)$$

This leads to the following posterior distribution

$$p(\mu, \Delta\mu, \sigma^2 | \mathbf{y}) \propto \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{1}{2\sigma^2}((y_j - (\mu + \Delta\mu))^2)\right) \quad (14)$$

where  $n_1$  and  $n_2$  are the number of measurements in each sample. We elect not to make use of the prior on  $\mu$  used in the Baldi work, Eq. (2), or its limiting form, Eq. (4). Instead, we assume a completely flat prior on both  $\mu$  and  $\Delta\mu$  while retaining the prior for  $\sigma^2$  given by Eq. (3). Similar to the earlier work, the observed dependence between  $\mu$  and  $\sigma^2$  is established by calculating  $\sigma_0^2$  based on the variance of similarly expressed genes. However, in this formulation, instead of taking the average standard deviation of similarly expressed genes we choose to estimate the prior variance by totaling the sum-squared differences for each similar gene and dividing by the total prior degrees of freedom. This is a more statistically rigorous way of incorporating prior information and leads to a more consistent test as will be discussed below.

*Marginal posterior for  $\Delta\mu$*

With the above definitions and assumptions in place, it can be shown (see Appendix) that the marginal posterior distribution follows a  $t$  distribution:

$$\frac{\Delta\mu - \Delta\bar{y}}{\sigma_n \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \Bigg| y \sim t_{v_n} \quad (15)$$

where

$$v_n = n_1 + n_2 + v_0 - 2 \quad (16)$$

$$v_n \sigma_n^2 = v_0 \sigma_0^2 + (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 \quad (17)$$

The sufficient statistics are

$$\Delta\bar{y} = \sum_{j=1}^{n_2} y_j / n_2 - \sum_{i=1}^{n_1} y_i / n_1 = \bar{y}_2 - \bar{y}_1, \text{ and the sum squared}$$

differences  $(n_1 - 1)s_1^2$  and  $(n_2 - 1)s_2^2$ , where

$$s_1^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 / (n_1 - 1) \quad \text{and}$$

$$s_2^2 = \sum_{j=1}^{n_2} (y_j - \bar{y}_2)^2 / (n_2 - 1). \text{ This distribution has the fea-}$$

ture that the standard  $t$ -test is obtained even when there are no prior degrees of freedom,  $v_0 = 0$ , in which case the observed variance for each gene dominates the posterior density. A hypothesis test is performed by asserting a null hypothesis that the true difference in expression levels is zero, i.e.  $\Delta\mu = 0$  (though other values for the true difference,  $\Delta\mu$ , can be used when attempting to identify genes that are differentially expressed by some threshold degree). When the posterior probability of no differential expression approaches zero,  $\Pr(\Delta\mu = 0 | \mathbf{y}) \rightarrow 0$ , the null is rejected. It is worth noting that while the posterior probability distribution for no differential expression,  $\Pr(\Delta\mu = 0 | \mathbf{y})$ , follows the same frequentist distribution for the data,  $\Pr(\mathbf{y} | \Delta\mu = 0)$  [25], the resulting probabilities have different interpretations. In the Bayesian approach the posterior probability is interpreted as the probability that there is no difference in expression levels. In the frequentist approach the probability is interpreted as the probability of observing such a difference from the null distribution of no differential expression given chance alone. When  $\Delta\mu = 0$  the Bayesian and frequentist methods give the same numerical results and one can ignore the interpretational differences when discussing such things as the false positive rate under the null. However, power analysis calculations ( $\Delta\mu \neq 0$ ) are best conducted under a frequentist framework using a non-central  $t$ -distribution [25].

*Estimating the prior variance*

For  $m$  similarly expressed genes each having  $n$  replicates, the prior degrees of freedom for the variance can be calculated as

$$v_0 = m(n - 1) \quad (18)$$

The prior variance is then calculated as the total sum-square differences for each sample of similarly expressed genes divided by the prior degrees of freedom, namely

$$\sigma_0^2 = \frac{\sum_{k=1}^m \sum_{i=1}^n (y_{k,i} - \bar{y}_k)^2}{v_0} \quad (19)$$

where  $\bar{y}_k$  is the mean response of gene  $k$  and  $y_{k,i}$  is the response  $i$  of gene  $k$ . In general the prior degrees of freedom and variance should be determined using these equations and not chosen separately from one another. While it is tempting to think of the prior variance as separate from the precision or credibility associated with that

**Table 1: Comparison of false positive rate and statistical power for fixed  $\alpha$ . For a given significance level ( $\alpha$ ) the observed false positive rate (FP) and power were determined by simulation using either a point estimate for the variance (PE) or the marginal posterior distribution for  $\Delta\mu$  (MP). Values are given for different sample sizes ( $n_1 = n_2 = n$ ) and prior degrees of freedom ( $v_0$ ).**

