A Bayesian calibration model for combining different pre-processing methods in Affymetrix chips

Marta Blangiardo*[1] and Sylvia Richardson[1]

[1]Centre for Biostatistics, Imperial College. St. Mary's Campus, Norfolk Place London W2 1PG, UK

Email: Marta Blangiardo*- m.blangiardo@imperial.ac.uk;

*Corresponding author

Abstract

In gene expression studies a key role is played by the so called "pre-processing", a series of steps designed to extract the signal and account for the sources of variability due to the technology used rather than to biological differences between the RNA samples. Many studies have shown how this choice can affect the results of subsequent analysis carried out to measure the influence of biological contrasts in differential expression. At the moment there is no commonly agreed gold standard method and each researcher has the responsibility to choose one pre-processing method, incurring the risk of false positive and false negative features associated with the pre-processing method chosen. We propose a Bayesian model that combines several pre-processing methods to assess the "true" unknown differential expression between two conditions and show how to estimate the posterior distribution of the differential expression values of interest. The model is tested both on simulated data and on a spike in data set and its biological interest is demonstrated through a real example on publicly available data.

Introduction

In gene expression studies one of the first steps of the statistical analysis is to estimate and correct the systematic noise through pre-processing, a series of steps designed to extract the signal and the sources of variability due to the technology used rather than to biological differences between the RNA samples. Many studies in the literature present the importance of pre-processing and show how this can influence the results in terms of differential expression (see, for example, [1] and [2]).

However, an agreed gold standard does not exist and as Allison and colleagues [3] discuss in a recent paper, researchers are torn between the different methods and usually end up restricting their analysis to using only one method (often the most commonly used or the most user-friendly). A simple alternative strategy is to perform the analysis using two different pre-processing methods and then compare the results in terms of differential expression, focusing attention on the genes in the intersection. The former strategy is reductive while the latter relies on the arbitrary choice of two methods and on that of considering only their intersection. Neither of these approaches make optimal use of all the information provided by the pre-processing methods.

In this work we aim at providing a more efficient analysis of a gene-expression experiment analyzed with several pre-processing to obtain a pooled estimate of differential expression. In the context of microarray, many researchers have proposed different strategies to pool together several independent studies. In particular, they have focused attention on the integration of each gene effect across studies [4] or on the evaluation of the consistency of differential expression across platforms [5]. Conlon *et al.* [6], [7] have proposed a different approach: they do not estimate a combined gene effect across studies, but they evaluate its differential expression through a pooled binary indicator for independent studies performed on the same platform. Their model is formulated in a Bayesian perspective and the posterior probability of differential expression is the quantity of interest. Later, Scharpf *et al.* [8] have extended Conlon's work to include the comparison of several platforms and to focus attention also on genes discordant across the experiments through the estimating of the sign of differential expression for each experiment. These papers share a *meta-analytical* framework, devoting their interest to synthesize several independent studies. We depart from this set up as we concentrate on a single experiment analyzed with several pre-processing techniques; thus we adopt a *measurement error* perspective assuming for each gene a latent (unmeasured) "true" differential expression and for each pre-processing method: (i) a measured value that departs from it (relative bias), (ii) a variance component.

Modeling measurement error is common practice in epidemiology, where errors in the recording of

explanatory variables is a frequent problem that has to be taken into account during the statistical analysis (see for example [9], [10], [11]); the formulation in a Bayesian framework has been discussed in the early 1990s by Thomas *et al.* [12], Richardson and Gilks [13], [14] and Richardson [15] among others, placing particular emphasis on the way their approach propagates coherently all sources of uncertainty in the data onto the estimation of the parameters of interest.

We follow Richardson and Gilks and specify a Bayesian calibration model for assessing differential expression in Affymetrix microarray, which combines the information from several pre-processing techniques. The freely available software WinBUGS [16] can be used to estimate the posterior distribution of the differential expression values of interest. The model performance is tested using simulated data and the Latin square data set provided by Affymetrix [17]; we also discuss a real biological example using an experiment publicly available to evaluate the effect of High Fat Diet versus Normal Fat Diet in mice adipose tissue.

Results

Bayesian Model

Our combined model for different pre-processing methods is characterized by two measurement error components: (i) an estimate of the relative bias for the "true" differential expression and (ii) a measure of variability around the mean gene expression.

Following [18], the observed log expression value for gene $g = 1, \dots G$, pre-processing $j = 1, \dots, J$, condition $k = 1, 2$ and replicate $r = 1, \dots, R$ is modeled as a Normal distribution:

$$y_{gj1r} \sim N(\alpha_{gj} - \frac{1}{2} \times \delta_g \times \phi_j, \sigma^2_{gj1}) \qquad y_{gj2r} \sim N(\alpha_{gj} + \frac{1}{2} \times \delta_g \times \phi_j, \sigma^2_{gj2}) \tag{1}$$

The parameter α_{gj} represents the global gene expression that is specific to the gene g and the pre-processing method j, whereas δ_g is the "true" (unknown) differential expression that we would like to capture for gene g. The method specific coefficient ϕ_j quantifies the relative bias of the method with respect to the latent quantity δ_g.

In (1) the variance σ^2_{gjk} for each gene, pre-processing and condition is the result of a gene and condition specific component σ^2_{gk} and an exponential error term specific to the pre-processing method and to the condition:

$$\sigma^2_{gjk} = \exp(\lambda_{1jk} + \lambda_{2jk} \times \overline{y}_g + \lambda_{3jk} \times \overline{y}^2_g) \times \sigma^2_{gk}. \tag{2}$$

The exponential component is allowed to depend on the global expression of the gene $\left(\overline{y}_g = \frac{1}{2JR}\sum_{j,k,r} y_{gjkr}\right)$ as it has often been noted that even after log transformation, the variability of the expression of a gene can be affected by its level of expression (see for example [19]). The use of a second order polynomial offers considerable flexibility though involving a limited number of parameters; a simplified version of equation (2) can be assumed when the coefficients λ_{2jk} and λ_{3jk} are equal to 0, so the exponential component is independent from the expression of the gene.

Following [18] we assign a hierarchical structure on the gene and condition specific component, to borrow strength from the entire set of genes:

$$\sigma^2_{gk} \quad \sim \quad Ga^{-1}(a_k, b_k) \tag{3}$$

To complete the model we need to specify the prior distributions for all the remaining parameters. The coefficients for the exponential component in (2) are assumed independent and to follow Normal non informative distributions $\lambda_{1jk} \sim N(0, 10^5)$, $\lambda_{2jk} \sim N(0, 10^5)$, $\lambda_{3jk} \sim N(0, 10^5)$, so that the function can model every trend. As identifiability constraint, we impose that the sum of the experimental parameters in (2) over the three pre-processing methods is equal to 0 ($\sum_j \lambda_{1jk} = 0$, $\sum_j \lambda_{2jk} = 0$, $\sum_j \lambda_{3jk} = 0$). We use the exponential parametrization to ensure positivity of (2).

The relative bias coefficients are modeled as $\phi_j \sim logN(0, 0.0001)$ independently for $j = 1, \ldots, J$, imposing the identifiability constraint that $\prod \phi_j = 1$, while we specify a non informative Normal distribution on α_{gj} and δ_g. Finally, $a_k \sim Ga(0.01, 0.01)$ and we model $1/\sqrt{b_k} \sim U[0, min_g(s^2_{gk})^{-0.5}]$, where s^2_{gk} is the sample variance for gene g and condition k; this choice ensures that the posterior distributions of a_k and b_k are proper and well adapted to the scale of the data, as justified in [20].

