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

Marta Blangiardo, Sylvia Richardson

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

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 the sources of variability due to the technology used rather than to biological differences between the RNA samples or between the printed probes. Many studies have shown how this can influence the results in terms of differential expression, but at the moment there is no commonly agreed gold standard 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 differential expression between two conditions. Our approach is based on the assumption that each method provides a measure of the “true” differential expression with a method specific error. The freely available software WinBUGS can be used to estimate the posterior distribution of the differential expression values of interest. The model is tested on simulated data and its biological interest is demonstrated through a real example on publicly available data.

Key words: Measurement error model, Bayesian models, affymetrix chips

1 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 or between the printed probes. In the literature there are many studies presenting the importance of pre-processing and showing how this can influ-
ence the results in terms of differential expression (see, for example, (7) and (9)).

However, an agreed gold standard does not exist and as Allison discusses in a recent paper (2), the researchers are torn between the different methods and usually end up restricting their analysis to using only one method (often the most used or 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 first strategy is reductive and the latter relies on the arbitrary choice of two methods and of looking only at their intersection. Both these approaches do not make optimal use of all the information provided by the pre-processing methods.

We apply ideas from Bayesian calibration to this issue, with the aim of providing a more efficient analysis of differential expression experiments: we assume that each method returns an unbiased measure of the true differential expression and is characterized by an error. Modeling measurement error is common practice in epidemiology, where errors in the measurement of explanatory variables is a frequent problem that has to be taken into account during the statistical analysis (see for example (8), (3)); the formulation in a Bayesian framework has been discussed in the early nineties among by Thomas (20), and Richardson and Gilks (19), among others, placing particular emphasis on the way the approach can propagate coherently every source of uncertainty in the data to the estimation of the parameters of interest. We adapt the work of Richardson and Gilks and specify a Bayesian calibration model for assessing the differential expression in Affymetrix microarray, which pools together the information from several pre-processing techniques. The freely available software WinBUGS can be used to estimate the posterior distribution of the differential expression values of interest. The model performance is tested using simulated data; we also present 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.

2 Methods

Bayesian Model

We introduce a Bayesian model which combines the different pre-processing methods as measures of the same “true” effect, each with an intrinsic error (measurement error). Following Lewin (13), the observed log expression value for gene $g = 1, \ldots, G$, pre-processing $j = 1, \ldots, J$, condition $k = 1, 2$ and
replicate $r = 1, \ldots, R$ is modeled as a Normal distribution:

$$y_{gj1r} \sim N(\alpha_{gj} - \frac{1}{2} \cdot \delta_g, \sigma_{gj}) \quad y_{gj2r} \sim N(\alpha_{gj} + \frac{1}{2} \cdot \delta_g, \sigma_{gj})$$ (1)

The parameter $\alpha_{gj}$ represents the global gene expression for gene $g$ and the pre-processing method $j$ whereas $\delta_g$ is the “true” (unknown) differential expression for gene $g$. In order to do this, we assume that each pre-processing gives an unbiased estimate of the differential expression $\delta_g$, but is characterized by a measure of variability we need to calibrate. The variance $\sigma_{gj}^2$ for each gene and pre-processing is then the result of a gene specific component and a measurement (pre-processing) error (indexed by $j$):

$$\sigma_{gj}^2 = \sigma_g^2 \times \lambda_j$$ (2)

This key multiplicative decomposition of the variability of gene $g$ for method $j$ will allow the borrowing of information across genes to estimate $\lambda_j$ and across methods to estimate $\sigma_g^2$. Generalization of 2 to allow more flexibility of modeling the component of variance will be discussed in Section 4. We assign a hierarchical structure on the gene specific component, to borrow strength from the entire set of genes:

$$\sigma_g^2 \sim Ga^{-1}(a, b)$$ (3)

where $a \sim Ga(0.01, 0.01)$ and we model $1/\sqrt{b} \sim U[0, \min_g(s_g^2)^{-0.5}]$, where $s_g^2$ is the sample variance for gene $g$; it ensures that the posterior distribution of $a$ and $b$ is proper, as described in Bochkina and Richardson (6).