| $n_1 = n_2 = n$ | $v_0$ | $\alpha_{PE}$ | $\alpha_{MP}$ | FP <sub>PE</sub> | FP <sub>MP</sub> | Power <sub>PE</sub> | Power <sub>MP</sub> |
|-----------------|-------|---------------|---------------|------------------|------------------|---------------------|---------------------|
| 2               | 0     | 0.05          | 0.05          | -                | 0.0509           | -                   | 0.2169              |
| 2               | 2     | 0.05          | 0.05          | 0.0044           | 0.0490           | 0.0450              | 0.3343              |
| 2               | 4     | 0.05          | 0.05          | 0.0146           | 0.0535           | 0.1814              | 0.3919              |
| 2               | 6     | 0.05          | 0.05          | 0.0211           | 0.0518           | 0.2747              | 0.4182              |
| 2               | 16    | 0.05          | 0.05          | 0.0385           | 0.0498           | 0.4186              | 0.4709              |
| 3               | 0     | 0.05          | 0.05          | 0.0030           | 0.0500           | 0.0709              | 0.4664              |
| 3               | 2     | 0.05          | 0.05          | 0.0167           | 0.0486           | 0.2833              | 0.5404              |
| 3               | 4     | 0.05          | 0.05          | 0.0261           | 0.0556           | 0.4153              | 0.5834              |
| 3               | 6     | 0.05          | 0.05          | 0.0280           | 0.0500           | 0.4838              | 0.6053              |
| 3               | 16    | 0.05          | 0.05          | 0.0389           | 0.0501           | 0.5951              | 0.6441              |

estimate (as represented by  $v_0$ ), a consistent hypothesis test requires that they be considered together based on the actual prior data collected.

**Results**

To assess the performance of the new test, a series of simulations were done to measure the false positive rate and statistical power under various conditions. The simulations consisted of the following steps:

1) Draw random samples from two normal distributions of size  $n_1$  and  $n_2$  distributed as  $N(\mu_1, \sigma_{test}^2)$  and  $N(\mu_2, \sigma_{test}^2)$  respectively. This step represents the selection of two different experimental regimes for a gene of interest in order to test for differential expression. For power testing, the standardized effect  $(\mu_1 - \mu_2)/\sigma_{test} = 2$  is used to generate the data by setting  $\mu_1 = 2$ ,  $\mu_2 = 0$ , and  $\sigma_{test} = 1$ . For false positive rate testing  $\mu_1 = 0$ .

2) Draw an estimate of the prior variance from the following Chi-square distribution:  $\sigma_0^2 \sim \sigma_{test}^2 \chi_v^2 / v_0$ . This step simulates the task of estimating the prior variance from similarly expressed genes. It is mathematically equivalent to calculating the sum-squared differences of similarly expressed genes and dividing by their prior degrees of freedom, but it is computationally more efficient.

3) Perform hypothesis tests using a) the method used by Baldi and Long [4] based on the mean posterior point estimate for the variance, PE( $\sigma^2$ ) or just PE, and b) the new test proposed here based on the marginal posterior for the difference in expression level, MP( $\Delta\mu$ ) or just MP. The hypothesis tests using each method are done at the  $\alpha = 0.05$  level using a two-sided, two-sample  $t$ -test. Test  $p$ -values below this level are counted as significant. To make

valid comparisons, both methods used the same number of prior degrees of freedom.

4) Steps 1 to 3 above are repeated 10,000 times.

Results of the power simulations for a number of sample sizes and prior degrees of freedom are shown in Table 1. Because the PE test cannot be used when  $v_0 + n_1 + n_2 \leq 4$ , the test has no power and no false positive rate when  $n_1 = n_2 = 2$  and  $v_0 = 0$ . Conversely, the MP test results in a classical  $t$ -test in this case and has a power of 0.2169 and a false positive rate of 0.0509, which is near the expected value given the  $\alpha = 0.05$  level of the test. Holding  $n_1 = n_2 = 2$  but increasing the number of prior degrees of freedom to  $v_0 = 2$  (equal to 2 other similarly expressed genes used to determine the prior variance) the PE test can now be used but has a very low power of 0.0450, compared to the MP power of 0.3343. Overall, when the estimate of the variance is balanced between the prior and observed variances (i.e. the prior number of degrees of freedom is not large compared to the number of replicates in the test samples) the MP test is significantly more powerful. The difference in power between the two methods is less pronounced when  $v_0 = 16$  (the default value used in the Cyber-T software when  $n_1 = n_2 = 2$ ). However, one should note that such a relatively high number for the prior degrees of freedom represents a strong degree of belief in the prior estimate of the variance relative to the variance in the genes of interest.

These simulations reveal that the observed false positive rate for the PE test is substantially out of line with the nominal value of 0.05 for small sample replicates ( $n_1 = n_2 \leq 3$ ) and few prior degrees of freedom ( $v_0 \leq 16$ ). Conversely, the MP test demonstrates consistent behavior for the observed false positive rate. However, it is possible to obtain the desired false positive rate for the PE method by iterating through values of  $\alpha$ . Table 2 shows that, by adjusting the critical  $\alpha$  value for the PE method until the