Models (1) and (2) allow the borrowing of information across genes to estimate ϕ_j and λ_{jk}, and across methods to estimate δ_g and σ^2_{gk}. The hierarchical model specified by (1), (2), (3) and the prior distributions are estimated using an MCMC algorithm coded in WinBUGS [16] to simulate the prior/posterior distribution of all unknown parameters. More details can be found in section *Materials and Methods* and in File 1 of Supplemental Material.

Tail posterior probability

For each gene we are interested in testing the hypothesis that the differential expression effect δ_g is different from 0:

$$H_0^g : \delta_g = 0 \qquad \text{vs} \qquad H_1^g : \delta_g \neq 0$$

and a variety of decision rules based on the output of the hierarchical model can be constructed. We chose to use the tail posterior probability statistic introduced in [20] to measure the strength of the evidence against H_0. This method considers a standardization of the differential expression measure $z_g = \frac{\delta_g}{\sqrt{(w_g)}}$ where w_g is a pooled measure of variability of δ_g:

$$w_g \quad = \quad \frac{2}{RJ^2} \sum_{j=1}^{J} (\sigma_{gj1}^2 + \sigma_{gj2}^2).$$

Here R is the number of replicates for each condition and J is the number of pre-processing methods considered. The tail posterior probability statistic is then defined as follows:

$$p(z_g; z_\alpha) \quad = \quad P(|z_g| > z_\alpha \mid \mathbf{y_g}) \tag{4}$$

where $\mathbf{y_g}$ denotes all the data available for gene g and z_α is the α quantile of the standard normal distribution (usually $\alpha = 0.05$ and consequently $z_\alpha = 1.96$). As discussed in [20] the histogram of $p(z_g; 1.96)$ is characterized by a local peak on the right tail in the presence of differentially expressed genes; this peak can be used to define a reasonable cut off for the differential expression (see the section that describes the results on real data for an illustrative example).

The tail posterior probability statistic can be loosely intended as a Bayesian analogy of the t test. It makes full use of the Bayesian output being a function of the differential expression (δ_g) and of the variability (σ_{gj1}^2 and σ_{gj2}^2), is easy to use and was shown to have good properties (see [20] for more details).

Posterior Predictive Check

One of the added benefits of working in a Bayesian framework is the ability to perform model checks by means of the predictive distribution of the parameters of interest. We use Mixed Posterior Predictive checks ([21], [22], [23]) , applied on gene expression data by Lewin *et al.* ([18], [24]) and focus attention on checking the gene specific variance, characterized by a hierarchical structure as described in equations (2) and (3). For each method we compare the observed sample variance, calculated for the expression

values, and the variance of the predicted expression values of each gene under the model using an empirical p-value.

Under the null hypothesis of the model being true, the distribution of the p-values should be approximatively uniform, while a poor model fit is indicated by the presence of a notable pattern in the plot, suggesting a systematic difference between the observed values and the predicted ones (see File 1 of Supplemental Material and [18] for more details).

Performance on simulated data

To first evaluate the benefits of using a model that combines several pre-processing methods we simulated log expression values for 1000 genes, 5 pre-processing and 2 conditions, following the approach described in section *Materials and methods*. We set 200 genes as differentially expressed and the remaining 800 genes as not differentially expressed. For the sake of clarity, we considered a simplification of equation (2) assuming (i) the same variance for the two conditions ($\sigma^2_{gj1} \equiv \sigma^2_{gj2}$) and (ii) the exponential component as independent from the global gene expression ($\lambda_{2jk} = 0, \lambda_{3jk} = 0$). To evaluate the consistency of our results we repeated the simulation 10 times and averaged the results.

The typical behavior of the combined method compared to each pre-processing one is presented in Figure 1: the ROC curve averaged over the 10 runs shows a greater sensitivity and specificity for the model that combines the five pre-processing approaches.

Ranking the pre-processing methods based on the ROC curve, Method 3, characterized by large relative bias ($\phi_3 = 2$) and small variability ($\exp(\lambda_3) = 0.5$), shows the best performance, while on the other end, Method 4, characterized by small relative bias ($\phi_4 = 0.5$) and high variability ($\exp(\lambda_4) = 2$), shows the worst ROC curve. Note that the "reference" method, characterized by $\exp(\lambda_1) = 1$ and $\phi_1 = 1$ shows an average specificity and sensitivity.

When doing pairwise comparisons of methods with the same variability (Method 2 vs Method 3 and Method 4 vs Method 5) they differ in their performance due to the relative bias coefficient ϕ_j: if it is set larger than 1 (Method 3 and Method 5), the corresponding pre-processing is assumed to inflate the "true" differential expression. This results in a stronger signal and consequently in a greater ability to discern true positives and true negatives.

On the other hand, comparing methods characterized by the same value for the relative bias coefficient (Method 2 vs Method 4 and Method 3 vs Method 5) the difference in their performance is explained by the

exponential component of variability $\exp(\lambda_j)$: a low value (Method 2 and Method 3) results in a higher precision of the estimates enhancing the performance of the method in terms of specificity and sensitivity.

FIGURE 1 HERE

Table 1 presents the operating characteristics for all the pre-processing methods and for the combined strategy. As we set 200 genes as differentially expressed in the simulation scenario, we consider the first 200 genes ranked accordingly to the tail posterior probability (4) and evaluate the number of False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN). The results are averaged over 10 repeats. The combined method is able to detect the maximum number of *truly differentially expressed genes* (179) and is characterized by only the 2.6% of False Positives. As already pointed out, the methods with a $\phi_j > 1$ (Method 3 and Method 5) have a higher signal, leading to a better performance, by means of a small percentage of False Positives and False Negatives.

TABLE 1 HERE

Latin square data set

We applied the model presented in (1), (2) and (3) to the 10621 genes present in the Latin square data set (experimental condition 2 versus 1), where only 64 genes are *truly differentially expressed*. The p-values histograms for the two conditions obtained from the posterior predictive checks are presented in the upper plots of Figure 2 and are characterized by a uniform behavior, suggesting a good model fit to the data. As a point of comparison we ran the simplified model characterized by $\lambda_{2jk} = \lambda_{3jk} = 0$ and present the corresponding posterior predictive checks in the lower plots of Figure 2. These plots show a deviation from uniformity with an over representation of small and large p-values, in particular for condition 1, suggesting a lack of flexibility of the simple model to account for all the variability present in the data. Hence we retain the variance model in (2) for our analysis.

FIGURE 2 HERE

Table 2 presents a synthesis of the results for the combined model and for each single pre-processing method. We evaluated the operating characteristics of each method based on the first 64 genes ranked accordingly to their tail posterior probability. The combined method shows an improvement in sensitivity and specificity, even if modest: out of the first 64 ranked genes only 12 are false positives, while the number increases to 14 for RMA and then jumps to 23 and 31 for MAS5 and dChip respectively. When the gene list size increases, the performance for the different methods tends to converge.

TABLE 2 HERE

Table 3 reports the posterior mean of the relative bias effect ϕ_j together with the 95% credibility interval. We see that MAS5 and RMA are characterized by a mean value greater than 1, meaning that they inflate the estimate δ_g, while $\phi < 1$ for dChip.

TABLE 3 HERE

Biological example: High Fat Diet versus Normal Fat Diet in mice adipose tissue

We applied the model presented in (1), (2) and (3) to the 12488 genes in the experiment to study the effect of high fat diet (HFD) versus normal fat diet (NFD) on mice adipose tissue, as part of the DGAP project (`www.diabetesgenome.org/arraydata.cgi`). As for the Latin Square data set the Mixed Posterior Predictive checks show a good fit, while the simpler model with $\lambda_{2jk} = \lambda_{3jk} = 0$ is not adequate (see Figure 3).