The errors are modeled as independent for the $j = 1, \ldots, J$ pre-processing methods: $\lambda_j \sim Ga(0.01, 0.01)$, imposing the identifiability constraint that $\lambda_1 = 1$. Thus the $\lambda_j$ should be interpreted as a relative error rate with respect to the method chosen as baseline.

**Tail Posterior Probability**

For each gene we are interested in testing the hypothesis that the differential expression effect $\delta_g$ is different from 0. We can formulate it as follow:

$$H_0^g : \delta_g = 0 \quad \text{vs} \quad H_1^g : \delta_g \neq 0$$
and we use the tail posterior probability to measure the strength of the evidence against \( H_0 \), as proposed in (6). We specify the following standardisation for \( \delta_g \):

\[
t_g = \frac{\delta_g}{\sqrt{w_g}}
\]

where \( w_g = \frac{2}{RJ^2} \sum_{j=1}^{J} \sigma^2_{gj} \), \( R \) is the number of replicates for each condition and \( J \) is the number of pre-processing methods considered; then the tail posterior probability that a gene is differentially expressed between two conditions can be calculated as follows:

\[
p(t_g; t^\alpha) = P(|t_g| > t^\alpha | y_{gjk})
\]

The procedure requires to define the quantile of the Normal distribution \( t^\alpha \); typically \( \alpha = 0.95 \) and consequently \( t^\alpha = 1.96 \). When a posterior probability \( p(t_g; t^\alpha) \) is obtained for each gene we need a cut off to select the differentially expressed genes. In order to do this, the tail posterior probability method allows to estimate a Bayesian False Discovery Rate (FDR), which gives an indication of the error we incur in selecting a certain cut off on the posterior probability scale. It is then possible to define the maximum level of error we want to allow (through the FDR) and define the corresponding cut off on the posterior probability scale.

The same formula for calculating the tail posterior probability can be easily adapted when the number of replicates \( R \) is different for each condition, or a variance for each method and condition is assumed (\( \sigma^2_{gjk} \)).

*Posterior Predictive Check*

The Bayesian framework allows to perform model checks by means of the predictive distribution of parameters of interest. We use the Mixed Posterior Predictive Check ((10), (15)), applied on gene expression data by (13) and focus attention on the gene specific variance, characterized by a hierarchical structure as described in Equation 3. We compare the observed sample variance, calculated on the expression values, and the variance of the predicted expression values of each gene under the model using an empirical p-value. First a new value for the gene specific component of the variance (\( \sigma^2_{g(\text{pred})} \)) is predicted from its prior distribution \( Ga^{-1}(a, b) \) and then new data are generated under the chosen model and current values of the parameters \( \alpha_{gj}, \delta_g \) and \( \lambda_j \), where the variance has been replaced by the predicted one:
Here \( \sigma_{\text{gj}}^2 = \sigma_{\text{gj}}^2 \times \lambda_j \). For each gene and pre-processing method the observed sample variance \( S_{\text{gj}}^2 = \frac{1}{2R-1} \sum_{r=1}^{2R} (y_{\text{gjkr}} - \bar{y}_{\text{gj}})^2 \) is compared to the predicted sample variance \( S_{\text{gj}}^{2(\text{pred})} = \frac{1}{2R-1} \sum_{r=1}^{2R} (y_{\text{gjkr}}^{(\text{pred})} - \bar{y}_{\text{gj}}^{(\text{pred})})^2 \) through the statistic \( S_{\text{gj}}^2 - S_{\text{gj}}^{2(\text{pred})} \) and a p-value is generated. Under the null hypothesis of the model being true, the distribution of the p-values should be uniform (5), while a poor model fit is indicated by the presence of a peak or a trend in the plot, suggesting a systematic difference between the observed values and the predicted ones (see Lewin et al., (13) for more details).