**Table 2: Comparison of  $\alpha$  and statistical power for iterated false positive rate. The values are the same as those given in Table 1 except the significance level ( $\alpha$ ) was iterated for the PE method until its power matched that of the MP method where  $\alpha$  was held constant.**

| $n_1 = n_2 = n$ | $v_0$ | $\alpha_{PE}$ | $\alpha_{MP}$ | FP <sub>PE</sub> | FP <sub>MP</sub> | Power <sub>PE</sub> | Power <sub>MP</sub> |
|-----------------|-------|---------------|---------------|------------------|------------------|---------------------|---------------------|
| 2               | 0     | -             | 0.05          | -                | 0.0505           | -                   | 0.2146              |
| 2               | 2     | 0.190         | 0.05          | 0.0514           | 0.0489           | 0.3374              | 0.3336              |
| 2               | 4     | 0.120         | 0.05          | 0.0483           | 0.0522           | 0.3783              | 0.3840              |
| 2               | 6     | 0.095         | 0.05          | 0.0496           | 0.0514           | 0.4302              | 0.4267              |
| 2               | 16    | 0.067         | 0.05          | 0.0491           | 0.0508           | 0.4794              | 0.4738              |
| 3               | 0     | 0.190         | 0.05          | 0.0507           | 0.0513           | 0.4576              | 0.4599              |
| 3               | 2     | 0.115         | 0.05          | 0.0531           | 0.0462           | 0.5324              | 0.5308              |
| 3               | 4     | 0.095         | 0.05          | 0.0475           | 0.0486           | 0.5611              | 0.5641              |
| 3               | 6     | 0.082         | 0.05          | 0.0536           | 0.0500           | 0.5994              | 0.6053              |
| 3               | 16    | 0.066         | 0.05          | 0.0511           | 0.0472           | 0.6531              | 0.6469              |

observed false positive rate matches that of the MP method, both tests have similar power. The disadvantage is that  $\alpha$  no longer represents the expected Type I error rate, and simulation or calibration is required whenever a new test is performed. Conversely, with the MP test, the critical value  $\alpha$  represents an unbiased estimate of the Type I error rate.

It is interesting to look for a simple correction to the degrees of freedom in order to make the two methods match each other. Unfortunately, there does not appear to be any simple correction available. The degrees of freedom used in the *t*-test itself would be easy to modify – though *ad hoc*, one could change the last term in the expression for the degrees of freedom used in the PE test ( $v_{t-test} = v_0 + n_1 + n_2 - 4$ ) from -4 to -2 by simply adding two to  $v_0$ . Unfortunately, the point estimate of the variance given by Eq. (10) is more problematic as it appears to give an increased estimate of the variance compared to the one used in the MP test, resulting in a decrease in both the *t*-statistic and the statistical power.

In the Cyber-T software [4] the prior variance is estimated by averaging the standard deviation over 100 similarly expressed genes. This has two problems associated with it. First, the measured standard deviation is a biased estimate and tends to be smaller than the true value for small sample sizes like  $n = 2$  or  $3$ , the average of these measured standard deviations will also be biased downward and in general  $E(\sqrt{\sigma^2}) \neq \sqrt{E(\sigma^2)}$ , where  $E(\cdot)$  represents the expected value of a random variable. The correct way to estimate the variance based on prior observations is to calculate the total sum square differences for *all* the prior samples and divide by the total degrees of freedom, as shown in Eq. (19). Second, if similarly expressed genes for each of the two samples are used to estimate the prior variance and they have the same number of replicates,  $n$ , as

the samples themselves, then estimating the standard deviation from  $m = 100$  similarly expressed genes is tantamount to asserting  $v_0 = 100 \cdot (n - 1)$  prior degrees of freedom. Setting  $v_0$  to less than the actual number of prior degrees of freedom used to calculate the prior variance results in a lower actual false positive rate than the rejection level  $\alpha$ . For consistency if we fix the number of prior degrees of freedom that are used in the test, then we must choose a corresponding number of similarly expressed genes that together have the same number of prior degrees of freedom. This dependence between  $v_0$  and  $m$  is more consistent with the formulation of the prior itself and leads to consistent Type I error rates. When we calculate the prior variance in the statistically rigorous manner (by taking the total sum square differences of prior observations and dividing by the total prior degrees of freedom) the PE method becomes overly conservative, generating false positives with a rate of 0.0044 at  $n_1 = n_2 = 2$ ,  $\alpha = 0.05$ , and  $v_0 = 2$ , this appears to be the result of using a point estimate for the prior variance that is larger than the expected value given similar genes, see Eq. (10).

As mentioned previously, to perform the full Bayesian integration required by the MP test, the assumption of a constant variance is required. To test the effect of violating this assumption, simulations were performed as described above, but in step 2 the variances for each sample were drawn from a uniform distribution in the range [0.05, 1]. The same uniform distribution was used to generate prior variances by calculating the total sum-squared differences of  $m = v_0 / (n - 1)$  different draws (corresponding to  $m$  different genes) for the variance and dividing by the prior degrees of freedom,  $v_0$ . The results for both the PE and MP methods are shown in Table 3, where the observed false positive rate for the MP method is used to calibrate the  $\alpha$  value for the PE test so as to achieve nearly the same false positive rate for each method. Though both methods are