FIGURE 3 HERE

We use the histogram of the tail posterior probability to identify a reasonable cut off for calling a gene differentially expressed. Contrary to what happens for each single pre-processing method, the histogram of the tail posterior probability for the combined model shows a local peak on the right tail of the distribution (see Figure 1 in File 1 of Supplemental Material), indicating more evidence of differential expression. As suggested in [20], we select a high cut off value, in our case equal to 0.98, corresponding to the local peak on the combined distribution, and obtained a list of 292 'top' genes classified as differentially expressed by the combined method. If we fixed the same cut off on the tail posterior probability for each pre-processing method, we would obtain substantially smaller lists with only 20 genes classified as differentially expressed by MAS5, 32 by RMA and 41 by dChip. This highlights the gain of confidence provided by the combined analysis: many more genes have high posterior probability of being differentially expressed in the combined model than if we proceeded for each method separately. In order to perform a comparison between the methods we also consider the first 292 genes ranked according to the tail posterior probability for each single method. Note that in doing so we lower the cut off on the tail posterior probability scale (0.78 for MAS5, 0.79 for RMA and 0.83 for dChip), introducing more noise in the list.

Figure 4 shows the Venn diagram for the three single methods and the combined model. All the 46 genes in the intersection of the three methods are included in the list of 'top' genes by the combined model. Additionally, there are 61 genes that are only found in the combined list, suggesting that combining several

pre-processing methods is more powerful than considering the intersection of lists for each method.

FIGURE 4 HERE

Figure 5 presents the plot of the log fold change versus the posterior probability (volcano plot) for the combined model and for each pre-processing method considered separately. The 292 genes called by the combined method are highlighted in red and are placed in the upper half of the plot for RMA and dChip, being characterized by values of the posterior probability far from 0. This indicates that in general when a gene shows some evidence of differential expression for the two methods with smaller variability, the combined approach strengthen this evidence and places these genes at the top of the list of differentially expressed genes. MAS5 is the method that contributes the least to the combined output, being associated with the largest variability. For this reason, some genes with a tail posterior probability < 0.5 for MAS5 can still be found at the top of the list for the combined method, if their posterior probability values for RMA and dChip are large enough. On the other hand, MAS5 shows the largest ϕ_j effect (presented on the right hand side of Table 3); thus, as already observed in the previous sections, the method is characterized by a larger signal that may provide additional information on the differential expression. This results in 7 genes called as differentially expressed only by MAS5 that are placed in the top list for the combined method. These genes would not end up in the list of differentially expressed genes when considering only RMA and dChip. We compare and discuss the results using combinations of two or three methods in the *Discussion* section.

FIGURE 5 HERE

Table 4 presents the number of annotated genes in the list of the first 292 ranked genes for each method. Looking specifically at the GO annotation at the level of Biological Processes, Molecular Functions and Cellular Components, the combined method seems to enrich it. The most represented biological processes are the metabolic ones, functions associated with the response of the body to a change in the diet. The number of genes involved in *cellular metabolism*, is 105 for the combined method, 70 for dChip, 87 for MAS5 and 88 for RMA. Similarly the number of genes involved in primary metabolism or macromolecular metabolism is larger for the combined model than for each pre-processing method.

KEGG pathways are also enriched through the combined strategy: these are mainly related to immune response and oxidation (*Antigen processing and presentation*, *MAPK signalling pathway*, *PPAR signaling pathway*), biologically regulators of physiological functions as energy metabolism, insulin action, immunity and inflammation and known from the literature to be associated with high fat diet (see [25] and [26]).

TABLE 4 HERE

Discussion

This paper describes a simple method to combine several pre-processing techniques on Affymetrix chips. It exploits the dependence between the results of the pre-processing methods and by using natural multiplicative assumptions on the structure of the measurement errors is able to (i) borrow information across the genes to estimate the method specific operational characteristics (λ_j, ϕ_j) and (ii) borrow information across the methods to estimate a gene specific component of variability (σ_g^2) and of differential expression (δ_g).

We want to stress the importance of having a way to assess whether the formulation of the synthetic model is an adequate representation of what is common and specific to the different methods. Bayesian model checks based on prediction allow the comparison of the observed data and the 'predicted' data under the model with respect to any feature of interest. We believe that the Mixed Posterior Predictive Checks that we considered are an effective way to carry out a model selection strategy. They have the advantage of calculating a measure of discrepancy (the empirical p-value) for each gene that can be easily displayed through a histogram. On the Latin square and the real data we used a *flexible and realistic* model, that allows a relation between the expression ($\overline{y}_g$) and the variance (σ_{gk}^2), instead of assuming no link. Comparing this formulation to a simpler version by means of the Posterior Predictive Checks, we clearly see a better fit for the more flexible one, which suggests that a variance model that is linked to the level of expression is better suited to the complexity of real data.

The contribution of MAS5 to the combined output has been investigated by implementing a version of our model that combines only RMA and dChip and comparing the results. In general, combining only these two methods results in 205 genes classified as differentially expressed for a cut off of 0.98, again chosen on the basis of the local peaks on the tail posterior probability histogram (see Figure 2 in File 1 of Supplemental Material). Out of these, 163 are in common with the model combining the 3 methods and 42 are specific to the two-methods calibration model. In the three-methods combined model, these 42 genes are either (i) borderline or (ii) far from the top of the list, characterized by high tail posterior probability for only one method. This shows that for some genes there remains a non negligible source of uncertainty when only two methods are considered. When a third method is included, evidence versus the null hypothesis of no differential expression is added and these genes are not classified as significant anymore. Moreover, including an additional method in the synthesis results in 129 new genes called differentially expressed: for these genes the evidence against H_0 is thus strengthened by MAS5. As a general comment we recommend to include several methods as this diminishes the impact of each pre-processing on the combined model.

Materials and Methods

Simulated data

To test the performance of our method we simulated log expression values for 1000 genes, two conditions and five pre-processing methods from the model previously described, and extracted 5 replicates for each combination of condition and pre-processing ($r = 1, \ldots, 5$). We specified 200 differentially expressed genes characterized by a log expression $y_{gjkr} \sim N(\alpha_{gj} + (-1)^k \frac{1}{2}\delta_g \times \phi_j, \sigma^2_{gj})$, with $k = 1, 2$, while for the remaining 800 genes the log expressions for both conditions came from the same distribution: $y_{gjkr} \sim N(\alpha_{gj}, \sigma^2_{gj})$. The parameters for simulating the distributions of $\alpha_{gj} \sim N(6.79, 4.77)$ and $\delta_g \sim N(0, 0.25)$ were obtained from real data we have analyzed (`www.bair.org.uk`); the model on the variance σ^2_{gjk} is presented in equation (2) with $(\lambda_{2jk} = 0, \lambda_{3jk} = 0)$ and we assume the same variance for the two conditions ($\sigma^2_{gj1} \equiv \sigma^2_{gj2}$), obtained from real data and characterized by the quartiles 0.02, 0.04, 0.09, 0.15.

The 5 pre-processing methods are described in the following table:

	Method 1	Method 2	Method 3	Method 4	Method 5
$exp(\lambda_j)$	1	0.5	0.5	2	2
ϕ_j	1	0.5	2	0.5	2

We assume the first method is the "reference", being characterized by $exp(\lambda_1) = 1$ and $\phi_1 = 1$. Method 2 and Method 4 have a smaller relative bias ($\phi_2 = \phi_4 = 0.5$), while that of Method 3 and Method 5 is larger ($\phi_3 = \phi_5 = 2$); Method 2 and Method 3 have a variability smaller than 1 ($exp(\lambda_2) = exp(\lambda_3) = 0.5$) while that of Method 4 and Method 5 is larger ($exp(\lambda_4) = exp(\lambda_5) = 2$).