3 Materials

Simulated data

To test the performance of our method we simulated log expression values for 1000 genes, two conditions and three 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 \( y_{\text{gjkr}} \sim N(\alpha_{\text{gj}} + \frac{1}{2} \delta_{\text{g}}, \sigma_{\text{gj}}^2) \) while for the remaining 800 genes the log expression for both conditions comes from the same distribution: \( y_{\text{gjkr}} \sim N(\alpha_{\text{gj}}, \sigma_{\text{gj}}^2) \). The distribution of \( \alpha_{\text{gj}} \sim N(6.79, 4.77) \) and \( \delta_{\text{g}} \sim N(-0.04, 0.25) \) have parameters obtained from real data we have analyzed; the measurement errors for the three pre-processing are \( \lambda_1 = 1 \) (for identifiability purpose), \( \lambda_2 = 3 \), \( \lambda_3 = 0.5 \) and the gene specific variability \( \sigma_{\text{g}}^2 \) is again obtained from real data and is characterized by the quartiles 0.02, 0.04, 0.09, 0.15.

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. Figure 1 of supplementary material presents the distribution of \( \lambda_2 \) and \( \lambda_3 \) with their credibility 95% intervals, showing that the estimates from different samples remain stable.

As a point of comparison we also run the model separately for each pre-processing method and evaluate their performance in terms of sensitivity and specificity (see Section 4 for the results on simulated data).
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 (12) and (23) for adipose tissue and (18) 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 run it on a real publicly available experiment to study the effect of high fat diet (HFD) versus Normal fat diet (NFD) on mice adipose tissue. The arrays are Affymetrix MGU74Av2 and contain 12488 probesets. 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 on http://www.diabetesgenome.org/arraydata.cgi.

We focused attention on the three of pre-processing methods most used: MAS5 (1), RMA (11) and dChip (14). There are many version of each of them, but we considered the default ones provided by the software R (www.r-project.org). The differences in the three methods are described in table 1:

| Method | Background correction | Perfect Match correction | Normalisation | Summary |
|--------|-----------------------|--------------------------|---------------|---------|
| MAS5   | Divide the chip in 16 regions | Ideal Mismatch correction | Scaling | 1 step Tukey Biweight |
|        | The lowest 2% is the background | Weighted average over all the regions |               |         |
| 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 |

Table 1
Characteristics of MAS5, RMA and Dchip: Perfect match correction, Normalization and Summarization used by each of them

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).
4 Results

Simulated Data

On simulated data the model estimates correctly the values of the parameters $\lambda_2$ and $\lambda_3$ specific to the pre-processing methods (recall that $\lambda_1$ is set to 1 for identifiability purpose). In Table 2 the posterior mean and 95% credibility intervals are reported together with the real values of the parameters set in the simulation: it shows that the posterior mean is equal to the real value of the parameters with small variability around it. The values are averaged over the 10 simulations. The Mixed Posterior Predictive Check shows a uniform distribution for the p-value, suggesting, as could be expected, a good fit of the model to the data (the histograms for the 10 simulations are presented in Figures 2 and 3 of supplementary material) The typical behavior of the

| Posterior Mean ($CI_{95\%}$) | True Value |
|-------------------------------|------------|
| $E(\lambda_2 \mid y) = 3.0 \ (2.89 - 3.10)$ | 3          |
| $E(\lambda_3 \mid y) = 0.5 \ (0.48 - 0.52)$ | 0.5        |

Table 2
Posterior Mean and true values for $\lambda_2$ and $\lambda_3$: the table reports the estimated posterior mean with the 95% credibility intervals for $\lambda_2$ and $\lambda_3$ together with the true values set in the simulation. For both $\lambda$, the posterior mean coincides with the true value of the parameter and is characterized by small variability around it. The values are average over the 10 runs.

combined method compared to each pre-processing one is presented in Figure 1: the ROC averaged over the 10 runs shows a greater sensitivity and specificity for the model that combines the three pre-processing. As expected, the method characterized by the smallest error ($\lambda_3 = 0.5$) performs better than the others and it is the closest to the curve of the combined model; the gap is larger for small values of the False Positive Rate ($FPR$) and start decreasing as the rate increases. Understandably, the method characterized by a large error ($\lambda_2 = 3$) shows the worst performance, reaching a value of 0.5 for the True Positive Rate ($TPR$) when the $FPR$ is around 0.2. Each run confirms the average results as show the ROCs plots presented in Figure 4 of supplementary material.