**Table 3: Comparison of  $\alpha$  and statistical power for iterated false positive rate and variable variance. The values are the same as those given in Table 1 and Table 2 except the variance for each sample was drawn from a uniform distribution between 0.05 and 1 during the simulation. The significance level ( $\alpha$ ) for the PE method was iterated so that the observed false positive rate matched that of the MP method where  $\alpha$  was held constant.**

| $n_1 = n_2 = n$ | $v_0$ | $\alpha_{PE}$ | $\alpha_{MP}$ | $FP_{PE}$ | $FP_{MP}$ | $Power_{PE}$ | $Power_{MP}$ |
|-----------------|-------|---------------|---------------|-----------|-----------|--------------|--------------|
| 2               | 0     | -             | 0.05          | -         | 0.0554    | -            | 0.4038       |
| 2               | 2     | 0.190         | 0.05          | 0.0593    | 0.0592    | 0.5876       | 0.5756       |
| 2               | 6     | 0.090         | 0.05          | 0.0522    | 0.0521    | 0.6838       | 0.6859       |
| 2               | 16    | 0.067         | 0.05          | 0.0573    | 0.0571    | 0.7563       | 0.7566       |

seen to have nearly the same power for a given false positive rate, the PE method again shows large deviations between  $\alpha$  and the observed false positive rate. However, the MP method shows more consistency between the observed and theoretical rates and is fairly robust to violations of the constant variance assumption.

To test whether these statistical improvements in the method translate to improvements in experimental data analysis, both methods were evaluated with the same microarray data [26] used by Baldi and Long [27]. The Arfin data matrix is composed of 8 columns by 1973 rows that contain gene expression levels used to study IHF regulated genes in *Escherichia coli*. Columns 1 to 4 correspond to 4 IHF<sup>+</sup> strains while columns 5 to 8 correspond to 4 IHF<sup>-</sup> strains. The 1973 rows correspond to the measured gene expression levels for each strain. There are actually more rows in the original data set but to be consistent with the earlier work, we only chose rows that had measurable expression levels for all 8 runs. As was done in the Long paper the 4 control and 4 treatment replicates were partitioned into 12 subsets that differ by at least two replicates so that quasi-independent pairwise comparisons could be made (i.e. 12v56, 12v78, 13v57, 13v68, 14v58, 14v67, 23v58, 23v67, 24v57, 24v68, 34v56, and 34v78). For the experimental evaluation we adopted the same strategy for pooling variances used by Baldi and Long. We sorted all the genes within each sample by the mean expression level and computed the prior variance by considering a window of the  $m/2$  most similarly expressed genes around each gene of interest. For the two samples this corresponds to a total of  $m$  similarly expressed genes and Eq. (18) gives the total prior degrees of freedom,  $v_0 = m$ .

To measure the consistency of a given testing method the number of common genes discovered between two quasi-independent subsets can be recorded. Genes that are discovered to be in common between subsets are more likely to be truly differentially expressed, thus higher a commonality percentage (genes discovered in common divided by the total number of genes discovered) is an indication that the testing method is more reliable. The number of common genes discovered can be recorded for the 66 pos-

sible ways of comparing the 12 subsets. Based on  $p$ -value rankings, the MP method achieved similar results (within sampling error) to those published using the Cyber-T software. The single hyperparameter ( $v_0$ ) was optimized for the MP method and a 95% confidence interval of [62.4, 67.8] common genes were identified out of the top 120. This compares well with the published results using Cyber-T where optimizing two parameters independently ( $v_0$  and  $m$ ) achieved an average common gene count of 67. These results represent a substantial improvement over the 33 common genes identified using a zero parameter classical  $t$ -test.

Because the  $p$ -value rankings by themselves are not actually statistical tests, a more stringent approach to assessing the performance of each method was accomplished by identifying genes that are differentially expressed at some confidence level  $\alpha$ . For the following tests we used a  $p$ -value of 0.01 for the null hypothesis cutoff (uncorrected for multiple tests). The results are presented in Table 4. In general, for a given number of prior degrees of freedom, the MP method consistently detects more genes as differentially expressed than does the PE method and more of these genes are commonly discovered when comparing them with quasi-independent subsets. In addition the fraction of common genes within those detected is generally higher for the MP method. The difference between the methods tends to decrease as the number of prior degrees of freedom is increased. The optimal performance (in terms of % commonality) is achieved at around  $v_0 = 400$ , which corresponds to basing the local variance on about 10% of the 1972 similarly expressed genes within each sample (there are 3944 similarly expressed genes when considering both samples). As  $v_0$  increases past 400 the commonality percentage drops somewhat. Basing the variance on the combined estimated of all 3944 similar genes results in a difference-in-means test, and the drop in performance is similar to that seen by Long and Baldi when a simple difference in means was used to rank the expression differences [27].