We used an MCMC algorithm with two chains to estimate the parameters of interest (we checked convergence for 10000 iterations and then extracted a sample of 1000 iterations).

To evaluate the consistency of our results we repeated the simulation process 10 times and performed our Bayesian estimate for each run. The model estimates well the values of the parameters λ and ϕ (see Table 5 for their posterior mean and 95% credibility intervals; see Figure 3 and 4 of File 1 of Supplemental Material for the associated posterior density plots).

TABLE 5 HERE

As a point of comparison we also ran the model separately for each pre-processing method and compared the performance of both combined and single pre-processing methods in terms of sensitivity and specificity (see the results on simulated data).

Spike-in example: Latin Square Affymetrix data

We tested our method by means of the Affymetrix Latin Square data set [17]. The array used is human $Hgu133a$ and contains 22300 genes. There are 14 experimental conditions and each has 3 replicates. We considered the experimental condition 2 versus 1 and the 42 spike-in probesets indicated by Affymetrix plus the 22 new spike in genes proposed by McGee *et al.* [27]. The ratio of the concentrations is 1:2 for 60 spike in genes and 0 for the remaining 4. Among the remaining 22236 genes we extracted only the ones present in at least one condition, evaluated using the present/absent call included in the Affy R package (`www.r-project.org`) and obtained 10621 genes.

We focused attention on the three most used pre-processing methods : MAS5 [28], RMA [29] and dChip [30]. There are many versions of each of them, but we considered the default ones provided by the software R. The differences in the three methods are described in Table 6.

TABLE 6 HERE

We ran the combined model, but also treated separately each pre-processing method. Again we performed the MCMC estimation with two chains (we checked convergence for 10000 iterations and then extracted a sample of 1000 iterations).

Biological example: High Fat Diet versus Normal Fat Diet in mice adipose tissue

There are many studies in the literature describing the effect of high fat diet on gene expression of several tissues in mice (see for example [31] and [32] for adipose tissue and [33] for liver). They are particularly interesting since the effect of diet can trigger obesity, hypertension and be related to major pathologies as diabetes.

In order to assess if our model leads to a more powerful analysis that improves the biological interpretability of the results we ran it on a publicly available experiment to study the effect of high fat diet (HFD) versus normal fat diet (NFD) on mice adipose tissue. The array used is mouse MG_U74Av2 and contains 12488 genes. The experiment analyzes the strain 129 and for each condition there are 4 replicates. The .CEL files and the description of the experiments are available at the DGAP project website [34]. Again we pre-processed the data using MAS5, RMA and dChip and ran the combined model, but also treated separately each pre-processing method. The MCMC estimation was performed with two chains (we checked convergence for 10000 iterations and then extracted a sample of 1000 iterations).

Implementation

The standard model built by equations (1), (2), (3) and by the prior distributions has been implemented in the free software WinBUGS and the code is provided in File 1 of Supplemental Material. Note that it is relatively quick to run with a small number of genes (it takes around 5 minutes to perform 1000 iterations on a DELL Precision workstation with 3.20 GHz for 1000 genes, 2 conditions, 3 pre-processing and 5 replicates), but the time increases linearly with the number of genes. To increase the speed on the Latin Square data set (that includes 22300 genes), we filtered the present genes for at least one condition, using the present/absent call implemented in the Affy R package; nevertheless it is possible to apply more stringent criteria (e.g. selecting the most variable genes between the two conditions), as long as the non differentially expressed genes are well represented in the subset, otherwise the null distribution of δ_g is not properly identified and it affects the correct estimate of the tail posterior probability.

Acknowledgments

The authors would like to thank Alex Lewin and Natalia Bochkina for valuable statistical discussion, Xinzhong Li, Peter Thomason and James Scott for useful discussions that motivated the development of this work. Marta Blangiardo and Sylvia Richardson's work was supported by BBSRC 'Exploiting Genomics' grant 28EGM16093.

References

1. Bolstad B, Irizarry R, Astrand M, Speed T: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19(2)**:185–193.

2. Cope L, Irizarry R, Jaffee H, Wu Z, Speed T: **A benchmark for affymetrix genechip expression measures.** *Bioinformatics* 2004, **20(3)**:55–65.

3. Allison D, Cui X, Page G, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nature Reviews* 2006, **7**:55–65.

4. Choi J, Yu U, Kim S, Yoo O: **Combining multiple microarray studies and modelling inter-study variation**. *Bioinformatics* 2003, :i84–i90.

5. Parmigiani G, Garrett-Mayer E, Anbazhagan R, Gabrielson E: **A cross-study comparison of gene expression studies for the molecular classification of lung cancer**. *Clinical Cancer Research* 2004, **5:81**.

6. Conlon E, Song J, Liu J: **Bayesian models for pooling microarray studies with multiple sources of replications**. *BMC Bioinformatics* 2006, **7(247)**:1–13.

7. Conlon, EM and Song, JJ and Liu, JS: **Bayesian meta-analysis models for microarray data: a comparative study**. *BMC Bioinformatics* 2007, **8(80)**:1–21.

8. Scharpf R, Tjelmeland H, Parmigiani G, Nobel A: **A Bayesian model for cross-study differential gene expression**. *Technical Report* 2008, :1–42.

9. Carroll R: **Covariance analysis in generalized linear measurement error models.** *Statistics in Medicine* 1989, **9**:1075–1093.

10. Armstrong B: **The effects of measurement error on relative risks regression.** *American Journal of Epidemiology* 1991, **132**:1176–1184.

11. Carroll R, Ruppert D, Stefanski L: *Measurement error in nonlinear models.* Chapman & Hall, London 1995.

12. Thomas D, Gauderman W, Kerber R: **A non-parametric Monte Carlo approach to adjustment for covariate measurement errors in regression problems.** *Technical report, Department of Preventive Medicine, University of Southern California* 1991.

13. Richardson S, Gilks W: **A Bayesian approach to measurement error problems in epidemiology using conditional independence models.** *American Journal of Epidemiology* 1993, **138**:430–442.

14. Richardson S, Gilks W: **Conditional independence models for epidemiological studies with covariate measurement error**. *Statistics in Medicine* 1993, **12**:1703–1722.

15. Richardson S: *Markov Chain Monte Carlo in practice*, Chapman & Hall, London 1996 chap. Measurement error, :400–417.

16. Lunn D, Thomas A, Best N, , Spiegelhalter D: **WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility**. *Statistics and Computing* 2000, **10**:325–337.

17. **Latin Square Data Set**. [www.affymetrix.com/support/technical/sample_data/datasets.affx].

18. Lewin A, Richardson S, Marshall C, Glazier A, Aitman T: **Bayesian Modelling of Differential Gene Expression.** *Biometrics* 2006, **62**:1–9.

19. 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**:509–519.

20. Bochkina N, Richardson S: **Tail Posterior Probability for Inference in Pairwise and Multiclass Gene Expression Data.** *Biometrics* 2007, **in press, doi:10.1111/j.1541-0420.2006.00807.x**.

21. Gelman A, Meng X, Stern H: **Posterior predictive assessment of model fitness via realized discrepancies.** *Statistica Sinica* 1996, **6**:733–807.

22. Marshall E, Spiegelhalter D: **Approximate cross-validatory predictive checks in disease mapping models.** *Statistics in Medicine* 2003, **22**:1649–1660.

23. Marshall E, Spiegelhalter D: **Identifying outliers in Bayesian hierarchical models: a simulation-based approach**. *Bayesian Analysis* 2007, **2(2)**:409–444.

24. Lewin A, Bochkina N, Richardson S: **Fully Bayesian mixture model for differential gene expression: simulations and model checks**. *Statistical Applications in Genetics and Molecular Biology* 2007, **6(1)**.