Table 3 reports a synthesis of the results averaged over the 10 simulations: for each cut off on the posterior probability we calculated the number of differentially and not differentially expressed genes ($DE$ and $\overline{DE}$), the number of false positives ($FP$), true positives ($TP$), false negatives ($FN$), true negatives ($TN$), the false discovery rate ($FDR = FP/DE$) and the false non discovery rate ($FNR = FN/\overline{DE}$). We report the information for the combined method as well as for each pre-processing. As we introduced in section 2 the $FDR$
can be used to select a cut off on the tail posterior probability scale and consequently to obtain a list of differentially expressed genes. Typically we do not want to introduce a rate of false positives larger than 1%, which would correspond to a posterior probability around 0.85 in Table 3. For a cut off of 0.85, 105 genes are found differentially expressed by the combined strategy, while each pre-processing calls a smaller list. In particular, the third method, characterized by the smallest $\lambda_j$ is very close to the combined, finding 104 genes, while the other two methods call only 79 and 40 genes, as expected, as they are characterized by a larger variability which influences the posterior probability estimate.

The combined method shows the best specificity and sensitivity, being characterized by the smallest number of false positives (73 genes) and false negatives (96 genes); again amongst the three pre-processing methods, the one characterized by the smallest error shows a better performance in terms of false positives and false negatives. For the sake of completeness we show also the information for a cut off smaller than 0.85, but we do not recommend to use it as the FDR increase, meaning that more false positives are included in the list of differentially expressed genes. On the other hand, considering a posterior probability cut off larger than 0.85, the number of differentially expressed genes decreases and the gap between the performance of the combined method and of each pre-processing increases, calling more true positives than the other methods.
Fig. 1. ROC curve for the Bayesian models averaged over the 10 simulated dataset: we have simulated 200 differentially expressed and 800 not differentially expressed genes. We have run the same Bayesian model combining the three pre-processing together (solid line) and treating each method separately. We have then ranked the tail posterior probability of differential expression. The ROC for the combining model is on top of each pre-processing method, indicating the best performance in terms of specificity and sensitivity.
| PostProb | Theoretical | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined | Combined |
|----------|-------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| DE       | DE         | FP       | FP (%)   | TP       | TP (%)   | FN       | FN (%)   | TN       | TN (%)   | FDR      | FNR      | FDR      | FNR      | FDR      | FNR      | FDR      | FNR      |
| 0.4      | 157.441    | 21.25   | 136.68   | 64.32    | 779.97   | 0.134    | 0.076    |          |          |          |          |          |          |          |          |          |
| 0.6      | 128.872    | 6.8     | 121.60   | 79.39    | 799.99   | 0.050    | 0.090    |          |          |          |          |          |          |          |          |          |
| 0.8      | 110.890    | 1.01    | 109.55   | 91.45    | 799.99   | 0.012    | 0.103    |          |          |          |          |          |          |          |          |          |
| 0.9      | 100.900    | 0.00    | 100.50   | 100.50   | 800.100  | 0.002    | 0.112    |          |          |          |          |          |          |          |          |          |
| 0.99     | 90.910     | 0.00    | 90.45    | 110.55   | 800.100  | 0.000    | 0.121    |          |          |          |          |          |          |          |          |          |
| 0.998    | 32.968     | 0.00    | 32.16    | 168.84   | 800.100  | 0.012    | 0.174    |          |          |          |          |          |          |          |          |          |
| 0.999    | 9.991      | 0.00    | 9.45     | 191.95   | 800.100  | 0.017    | 0.193    |          |          |          |          |          |          |          |          |          |
| 0.9999   | 58.944     | 0.00    | 55.27    | 145.72   | 800.100  | 0.004    | 0.153    |          |          |          |          |          |          |          |          |          |
| 1        | 55.945     | 0.00    | 55.27    | 145.72   | 800.100  | 0.000    | 0.154    |          |          |          |          |          |          |          |          |          |
| 1.000    | 12.988     | 0.00    | 12.00    | 188.94   | 800.100  | 0.000    | 0.191    |          |          |          |          |          |          |          |          |          |
| 1.0000   | 2.999      | 0.00    | 2.00     | 198.99   | 800.100  | 0.000    | 0.199    |          |          |          |          |          |          |          |          |          |
| 1.00000  | 26.974     | 0.00    | 26.13    | 174.87   | 800.100  | 0.000    | 0.187    |          |          |          |          |          |          |          |          |          |