It is important to stress that the above results for the PE method are not what one would expect using the implementation found in the Cyber-T software [4]. As men-

**Table 4: Experimental comparison of hypothesis testing methods. The microarray data consists of 1973 genes partitioned into 12 quasi independent  $n_1 = 2$  and  $n_2 = 2$  subsets using replicate experiments of 4 control and 4 treatment runs [26, 27]. The number of genes detected as differentially expressed using each method is listed for an  $\alpha = 0.01$  significance level (there was no correction for multiple testing) and different prior degrees of freedom ( $v_0$ ). The average number of common genes for the 66 unique comparisons of the 12 subsets is also listed along with the % of commonality (common/detected) for each test.**

| $v_0$ | Detected <sub>PE</sub> | Detected <sub>MP</sub> | Common <sub>PE</sub> | Common <sub>MP</sub> | % Common <sub>PE</sub> | % Common <sub>MP</sub> |
|-------|------------------------|------------------------|----------------------|----------------------|------------------------|------------------------|
| 0     | 0                      | 46.2                   | 0                    | 6.4                  | 0                      | 13.8                   |
| 2     | 3.3                    | 72.5                   | 0.2                  | 21.0                 | 6.3                    | 29.0                   |
| 4     | 29.7                   | 88.0                   | 8.7                  | 30.9                 | 29.2                   | 35.1                   |
| 8     | 61.5                   | 95.2                   | 23.4                 | 37.9                 | 38.1                   | 39.8                   |
| 16    | 83.3                   | 103.9                  | 34.0                 | 42.6                 | 40.1                   | 40.1                   |
| 400   | 107.0                  | 108.1                  | 51.2                 | 51.5                 | 47.9                   | 47.6                   |
| 1972  | 118.8                  | 119.0                  | 54.7                 | 54.7                 | 46.0                   | 46.0                   |
| 3944  | 109.1                  | 109.2                  | 48.6                 | 48.7                 | 44.6                   | 44.6                   |

tioned before, this is because the existing software incorrectly calculates the pooled variance by averaging the standard deviation of similarly expressed genes. This results in a substantial error for example, when  $n_1 = n_2 = 2$  and  $v_0 = 400$ , the Type I error rate is approximately 0.125 (2.5 times higher than that predicted by the nominal rate  $\alpha = 0.05$ ). The authors of Cyber-T recommend smaller values of  $v_0$  where the overestimate of variance given by their point estimate, Eq. (10), is balanced by the underestimate of variance owing to incorrect pooling, thereby achieving (approximately) correct false positive rates by offsetting two competing biases.

**Discussion and conclusion**

The new marginal posterior (MP) formulation proposed here is uniformly at least as powerful as the previous point estimate (PE) method which it is based on and more powerful when the number of similarly expressed genes (and therefore prior degrees of freedom  $v_0$ ) is small. Moreover, unlike the existing regularization methods, the new formulation consistently maintains the proper false positive rates under the null hypothesis across the entire range of pooling. The MP formulation also makes clear the dependence between the prior degrees of freedom and prior variance and thus methods to estimate the single hyperparameter  $v_0$  are expected to be more robust and interpretable.

As mentioned previously, the key to the power of these Bayesian methods is the use of similarly-expressed genes to estimate the expected variance. Naturally, the more similar the variance in expression of the other genes is to the gene of interest, the greater the statistical consistency of the method. In generating the simulated data we have attempted to remain agnostic on the definition of "similarly-expressed", using variances that are either constant or drawn from a uniform distribution. Another Bayesian method [28] utilizes a clustering technique based on a series of *t*-statistics to group genes into prior categories in

order to pool their prior variances. Because the Bayesian posterior probabilities used by that method are not interpretable as frequentist Type I error rates, it is difficult to set the desired false positive rate using such a method without resorting to simulation and iteration. Other clustering strategies include those based on non-parametric regression to obtain local variance estimates [29] and mixture models that group together similar variances [30] using maximum likelihood. These clustering strategies assume the variance is known based on the large number of observations in each group. This allows normal distributions to be used in deriving test statistics. Unfortunately, when the number of similarly expressed genes is small the use of a normal distribution is not justified as it assumes there is no uncertainty in the variance. The advantage of the MP formulation is that the normal approximation for the test statistic is not required – it is statistically robust across the entire range of pooling or clustering strategies (assuming the strategies work properly to group together genes of similar variance). In our experimental application of the method we have adopted a pooling strategy based on grouping together genes with similar expression levels. Thus the MP formulation can be used in situations where two or more control replicates can be compared with only single observations under treatment conditions; this is not possible with clustering strategies that require at least two replicates in each sample [30].

Though the MP test was developed in a Bayesian framework, there is a strong analogy with frequentist statistics. One person's prior is another person's data; if we treat the prior estimates of variance as just additional data collection then our hypothesis test using the *t* distribution is the same one as that used in any introductory statistics course. In either situation, one considers the estimate of variance as the total sum-of-squared differences for all observations of the variance divided by the total degrees of freedom. By "all observations" we mean both the two samples involved in the hypothesis test plus some number of rep-

licates from other genes we believe to be similarly expressed and are good representatives of the variance. This is just like estimating the method standard deviation and conducting a normal  $t$ -test with that estimate along with the additional degrees of freedom that come with it. All of the power analysis calculations can then be done without recourse to simulations; for example, built-in functions in the desktop statistical package R (using the non-central  $t$  distribution) gave the same results we present by simulation for the MP method. It should be stressed that this Bayesian hypothesis test is not limited in any way to the analysis of microarray data. It can be applied to any experimental situation where prior beliefs about the variance associated with small sample sizes are used to achieve higher statistical power while maintaining consistent Type I error rates. This is particularly important when only a few similar observations of the variance are available, for example when only a handful of genes are under study across a range of treatment conditions.

The method is implemented in R and is freely available for download [31] or by contacting the corresponding author.

