BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 6 2011, pages 807–814
doi:10.1093/bioinformatics/btr018

Gene expression

Advance Access publication January 19, 2011

Biological assessment of robust noise models in microarray data analysis

A. Posekany[1], K. Felsenstein[2] and P. Sykacek[1,*]

[1]Chair of Bioinformatics, Department of Biotechnology, University of Natural Resources and Life Sciences, Gregor Mendel Straße 33, 1180, Vienna and [2]Department of Statistics, Vienna University of Technology, Karlsplatz 13, 1040 Vienna, Austria

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Although several recently proposed analysis packages for microarray data can cope with heavy-tailed noise, many applications rely on Gaussian assumptions. Gaussian noise models foster computational efficiency. This comes, however, at the expense of increased sensitivity to outlying observations. Assessing potential insufficiencies of Gaussian noise in microarray data analysis is thus important and of general interest.

Results: We propose to this end assessing different noise models on a large number of microarray experiments. The goodness of fit of noise models is quantified by a hierarchical Bayesian analysis of variance model, which predicts normalized expression values as a mixture of a Gaussian density and *t*-distributions with adjustable degrees of freedom. Inference of differentially expressed genes is taken into consideration at a second mixing level. For attaining far reaching validity, our investigations cover a wide range of analysis platforms and experimental settings. As the most striking result, we find irrespective of the chosen preprocessing and normalization method in all experiments that a heavy-tailed noise model is a better fit than a simple Gaussian. Further investigations revealed that an appropriate choice of noise model has a considerable influence on biological interpretations drawn at the level of inferred genes and gene ontology terms. We conclude from our investigation that neglecting the over dispersed noise in microarray data can mislead scientific discovery and suggest that the convenience of Gaussian-based modelling should be replaced by non-parametric approaches or other methods that account for heavy-tailed noise.

Contact: peter.sykacek@boku.ac.at

Availability: http://bioinf.boku.ac.at/alexp/robmca.html.

Received on July 15, 2010; revised on January 5, 2011; accepted on January 6, 2011

1 INTRODUCTION

The importance of microarray data for the biological sciences has generated a large number of sophisticated analysis methods. Approaches like *t*-tests (Baldi and Long, 2001; Tusher *et al.*, 2001), linear models (Smyth, 2005) and many Bayesian methods (Bae and Mallick, 2004; Ibrahim *et al.*, 2002; Ishwaran and Rao, 2003; Lewin *et al.*, 2007; Zhao *et al.*, 2008) consider data to be approximately Gaussian distributed. Recent investigations have, however, cast doubt on the correctness of the Gaussian assumption. By testing for Gaussianity, Hardin and Wilson (2009) find that microarray data does not follow a Gaussian distribution. The observed overdispersion leads to a large number of outlying values which can have a considerable influence on the inference results. The cost of measurements and the possibility that outlying data points are caused by biological processes rule out that such samples get removed. All samples must thus be taken into account carefully, as excluding outlying values or including them based on incorrect distribution assumptions would falsify the biological findings. The adverse effects of outliers in microarray data can be overcome with non-parametric approaches (cf. de Haan *et al.*, 2009; Gao and Song, 2005; Lee *et al.*, 2005; Troyanskaya *et al.*, 2002; Tusher *et al.*, 2001; Zhao and Pan, 2003). Non-parametric methods replace the restrictive assumptions linked with the Gaussian distribution with very general ones, however, at the expense of losing some power of tests (cf. Whitley and Ball, 2002). Alternatively, we can analyze overdispersed data with robust parametric noise models like Student's-*t* distributions (cf. Gottardo *et al.*, 2006).

The issue of appropriate noise models led to an ongoing discussion, with Giles and Kipling (2003) arguing that microarray data are Gaussian distributed. Similar methods let Hardin and Wilson (2009) conclude that microarray data require heavy-tailed noise models. The conclusion of Novak *et al.* (2006) was that 5–15% of genes are non-Gaussian distributed, with the majority following Gaussian distributions. Finding such diverse conclusions about noise in microarray data suggest an in-depth investigation of this issue. We propose to this end inferring the appropriate degree of over-dispersion in microarray data with a hierarchical Bayesian model, which is inspired by the proposal of Gottardo *et al.* (2006). Built-in means for ranking genes according to differential expression enable investigations of the biological implications of deviating from the optimal noise model. The essential components of the proposed model are thus two indicator variables, one decoding whether a gene is differentially expressed, the other decoding the most appropriate noise model. These variables are built into a hierarchical Bayesian analysis of variance (ANOVA) model which can be used for analyzing a variety of experimental designs.