**Table 3:**
Average simulation results: we show the results of the Bayesian model combining the pre-processing methods, as well as considering each pre-processing separately. For each cut off on the posterior probability scale, we present the number of genes called differentially and not differentially expressed (FP, DE and DE), the false positives (FP), the false negatives (FN), the true positives (TP), the true negatives (TN), the False Discovery Rate (FDR) and the False non Discovery rate (FNR). The combining strategy has greater sensitivity and specificity than each pre-processing methods. The results are averaged over 10 simulations.
**Biological example: High Fat Diet versus Normal Fat Diet in mice adipose tissue**

The Mixed Posterior Predictive Check does not show an optimal fitting for the High Fat Diet experiment. The upper plots in Figure 2 present the p-values distribution for the 12488 genes and the two conditions (HFD and NFD). It shows a deviation from uniformity for very small and very large p-values indicating that for some genes the observed variability is much smaller or much higher than the predicted one. This suggests that the one parameter model of variability presented in Equation 2 is too simplistic and needs to be extended to model properly real data, where the degree of complexity is much higher. To obtain a more flexible model, we allow the error ratios between the methods (λ) to be linked to the global gene expression, as it has often been noted that even after log transformation, the variability can be affected by the level of global expression (see for example (4)). To be precise for each pre-processing \( j = 1, 2, 3 \) and condition \( k = 1, 2 \) we now specify \( \sigma_{gjk}^2 \) as the exponential of a polynomial of the sampling mean log expression \( \bar{y}_g \) as follow:

\[
\sigma_{gjk}^2 = \exp(\lambda_{1jk} + \lambda_{2jk} \times \bar{y}_g + \lambda_{3jk} \times (\bar{y}_g)^2) \times \sigma_{gk}^2. \tag{4}
\]

In this way the coefficient that multiply the gene expression variance depends on the global expression of that gene. The coefficients are assumed independent and follow a Normal non informative distribution \( \lambda_{1jk} \sim N(0, 0.00001), \lambda_{2jk} \sim N(0, 0.00001), \lambda_{3jk} \sim N(0, 0.00001) \), so that the function can range over all the \( \mathbb{R} \) space and model every trend. We use the exponential parametrization to ensure positivity. As identifiability constraint, we impose that the mean of each parameter over the three pre-processing to be equal to 0 (\( E[\lambda_{1,k}] = 0, E[\lambda_{2,k}] = 0, E[\lambda_{3,k}] = 0 \)). The use of a second order polynomial seems a good candidate as it offers a great improvement in flexibility from the linear function but at the same time it involves the estimate of a limited number of parameter.

The Mixed Posterior Predictive Check on this model results in a clear improvement of the plot, with a disappearance of the two peaks for extreme p-values and a more uniform distribution (see Figure 3, bottom). Note that assuming the same variance \( (\sigma_{g1}^2 = \sigma_{g2}^2) \) for the two conditions causes a worse fitting, with the p-value distribution characterized by a defined trend deviating from the uniformity (see Figure 5 of supplementary material for the plot obtained from the Mixed Posterior Predictive Check). We considered a cut off of 0.95 on the posterior probability scale and obtained 300 genes differentially expressed by the combined method, 41 by MAS5, 58 by RMA and 194 by DChip. Figure 3 presents the plot of the log fold change versus the posterior probability (volcano plot) for the combined model and for each pre-processing.
Fig. 2. Mixed Posterior Predictive Check for the HFD example. The upper plots present the distribution of the p-values for the two conditions (HFD and NFD) fitting the model with only one parameter ($\lambda_j$) for each pre-processing method. The histograms indicate the presence of two peaks corresponding to very small and very large p-values. The bottom plots present the distribution of the p-values for the two conditions (HFD and NFD) fitting the model with the polynomial function of the global expression for each gene and show a uniform behavior for both conditions.

separately. The genes found only by the combined method (highlighted in red) 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 when the gene shows some evidence of differential expression for two methods the combined approach strengthen that evidence and pushes its posterior probability at the top of the list of differentially expressed genes. This results is confirmed by the genes called differentially expressed by two or three methods (highlighted in green and blue respectively), that are always found at the top of the list for the combined model. On the other hand, MAS5 is the method that contributes least to the combined output, being the one characterized by the largest variability. For this reason, even though some genes show a low posterior probability for it, they can still be pushed to the top of the list by the combined method, if their values for RMA and Dchip are large enough.