**Authors' contributions**

RJF conceived of and derived the statistical formulation, applied the method to various data sets and drafted the manuscript. MWD contributed major revisions to the manuscript and suggested ways of improving the study. Both authors contributed to the final version of the manuscript and approved it.

**Appendix**

**Determining the marginal posterior distribution of  $\Delta\mu$**

The posterior is the product of the prior and the likelihood:

$$p(\mu, \Delta\mu, \sigma^2 | \gamma) \propto p(\mu, \Delta\mu, \sigma^2) p(\gamma | \mu, \Delta\mu, \sigma^2) \quad (20)$$

The likelihood consists of the product of two normal distributions:

$$y_i \sim N(\mu, \sigma^2) \quad (21)$$

$$y_j \sim N(\mu + \Delta\mu, \sigma^2) \quad (22)$$

$$p(\mu, \Delta\mu, \sigma^2 | \gamma) \propto p(\mu, \Delta\mu, \sigma^2) \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \prod_{j=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_j - (\mu + \Delta\mu))^2\right) \quad (23)$$

Combining the joint likelihood into a common exponential gives:

$$p(\mu, \Delta\mu, \sigma^2 | \gamma) \propto p(\mu, \Delta\mu, \sigma^2) \sigma^{-(n_1+n_2)} \exp\left(-\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n_1} (y_i - \mu)^2 + \sum_{j=1}^{n_2} (y_j - (\mu + \Delta\mu))^2 \right] \right) \quad (24)$$

Next we can define a variable that only contains terms for the expression levels:

$$C_\mu = \sum_{i=1}^{n_1} (y_i - \mu)^2 + \sum_{j=1}^{n_2} (y_j - (\mu + \Delta\mu))^2 \quad (25)$$

$$p(\mu, \Delta\mu, \sigma^2 | \gamma) \propto p(\mu, \Delta\mu, \sigma^2) \sigma^{-(n_1+n_2)} \exp\left(-\frac{C_\mu}{2\sigma^2}\right) \quad (26)$$

Expanding the square leads to the following:

$$C_\mu = \sum_{i=1}^{n_1} (y_i^2 - 2y_i\mu + \mu^2) + \sum_{j=1}^{n_2} (y_j^2 - 2y_j\mu - 2y_j\Delta\mu + \mu^2 + \Delta\mu^2 + 2\mu\Delta\mu) \quad (27)$$

$$\bar{y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} y_i \quad (28)$$

$$\bar{y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} y_j \quad (29)$$

$$C_\mu = (n_1 + n_2)\mu^2 - 2(n_1\bar{y}_1 + n_2\bar{y}_2 - n_2\Delta\mu)\mu + \sum_{i=1}^{n_1} y_i^2 + \sum_{j=1}^{n_2} y_j^2 + n_2\Delta\mu^2 - 2n_2\bar{y}_2\Delta\mu \quad (30)$$

Next, we look to complete the square by adding and subtracting the same value  $k$ :

$$k = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 - n_2\Delta\mu}{n_1 + n_2} \quad (31)$$

$$C_\mu = (n_1 + n_2)(\mu^2 - 2k\mu + k^2 - k^2) + \sum_{i=1}^{n_1} y_i^2 + \sum_{j=1}^{n_2} y_j^2 + n_2\Delta\mu^2 - 2n_2\bar{y}_2\Delta\mu \quad (32)$$

Now collapsing back to a single squared term in the exponent and defining a new variable,  $C_\sigma$ , we have

$$C_\mu = (n_1 + n_2)(\mu - k)^2 + C_\sigma \quad (33)$$

where

$$C_\sigma = \sum_{i=1}^{n_1} y_i^2 + \sum_{j=1}^{n_2} y_j^2 + n_2\Delta\mu^2 - 2n_2\bar{y}_2\Delta\mu - (n_1 + n_2)k^2 \quad (34)$$

The posterior density can then be written as

$$p(\mu, \Delta\mu, \sigma^2 | \gamma) \propto p(\mu, \Delta\mu, \sigma^2) \sigma^{-(n_1+n_2)} \exp\left(-\frac{(n_1 + n_2)(\mu - k)^2}{2\sigma^2}\right) \exp\left(-\frac{C_\sigma}{2\sigma^2}\right) \quad (35)$$

We can now look to integrate out the true expression level,  $\mu$ :

$$p(\Delta\mu, \sigma^2 | \gamma) \propto \int_{-\infty}^{+\infty} p(\mu, \Delta\mu, \sigma^2 | \gamma) d\mu \quad (36)$$

We have flat priors on  $\mu$  and  $\Delta\mu$ , allowing us to complete the integration over all  $\mu$ .

$$p(\mu, \Delta\mu, \sigma^2) \propto p(\sigma^2) \quad (37)$$

$$p(\Delta\mu, \sigma^2 | \gamma) \propto p(\sigma^2) \sigma^{-(n_1+n_2)} \exp\left(-\frac{C_\sigma}{2\sigma^2}\right) \int_{-\infty}^{+\infty} \exp\left(-\frac{(n_1+n_2)(\mu-k)^2}{2\sigma^2}\right) d\mu \quad (38)$$