25. Verwaerde C, Delanoye A, Macia L, Tailleux A, Wolowczuk I: **Influence of high-fat feeding on both naive and antigen-experienced T-cell immune response in** *DO*10.11 **mice.** *Scandinavian Journal of Immunology* 2006, **64**(5):457–466.

26. Masternak M, Bartke A: **PPARs in Calorie Restricted and Genetically Long-Lived Mice.** *PPAR Research* 2007, **ID28436**:1–7.

27. McGee M, Chen Z: **New Spiked-In Probe Sets for the Affymetrix HGU-133A Latin Square Experiment**. *COBRA preprint Series* 2006.

28. Affymetrix: *Statistical Algorithms Description Document* 2001.

29. Irizarry R, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T: **Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data.** *Biostatistics* 2003, **4**(2):249–264.

30. Li C, Wong W: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. *Genome Biology* 2001, **2**(8):32.1.

31. Koza R, Nikonova L, Hogan J, Rim J, Mendoza T, Faulk C, Skaf J, Kozak L: **Changes in gene expression foreshadow diet-induced obesity in genetically identical mice.** *PLoS Genetics* 2006, **2**(5):e81.

32. Yang X, Schadt E, Wang S, Wang H, Arnold A, Ingram-Drake L, Drake T, Lusis A: **Tissue-specific expression and regulation of sexually dimorphic genes in mice.** *Genome Research* 2006, **16**(8):995–1004.

33. Recinos A, Carr B, Bartos D, Boldogh I, Carmical J, Belalcazar L, Brasier A: **Liver gene expression associated with diet and lesion development in atherosclerosis-prone mice: induction of components of alternative complement pathway.** *Physiological Genomics* 2004, **19**:131–142.

34. **Diabetes Genome Anatomy Project website**. [www.diabetesgenome.org/arraydata.cgi].

35. Wang L, Balas B, Christ-Roberts C, Yeo Kim R, Ramos F, Kikani C, Li C, Deng C, Reyna S, Musi N, Dong L, DeFronzo R, Liu F: **Peripheral Disruption of Grb10 Gene Enhances Insulin Signaling and Sensitivity in vivo.** *Molecular and Cellular Biology* 2007, **MCB.00679-07**:1–33.

36. Ohtsubo K, Takamatsu S, Minowa M, Yoshida A, Takeuchi M, Marth J: **Dietary and Genetic Control of Glucose Transporter 2 Glycosylation Promotes Insulin Secretion in Suppressing Diabetes.** *Cell* 2005, **123**(7):1307–1321.

Figures

Figure 1: ROC curve for the simulation study

The plot shows the ROC curve for the Bayesian models averaged over the 10 simulated dataset: in each case we simulated 200 differentially expressed and 800 not differentially expressed genes. We have implemented Bayesian models either combining the five pre-processing methods together (solid line) or analyzing each one separately and ranked the tail posterior probability of differential expression. The ROC curve for the combined model is above that of each pre-processing method, highlighting the benefit of using a combined model in terms of specificity and sensitivity.

Figure 2: Mixed Posterior Predictive checks for the Latin square data set

The upper plots present the distribution of the mixed predictive p-values for the two conditions fitting the calibration model presented in (2) which assumes that the variability is specified by a polynomial function of the level of expression. The mixed predictive p-values show a uniform behavior for both conditions, indicating an excellent model fit. The values for λ_{1jk}, λ_{2jk} and λ_{3jk} are presented in Table 1 of File 1 in

Supplemental Material. The bottom plots present the distribution of the p-values for the two conditions after fitting the simpler model assuming $\lambda_{2jk} = \lambda_{3jk} = 0$. They indicate, in particular for condition 1, that the variance model is not appropriate.

Figure 3: Mixed Posterior Predictive checks for the HFD example

The upper plots present the distribution of the p-values for the two conditions (HFD and NFD) after fitting the calibration model presented in (2) which assumes a polynomial function of the level of expression. They show a uniform behavior for both conditions, indicating good model fit. The values for λ_{1jk}, λ_{2jk} and λ_{3jk} are presented in Table 2 of File 1 in Supplemental Material. The bottom plots present the distribution of the p-values for the two conditions (HFD and NFD) fitting the simpler calibration model assuming $\lambda_{2jk} = 0$ and $\lambda_{3jk} = 0$. The histograms indicate the presence of two peaks corresponding to very small and very large p-values, suggesting a poor fit.

Figure 4: Venn diagram

The figure shows the number of genes in the top list of 292 features for each method according to their relation with all the other methods. The number of genes in common between 1,2 or all the methods and the combined one are shown in parenthesis. For instance, in the intersection of MAS5 and RMA there are 7 genes; out of them 6 (shown in parenthesis) are in common also with the combined method.

Figure 5: Volcano plot for the HFD experiment

The plot shows the different behavior of the pre-processing methods: dChip has a small variability resulting in a compact volcano, while MAS5 is characterized by a large variability, that causes some genes with very different values of the log fold change to have very similar posterior probability. The combined method shows a distribution close to dChip. The 292 genes called differentially expressed by the combined model using a cut off of 0.98 on the tail posterior probability scale are highlighted in red in all the plots.

Tables

Table 1: Operating characteristics for simulated data

The table presents the number of False Positives (FP), True Negatives (TN), True Positives (TP) and False Negatives (FN) in the first 200 genes ranked accordingly to their tail posterior probability. Note that $FP = FN$ since the size of the list of differentially expressed genes is equal to the number of *true positives*. The combined method shows the smallest number of FP and FN. Out of the 5 pre-processing methods, the one characterized by a relative bias parameter $\phi_j > 1$ and a variability parameter $exp(\lambda_j) < 1$ (Method 3) shows the best performance, due to the combination of high signal and low variability around the mean gene expression.

	DE	Non DE	FP (%)	TN (%)	TP (%)	FN (%)
Combined	200	800	21 (2.6)	779 (97.4)	179 (89.5)	21 (10.5)
Method 1 ($exp(\lambda) = 1, \phi = 1$)	200	800	61 (7.6)	739 (92.4)	139 (69.5)	61 (30.5)
Method 2 ($exp(\lambda) = 0.5, \phi = 0.5$)	200	800	80 (10.0)	720 (90.0)	120 (60.0)	80 (40.0)
Method 3 ($exp(\lambda) = 0.5, \phi = 2$)	200	800	26 (3.2)	774 (96.8)	174 (87.0)	26 (13.0)
Method 4 ($exp(\lambda) = 2, \phi = 0.5$)	200	800	122 (15.2)	678 (84.8)	78 (49.0)	122 (61.0)
Method 5 ($exp(\lambda) = 2, \phi = 2$)	200	800	45 (5.6)	755 (94.4)	155 (77.5)	45 (22.5)

Table 2: Operating characteristics for simulated data

The table presents the operating characteristics of the combined method and of each pre-processing method on the first 64 genes ranked accordingly to their tail posterior probability. The combined strategy is more able to recognize true positives and true negatives that each single method. Note that $FP = FN$ since the size of the list of differentially expressed genes is equal to the number of *true positives*.

	First 64 ranked genes			
	FP (%)	TN (%)	TP (%)	FN (%)
Combined	12 (0.1)	10545 (99.9)	52 (81.3)	12 (18.7)
MAS5	23 (0.2)	10534 (99.8)	41 (64.1)	23 (35.9)
RMA	14 (0.1)	10543 (99.9)	50 (78.1)	14 (21.9)
dChip	31 (0.4)	10526 (99.6)	33 (51.6)	31 (48.4)

Table 3: Posterior mean and credibility interval for ϕ_j

Posterior mean and 95% credibility intervals for the relative bias effect ϕ_j in the Latin Square data set (left) and in the real data analysis that compares the effect of high fat diet (HFD) and normal fat diet (NFD) on adipose tissue of mice (right).