Table 4 presents the number of annotated genes in each differentially expressed gene list. Looking specifically at the GO annotation at the level of Biological Processes, Molecular Functions and Cellular Components, the combined method seems to enrich it. In particular, the most represented biological pro-
Fig. 3. Volcano for HFD experiment. The plot shows 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 genes called differentially expressed only by the combined model are highlighted in red, while the one called by two and three methods are in green and blue respectively.

cesses are the metabolic ones, representing function associated with the response of the body to a change in the diet. The number of genes involved in cellular metabolism, primary metabolism or macromolecule metabolism is between 79 and 109 for the combined method, but it only varies from 47 to 70 for DChip, from 5 to 10 for MAS5 and from 12 to 17 for RMA. Also the combined model enriches some processes related to inflammation and immune response of the organism (immune system process, response to stress, response to external stimulus, defense response).

Moreover, the KEGG pathways are enriched through the combined method, are 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 (21) and (16)). 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 finds some pathways specifically related to diabetes (Insulin signalling pathway and Type I
diabetes mellitus), one of the diseases recently discovered to be associated to high fat diet (see for example (22) and (17)). Both these pathways are not found by MAS5, only 1 genes for each of them is called by RMA, while 4 genes belonging to Insulin signalling pathway and only 1 to Type I diabetes mellitus are found by Dchip.

| Biological Processes | Combining method | MAS5 | RMA | DChip |
|----------------------|------------------|------|-----|-------|
| 182                  | 19               | 41   | 118 |
| Molecular Functions  | 194              | 22   | 39  | 124   |
| Cellular Components  | 190              | 25   | 38  | 114   |
| KEGG pathways        | 233              | 29   | 45  | 144   |

Table 4
Annotation for the list of differentially expressed genes found by the different methods. The table presents the number of genes with GO or KEGG annotation found in the list of differentially expressed genes by each method and by the combined model.

5 Discussion

The paper describes a simple method to combine several pre-processing techniques on affymetrix chips. It is easy to implement and relatively quick to run (300 secs on a DELL Precision workstation with 3.20 GHz for 1000 genes, 2 conditions, 3 pre-processing and 5 replicates). When the data are simulated from the model, combined several pre-processing methods seems a good strategy to improve the performance of the model both in terms of specificity and sensitivity. However the user should be very careful in identifying the right model. The Mixed Posterior Predictive Check on the HFD example showed that the use of a single parameter for each pre-processing was too simplistic; to overcome this issue we defined a polynomial function of the global expression. This modeling besides being more flexible is also more realistic, allowing a relation between the expression ($\bar{y}_g$) and the variance ($\sigma^2_{gk}$, which usually are assumed independent.

We want to stress the importance of a selection strategy in the process of finding the right model; in particular, Bayesian methods based on prediction permit to compare the observed data and the estimated data under the model in terms of any feature of interest. We believe the Mixed Posterior Predictive Check is a good way to carry out a model selection strategy and has the advantage of calculating a measure of discrepancy (the empirical p-value) for each gene that can be easily displayed through a histogram.

We used two types of constraints on $\lambda_j$ to ensure model identifiability: in the
model for the simulated data we used “the corner point”, fixing $\lambda_1 = 1$, while in the real dataset we used the constraint on the mean ($E[\lambda_j] = 0$). The first constraint runs quicker, but the second has the advantage of being more natural and more interpretable (in this case each $\lambda_j$ is the difference from overall mean, while in the former case each $\lambda_j$ is the ratio from the corner point).

The model is very general and could include a large number of pre-processing methods. It can be easily adapted for comparison on different type of arrays (for instance CGH or other two color arrays) and also for comparing the effect of using different platforms on the same experiment in terms of differential expression.

Acknowledgments

Marta Blangiardo and Sylvia Richardson’s work was supported by BBSRC ‘Exploiting Genomics’ grant 28EGM16093.