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{(n_1+n_2)(\mu-k)^2}{2\sigma^2}\right) d\mu \propto \sigma \quad (39)$$

The posterior for the joint density of  $\Delta\mu$  and  $\sigma^2$  is then

$$p(\Delta\mu, \sigma^2 | \gamma) \propto p(\sigma^2) \sigma^{-(n_1+n_2-1)} \exp\left(-\frac{C_\sigma}{2\sigma^2}\right) \quad (40)$$

We next look to integrate out  $\sigma^2$  by defining the sufficient statistics associated with the variance:

$$(n_1-1)s_1^2 = \sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2 = \sum_{i=1}^{n_1} y_i^2 - 2n_1\bar{y}_1^2 + n_1\bar{y}_1^2 = \sum_{i=1}^{n_1} y_i^2 - n_1\bar{y}_1^2 \quad (41)$$

$$(n_2-1)s_2^2 = \sum_{j=1}^{n_2} y_j^2 - n_2\bar{y}_2^2 \quad (42)$$

$$C_\sigma = (n_1-1)s_1^2 + (n_2-1)s_2^2 + C_{\Delta\mu} \quad (43)$$

where

$$C_{\Delta\mu} = n_1\bar{y}_1^2 + n_2\bar{y}_2^2 + n_2\Delta\mu^2 - 2n_2\bar{y}_2\Delta\mu - (n_1+n_2)k^2 \quad (44)$$

The last term in Eq. (43)  $C_{\Delta\mu}$ , now contains all references to the null difference in means,  $\Delta\mu$ . The term can be expanded by multiplying out  $k^2$  and then simplified:

$$\begin{aligned} C_{\Delta\mu} &= n_1\bar{y}_1^2 + n_2\bar{y}_2^2 + n_2\Delta\mu^2 - 2n_2\bar{y}_2\Delta\mu - \\ &\frac{1}{n_1+n_2} \left( n_1^2\bar{y}_1^2 + 2n_1n_2\bar{y}_1\bar{y}_2 + n_2^2\bar{y}_2^2 - 2n_1n_2\bar{y}_1\Delta\mu + n_2^2\Delta\mu^2 - 2n_2^2\bar{y}_2\Delta\mu \right) \quad (45) \\ &= \frac{n_1n_2}{n_1+n_2} (\Delta\mu - (\bar{y}_2 - \bar{y}_1))^2 \end{aligned}$$

This can be inserted back into Eq. (43) to give

$$C_\sigma = (n_1-1)s_1^2 + (n_2-1)s_2^2 + \frac{n_1n_2}{n_1+n_2} (\Delta\mu - (\bar{y}_2 - \bar{y}_1))^2 \quad (46)$$

The prior for variance is said to follow a scaled inverse Chi-square distribution:

$$p(\sigma^2) = I(\sigma^2; \nu_0, \sigma_0^2) \propto \sigma^{-(\nu_0+2)} \exp\left(-\frac{1}{2\sigma^2} \nu_0 \sigma_0^2\right) \quad (47)$$

The posterior density is then

$$\sigma^{-(n_1+n_2+\nu_0+1)} \exp\left(-\frac{1}{2\sigma^2} \left[ \nu_0 \sigma_0^2 + (n_1-1)s_1^2 + (n_2-1)s_2^2 + \frac{n_1n_2}{n_1+n_2} (\Delta\mu - (\bar{y}_2 - \bar{y}_1))^2 \right] \right) \quad (48)$$

But now we can integrate out  $\sigma^2$

$$p(\Delta\mu | \gamma) \propto \int_0^\infty p(\Delta\mu, \sigma^2 | \gamma) d\sigma^2 \quad (49)$$

With a change of variables

$$A = \nu_0 \sigma_0^2 + (n_1-1)s_1^2 + (n_2-1)s_2^2 + \frac{n_1n_2}{n_1+n_2} (\Delta\mu - (\bar{y}_2 - \bar{y}_1))^2 \quad (50)$$

$$z = \frac{A}{2\sigma^2} \quad (51)$$

$$p(\Delta\mu | \gamma) \propto \int_0^\infty \left(\frac{A}{2z}\right)^{-(n_1+n_2+\nu_0+1)/2} \exp(-z) \left(\frac{A}{2z^2}\right) dz \quad (52)$$

$$p(\Delta\mu | \gamma) \propto A^{-(n_1+n_2+\nu_0-1)/2} \int_0^\infty z^{(n_1+n_2+\nu_0-3)/2} \exp(-z) dz \quad (53)$$

The posterior density then becomes the well known  $t$  distribution.

$$p(\Delta\mu | \gamma) \propto \left[ 1 + \frac{(\Delta\mu - (\bar{y}_2 - \bar{y}_1))^2}{\nu_n \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \sigma_n^2} \right]^{-(\nu_n+1)/2} \quad (54)$$

where

$$\nu_n = n_1 + n_2 + \nu_0 - 2 \quad (55)$$

$$\nu_n \sigma_n^2 = \nu_0 \sigma_0^2 + (n_1-1)s_1^2 + (n_2-1)s_2^2 \quad (56)$$

$$\frac{\Delta\mu - \Delta\bar{y}}{\sigma_n \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \Bigg| \gamma \sim t_{\nu_n} \quad (57)$$

### Acknowledgements

We would like to thank Magda Bartilson and Stephen DeCardayre for giving us the inspiration to develop the method. We also thank John Grate, Lori Giver and John Whitman for their generous support and feedback and for reviewing various versions of the manuscript.