	Latin Square data set		Real Experiment: HFD vs NFD	
	$E(\phi_j \mid \mathbf{y})$	$CI95\%$	$E(\phi_j \mid \mathbf{y})$	$CI95\%$
MAS5	1.382	$[1.345 - 1.417]$	1.449	$[1.438 - 1.462]$
RMA	1.144	$[1.124 - 1.165]$	1.064	$[1.056 - 1.074]$
dChip	0.632	$[0.623 - 0.644]$	0.648	$[0.643 - 0.653]$

Table 4: Annotation for differentially expressed genes

The table presents the number of differentially expressed genes annotated by GO or KEGG considering the first 292 genes with the largest tail posterior probability for each method and for the combined one. The enriched KEGG pathways are mainly related to immune response and oxidation (*Antigen processing and presentation*, *MAPK signalling pathway*, *PPAR signaling pathway*). For these pathways the number of genes found by the combined method is always larger than the one found by each pre-processing method. Moreover the combined model also calls 6 genes that are involved in the *Insulin signalling pathway*, specifically related to diabetes, one of the diseases recently highlighted to be associated with high fat diet in mice (see for example [35] and [36]). The same pathway is present with only 3 genes in MAS5 and with 2 genes in RMA and dChip.

First 292 genes with the largest posterior probability

	Combined model	MAS5	RMA	dChip
Biological Processes	181	146	171	168
Molecular Functions	193	154	180	173
Cellular Components	179	145	173	169
KEGG pathways	223	182	215	205

Table 5: Performance of simulated data on λ and ϕ

The table reports the posterior mean with the 95% credibility interval for the measurement error parameters, together with the true values set in the simulation. For all the parameters the posterior mean coincides with the true value and is characterized by a small variability around it. The values are averaged over 10 runs.

	$E(\lambda_j \mid \mathbf{y})[CI95\%]$	$E(\phi_j \mid \mathbf{y})[CI95\%]$
Method 1 ($\lambda = 1, \phi = 1$)	1.0 [1.0-1.0]	1.0 [1.0-1.0]
Method 2 ($\lambda = 0.5, \phi = 0.5$)	0.5 [0.49-0.51]	0.5 [0.49-0.51]
Method 3 ($\lambda = 0.5, \phi = 2$)	0.50 [0.49-0.51]	2.0 [1.97-2.04]
Method 4 ($\lambda = 2, \phi = 0.5$)	2.0 [1.96-2.07]	0.5 [0.48-0.51]
Method 5 ($\lambda = 2, \phi = 2$)	2.0 [1.97-2.04]	2 [1.99-2.03]

Table 6: Characteristics of MAS5, RMA and dChip

Characteristics of MAS5, RMA and dChip in terms of perfect match correction, normalization and summarization method.

	Background correction	Perfect Match correction	Normalisation	Summary
MAS5	Divide the chip in 16 regions. The lowest 2% is the background. Weighted average over all the regions	Ideal Mismatch	Scaling	1 step Tukey Biweight
RMA	Global model for the distribution of the probe intensities	No correction	Quantile	Median Polish
dChip	No correction	No correction	Invariant set using one array as default	Multi-chip linear model

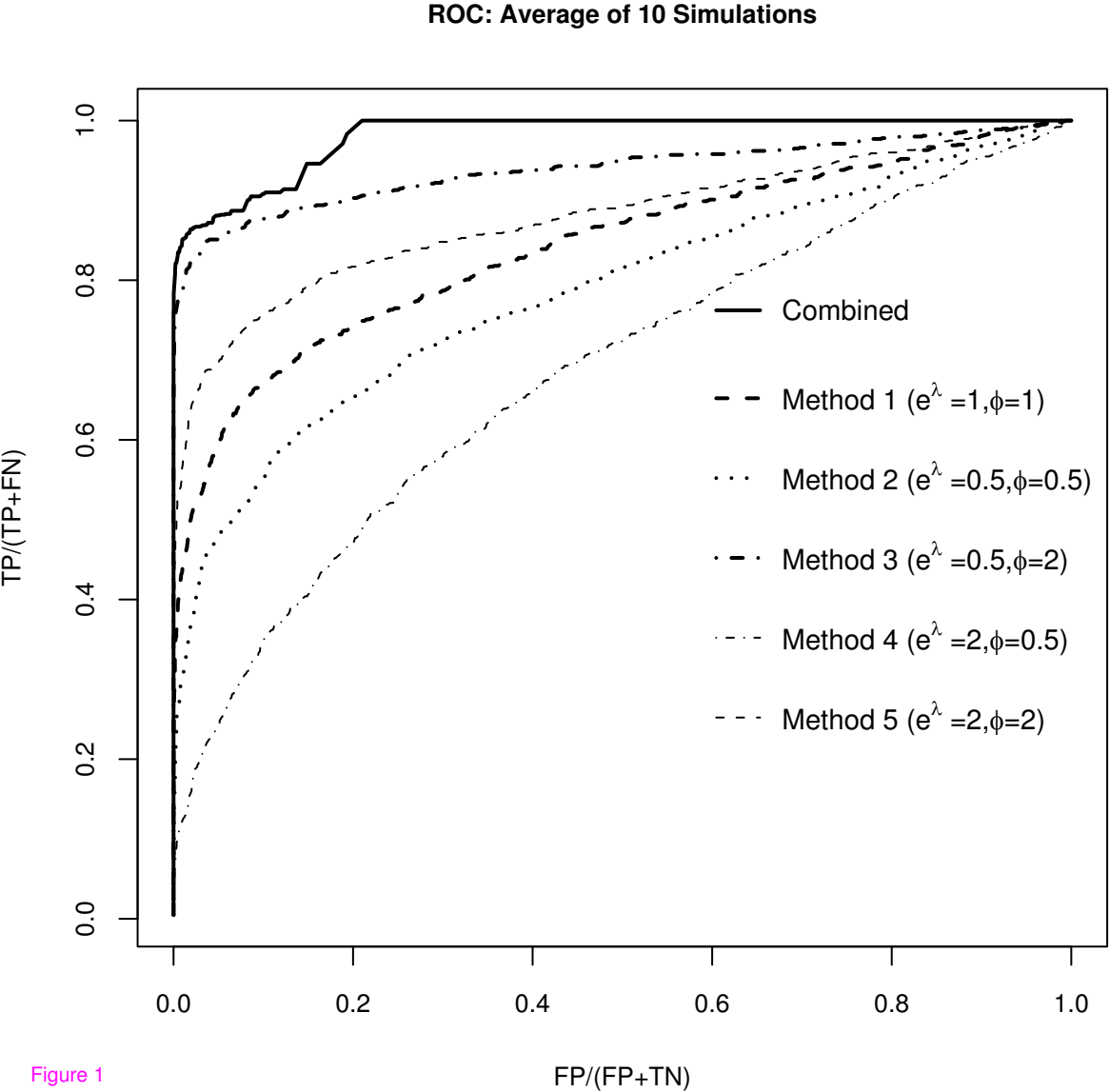
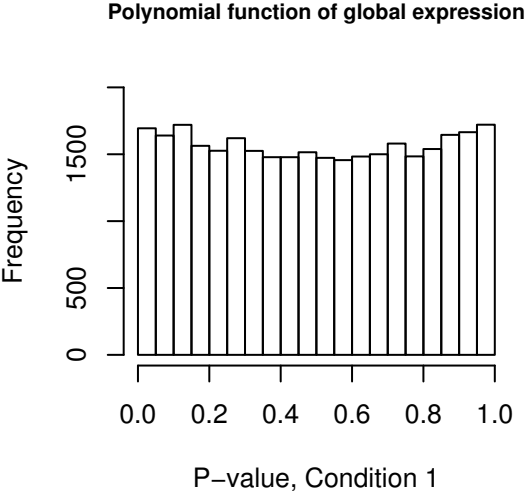
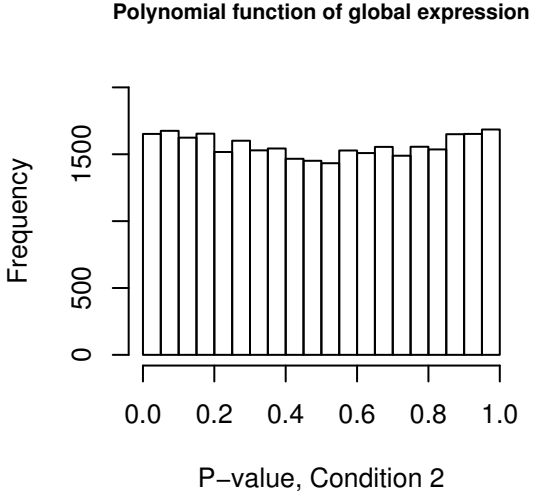
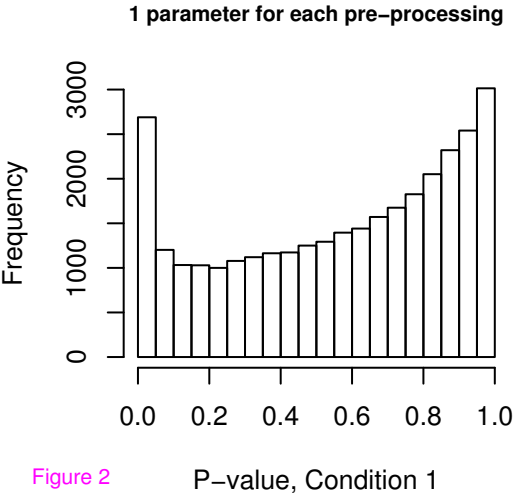