References

[1] Affymetrix, Statistical Algorithms Description Document (2001).
[2] D. Allison, X. Cui, G. Page, M. Sabripour, Microarray data analysis: from disarray to consolidation and consensus., Nature Reviews 7 (2006) 55–65.
[3] B. Armstrong, The effects of measurement error on relative risks regression., American Journal of Epidemiology 132 (1991) 1176–1184.
[4] P. Baldi, A. Long, A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes., Bioinformatics 17 (2001) 509–519.
[5] M. J. Bayarri, J. Berger, P-values for composite null models, Journal of the American Statistical Association 95 (2000) 1127–1142.
[6] N. Bochkina, S. Richardson, , tail posterior probability for inference in pairwise and multiclass gene expression data., Biometrics in press.
[7] B. Bolstad, R. Irizarry, M. Astrand, T. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias., Bioinformatics 19(2) (2003) 185–193.
[8] R. Carroll, Covariance analysis in generalized linear measurement error models., Statistics in Medicine 9 (1989) 1075–1093.
[9] L. Cope, R. Irizarry, H. Jaffee, Z. Wu, T. Speed, A benchmark for affymetrix genechip expression measures., Bioinformatics 20(3) (2004) 55–65.

15
[10] A. Gelman, X. Meng, H. Stern, Posterior predictive assessment of model fitness via realized discrepancies., Statistica Sinica 6 (1996) 733–807.
[11] R. Irizarry, B. Hobbs, F. Collin, Y. Beazer-Barclay, K. Antonellis, U. Scherf, T. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, Biostatistics 4 (2) (2003) 249–264.
[12] R. Koza, L. Nikonova, J. Hogan, J. Rim, T. Mendoza, C. Faulk, J. Skaf, L. Koza, Changes in gene expression foreshadow obesity in genetically identical mice, PLoS Genetics 2 (5) (2006) e81.
[13] A. Lewin, S. Richardson, C. Marshall, A. Glazier, T. Aitman, Bayesian modelling of differential gene expression., Biometrics 62 (2006) 1–9.
[14] C. Li, W. Wong, Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, Genome Biology 2 (8) (2001) 32.1.
[15] E. Marshall, D. Spiegelhalter, Approximate cross-validatory predictive checks in disease mapping models, Statistics in Medicine 22 (2003) 1649–1660.
[16] M. Masternak, A. Bartke, Ppars in calorie restricted and genetically long-lived mice, PPAR Research ID28436 (2007) 1–7.
[17] K. Ohtsubo, S. Takamatsu, M. Minowa, A. Yoshida, M. Takeuchi, J. Marth, Dietary and genetic control of glucose transporter 2 glycosylation promotes insulin secretion in suppressing diabetes., Cell 123 (7) (2005) 1307–1321.
[18] A. Recinos, B. Carr, D. Bartos, I. Boldogh, J. Carmical, L. Belalcazar, A. Brasier, Liver gene expression associated with diet and lesion development in atherosclerosis-prone mice: induction of components of alternative complement pathway, Physiological Genomics 19 (1) (2004) 131–142.
[19] S. Richardson, W. Gilks, A bayesian approach to measurement error problems in epidemiology using conditional independence models., American Journal of Epidemiology 138 (1993) 430–442.
[20] D. Thomas, W. Gauderman, R. Kerber, 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.
[21] C. Verwaerde, A. Delanoye, L. Macia, A. Tailleux, I. Wolowczuk, Influence of high-fat feeding on both naive and antigen-experienced t-cell immune response in do10.11mice, Scandinavian Journal of Immunology 64 (5) (2006) 457–466.
[22] L. Wang, B. Balas, C. Christ-Roberts, R. Yeo Kim, F. Ramos, C. Kikani, C. Li, C. Deng, S. Reyna, N. Musi, L. Dong, R. DeFronzo, F. Liu, Peripheral disruption of grb10 gene enhances insulin signaling and sensitivity in vivo, Molecular and Cellular Biology MCB.00679-07 (2007) 1–33.
[23] X. Yang, E. Schadt, S. Wang, H. Wang, A. Arnold, L. Ingram-Drake, T. Drake, A. Lusis, Tissue-specific expression and regulation of sexually dimorphic genes in mice., Genome Research 16 (8) (2006) 995–1004.