## References

1. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *J Biomed Optics* 1997, **2(4)**:364-374.
2. Cui X, Churchill G: **Statistical tests for differential expression in cDNA microarray experiments.** *Genome Biology* 2003, **4(4)**:210.0-210.1.
3. Nadon R, Shoemaker J: **Statistical issues with microarrays: processing and analysis.** *Trends in Genetics* 2002, **18(5)**:265-271.
4. Baldi P, Long A: **A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes.** *Bioinformatics* 2001, **17(6)**:509-519.
5. Townsend J: **Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays.** *BMC Genomics* 2003, **4(41)**.
6. Vinciotti V, Khanin R, xD'Alimonte R, Liu X, Cattini N, Hotchkiss G, G. B, de Jesus O, Rasaiyaah J, Smith CP, Kellam P, Wit E: **An experimental evaluation of a loop versus a reference design for two-channel microarrays.** *Bioinformatics* 2005, **21(4)**:492-501.
7. Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3(8)**:579-588.
8. Orian A, van Steensel B, Delrow J, Bussemaker HJ, Li L, Sawado T, Williams E, Loo LW, Cowley SM, Yost C, Pierce S, B.A. E, Parkhurst SM, Eisenman RN: **Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network.** *Genes Dev* 2003, **17(9)**:1101-1114.
9. Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, Brivanlou AH: **Molecular signature of human embryonic stem cells and its comparison with the mouse.** *Dev Biol* 2003, **260(2)**:404-413.
10. Tompa R, McCallum CM, Delrow J, Henikoff JG, van Steensel B, Henikoff S: **Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3.** *Curr Biol* 2002, **12(1)**:65-68.
11. Hu L, Wang J, Baggerly K, Wang H, Fuller GN, Hamilton SR, Coombes KR, Zhang W: **Obtaining reliable information from minute amounts of RNA using cDNA microarrays.** *BMC Genomics* 2002, **3(16)**.
12. Gu J, Gu X: **Induced gene expression in human brain after the split from chimpanzee.** *Trends Genet* 2003, **19(2)**:63-65.
13. Pavlidis P, Li. Q, Noble WS: **The effect of replication on gene expression microarray experiments.** *Bioinformatics* 2003, **19(13)**:1620-1627.
14. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A: **False discovery rate, sensitivity and sample size for microarray studies.** *Bioinformatics* 2005, **21(13)**:3017-3024.
15. Thomas JG, Olson JM, Tapscott SJ, Zhao LP: **An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles.** *Genome Res* 2001, **11(7)**:1227-1236.
16. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comp Biol* 2001, **8(1)**:37-52.
17. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2000, **98(9)**:5116-5121.
18. Theilhaber J, Bushnell S, Jackson A, Fuchs R: **Bayesian estimation of fold-changes in the analysis of gene expression: the PFOLD algorithm.** *J Comp Biol* 2001, **8(6)**:585-614.
19. Townsend J: **Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays.** *BMC Bioinformatics* 2004, **5(54)**.
20. Townsend JP, Hartl DL: **Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments.** *Genome Biol* 2002, **3(12)**.
21. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.** *Nucleic Acids Res* 2001, **29(12)**:2549-2557.
22. Reiner A, Yekutieli D, Benjamini Y: **Identifying differentially expressed genes using false discovery rate controlling procedures.** *Bioinformatics* 2003, **19(3)**:368-375.
23. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18(Suppl. 1)**:S96-S104.
24. Durbin BP, Hardin JS, Hawkins DM, Rocke DM: **A variance-stabilizing transformation for gene-expression microarray data.** *Bioinformatics* 2002, **18(Suppl. 1)**:S105-S110.
25. DeGroot MH, Schervish MJ: **Probability and Statistics.** Addison Wesley; 2002.
26. Arfin S, Long D, Ito E, Toller L, Riehle M, Paegle E, Hatfield GW: **Global gene expression profiling in Escherichia coli K12.** *J Biol Chem* 2000, **275(38)**:29672-29684.
27. Long D, Mangalam H, Chan B, Toller L, Hatfield GW, Baldi P: **Improved statistical inference from DNA microarray data using analysis of variance and a Bayesian statistical framework.** *J Biol Chem* 2001, **276(23)**:19937-19944.
28. Gottardo R, Pannucci J, Kuske C, Brettin T: **Statistical analysis of microarray data: a Bayesian approach.** *Biostatistics* 2003, **4(4)**:597-620.
29. Jain N, Thatte J, Braciale T, Ley K, O'Connell M, Lee J: **Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays.** *Bioinformatics* 2003, **19(15)**:1945-1951.
30. Delmar P, Robin S, Daudin JJ: **VarMixt: efficient variance modeling for the differential analysis of replicated gene expression data.** *Bioinformatics* 2005, **21(4)**:502-508.
31. Fox RJ, Dimmic MW: **A Bayesian two-sample t-test.** 2006 [<http://www.dimmic.net/supplement/>].

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