ROC: Average of 10 Simulations
TP/(TP+FN)
FP/(FP+TN)
0.0
0.2
0.4
0.6
0.8
1.0
Combined
Method 1 (e^λ =1,ϕ=1)
Method 2 (e^λ =0.5,ϕ=0.5)
Method 3 (e^λ =0.5,ϕ=2)
Method 4 (e^λ =2,ϕ=0.5)
Method 5 (e^λ =2,ϕ=2)

Figure 1


Polynomial function of global expression
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
P−value, Condition 1


Polynomial function of global expression
Frequency
0
500
1500
0.0
0.2
0.4
0.6
0.8
1.0
P−value, Condition 2


1 parameter for each pre−processing
Frequency
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
P−value, Condition 1

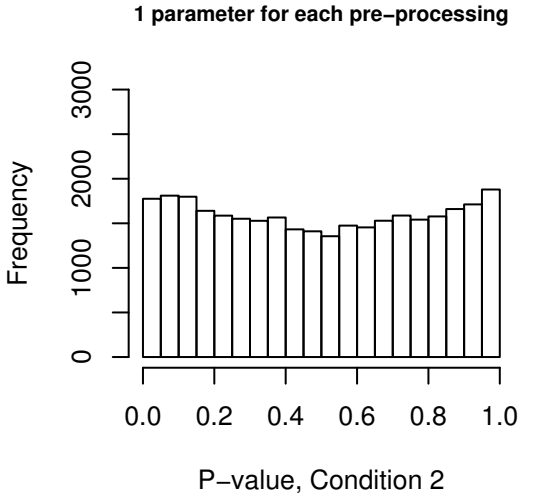
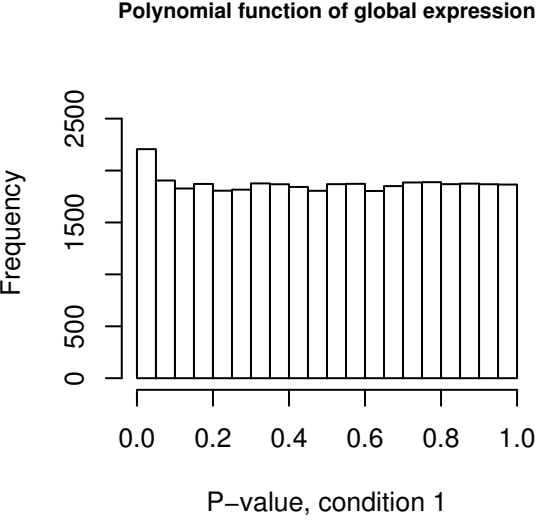
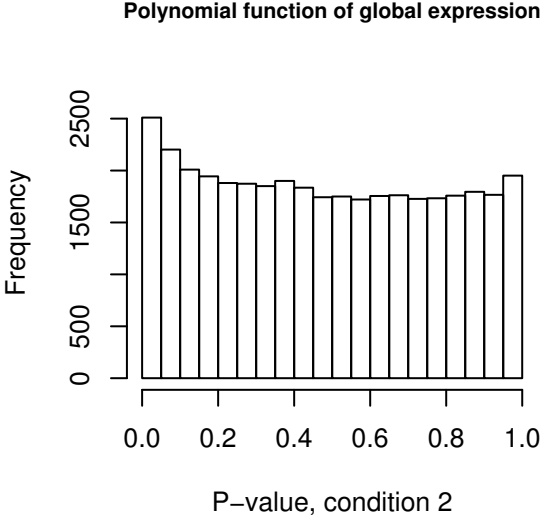
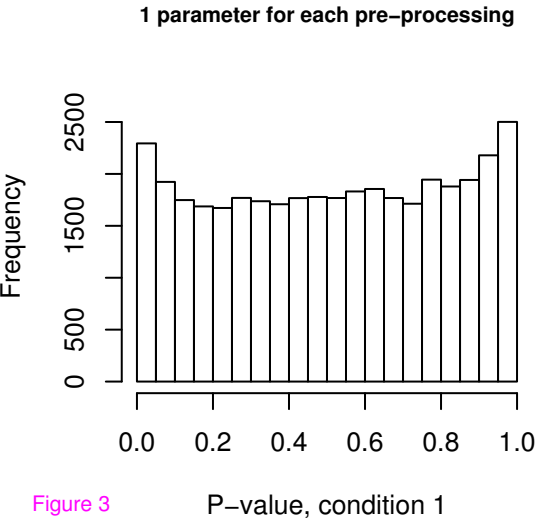

1 parameter for each pre−processing
Frequency
0
1000
2000
3000
0.0
0.2
0.4
0.6
0.8
1.0
P−value, Condition 2

Figure 2


Polynomial function of global expression
Frequency
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
1.0
P−value, condition 1


Polynomial function of global expression
Frequency
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
1.0
P−value, condition 2


1 parameter for each pre−processing
Frequency
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
1.0
P−value, condition 1

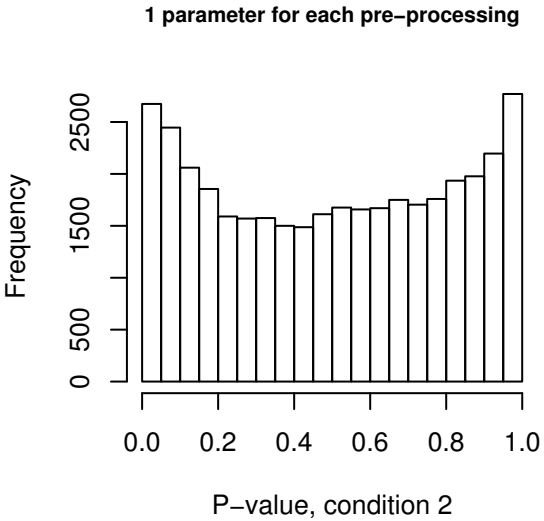

1 parameter for each pre−processing
Frequency
0
500
1500
2500
0.0
0.2
0.4
0.6
0.8
1.0
P−value, condition 2

Figure 3

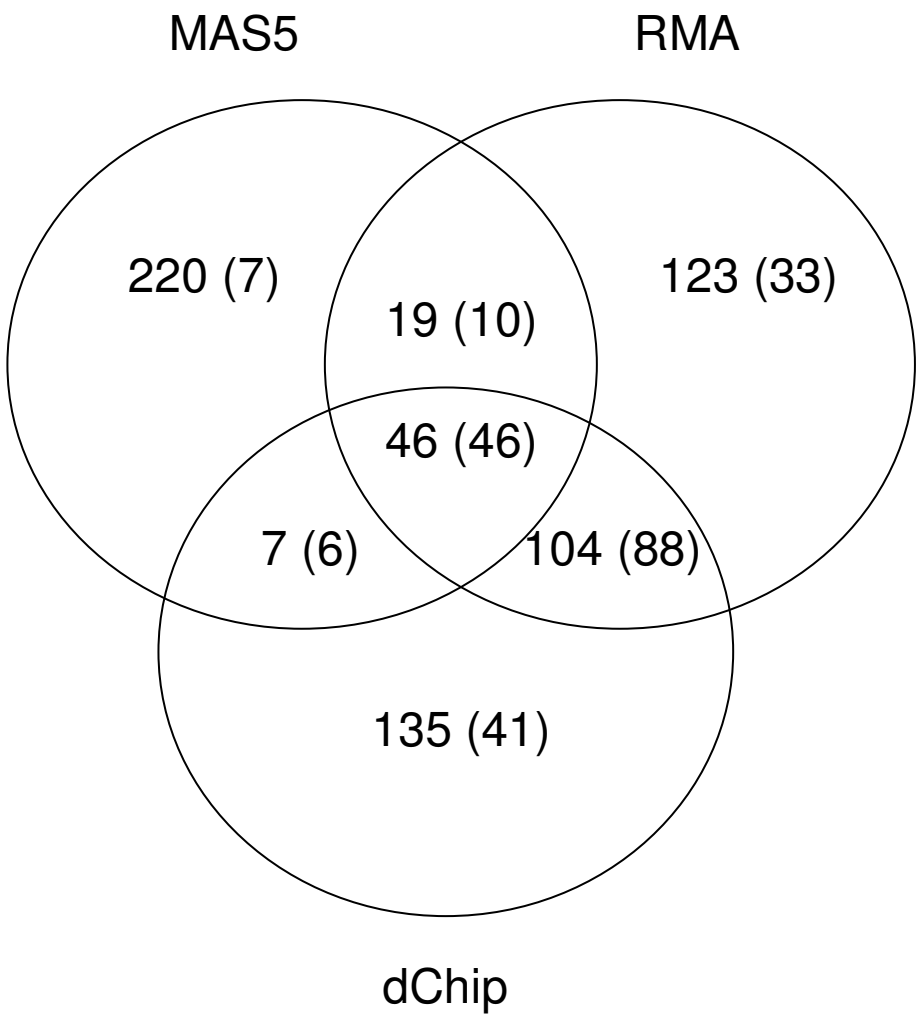


Figure 4

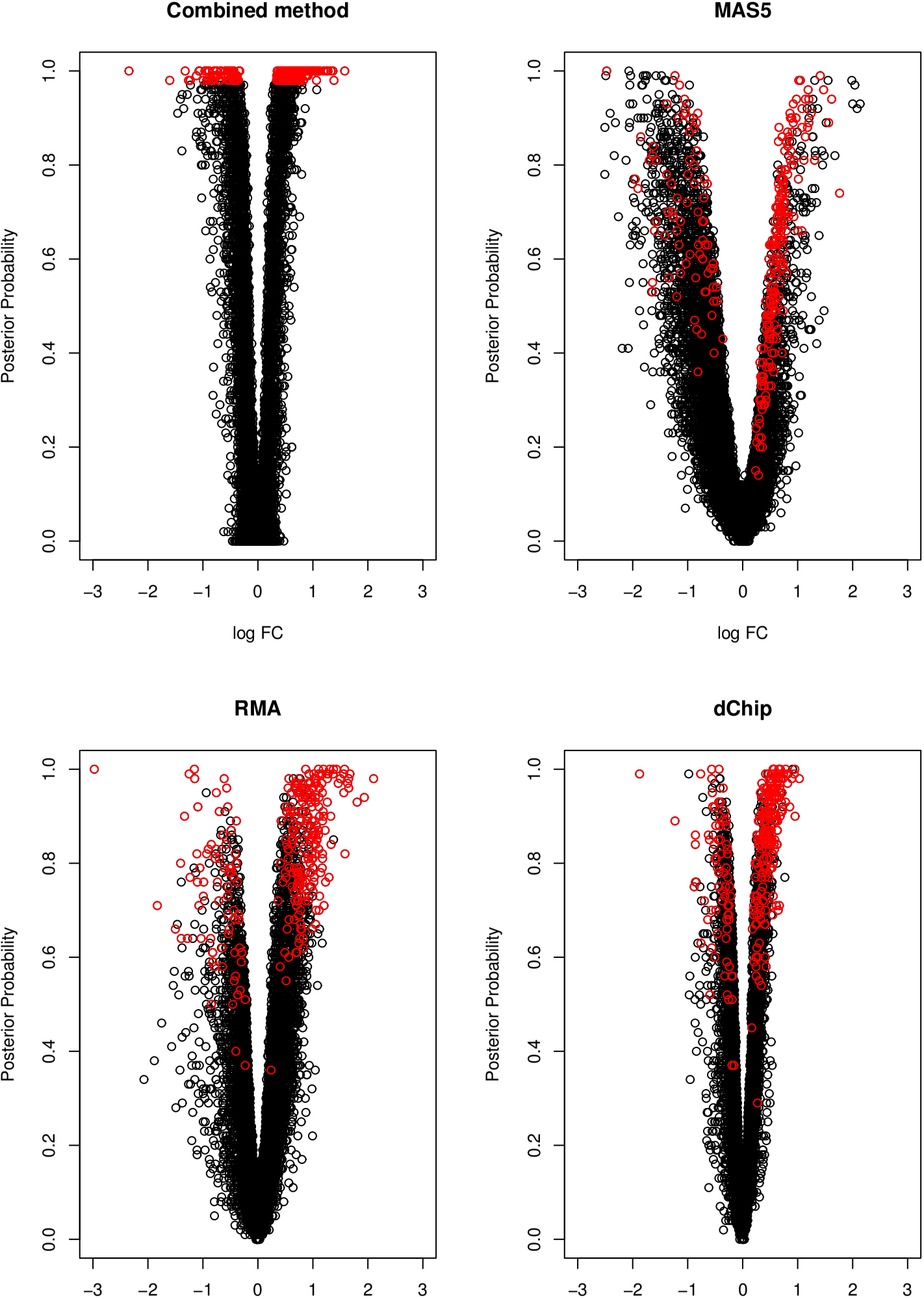

Combined method
MAS5
RMA
dChip
Posterior Probability
log FC
0.0
0.2
0.4
0.6
0.8
1.0
−3
−2
−1
0
1
2
3

Figure 5

Additional files provided with this submission:

Additional file 1: file1.pdf, 143K
http://www.biomedcentral.com/imedia/8872108622000024/supp1.pdf