Inferring the proposed model with uninformative prior settings provides reliable probability measures, which quantify the suitability of competing noise models. This mode of operation compares the goodness of fit of a Gaussian noise model with *t*-distributions of different degrees of freedom and infers the appropriate robustness level required for analyzing a microarray dataset. The ultimate

[*] To whom correspondence should be addressed.

goal of microarray data analysis is, however, obtaining sound biological conclusions about which transcripts are involved in a particular process. Judgements about different noise models should therefore be linked with their implications on biological findings. The proposed model provides for this purpose a second mode of operation, in which we fix the noise model either to a Gaussian density or to a t-distribution with optimal degrees of freedom as found in the adaptive mode of operation. The biological implications of deviating from the optimal noise model can then be assessed from the noise model-dependent gene rankings.

To warrant reliable conclusions, we calibrated the model on synthetic data and the golden-spike experiment from Choe *et al.* (2005), before analysing 14 microarray datasets. Independent of normalization and preprocessing, we found in every case that a t-distribution with small degrees of freedom provides a much better fit of the noise characteristics than a Gaussian density. The importance of robust inference is apparent from our observation that exchanging the optimal Student's-t density with a Gaussian leads to between 119 and 3561 differences in gene lists and to between 14 and 316 differences in Gene Ontology (GO) (cf. Ashburner *et al.*, 2000) term lists. We have thus strong evidence that opting for Gaussian noise models in microarray data analysis may result in seriously misleading biological leads. Microarray data analysis should thus preferably use non-parametric approaches (cf. de Haan *et al.*, 2009; Gao and Song, 2005; Lee *et al.*, 2005; Troyanskaya *et al.*, 2002; Tusher *et al.*, 2001; Zhao and Pan, 2003) or approaches that allow for heavy-tailed noise models (cf. Gottardo *et al.*, 2006).

2 METHODS

The methods in this article provide a framework for thoroughly investigating whether microarray data analysis requires robust approaches, or whether we may safely rely on Gaussian assumptions. The Bayesian ANOVA model shown in Figure 1 as directed acyclic graph (DAG) infers to this end optimal robustness levels and a measure whether genes are differentially expressed. The proposed approach achieves robustness by using a parametric heavy-tailed noise model, with non-parametric methods (cf. de Haan *et al.*, 2009; Gao and Song, 2005; Lee *et al.*, 2005; Troyanskaya *et al.*, 2002; Tusher *et al.*, 2001; Zhao and Pan, 2003) being popular alternatives. To put our investigation into the context of these tools, we include the two methods by Lee *et al.* (2005) and de Haan *et al.* (2009) in our assessment. Similar to the approach in Gottardo *et al.* (2006), we propose inferring differentially expressed genes, while at the same time inferring the most appropriate noise model from a set of Student's t-distributions, which include the Gaussian as a special non-robust case. Whereas Gottardo *et al.* (2006) allow for all possible ANOVA contrasts simultaneously and infer a *per* gene posterior probability over all contrasts, our model follows the conventional strategy in microarray data analysis and infers differential expression with one common contrast.

An important aspect of our investigation is assessing the practical relevance of deciding for appropriate noise models. We propose to this end repeating inference of gene lists twice, once using the inferred noise characteristics and once using a Gaussian instead. When leaving all other settings identical, the differences in gene and GO term lists are indicative for the effect of using suboptimal noise models. Gaining far reaching validity requires analysing a representative collection of microarray datasets covering important organisms and measurement platforms and repeating assessments with different normalization and preprocessing methods.

2.1 Bayesian ANOVA with flexible noise model

The Bayesian one-way ANOVA model shown in Figure 1 as DAG constitutes the core of our evaluation. ANOVA models are commonly used for analysing multi-level microarray experiments like time course data. The model is based on a linear relation between the gene expression $y_{n,g}$, measured for sample n and gene g and the mean expression β_g. The S-dimensional vector x_n is an indicator for the biological state. If sample n belongs to state s, x_n has a 1 at the s-th position and zeros everywhere else. Depending on whether gene g is differentially expressed or not, the latent indicator variable I_g switches between two different dimensional representations of β_g (cf. Holmes and Held, 2006; Sykacek *et al.*, 2007). The case that gene g is not differentially expressed is coded by $I_g = 0$ and corresponds to the null hypothesis of the classical ANOVA that all groups have the same mean. Vector β_g contains in this case S identical entries of mean expression $\beta_{g,0}$, the latter being equipped with a Gaussian prior with mean μ and prior precision λ. The alternative hypothesis that gene g is differentially expressed is coded by $I_g = 1$ with β_g being multivariate with a Gaussian prior with mean μ and diagonal precision matrix λ. The indicator I_g is *a priori* binomially distributed with probability p of differential expression. To reduce the sensitivity of the approach, the model is extended hierarchically by allowing for a beta prior over p with hyperparameters a and b. The observations $y_{n,g}$ follow a symmetric distribution centred around $\hat{y}_{n,g} = x_n^T \beta_g$ with precision $\varphi_{n,g}\tau$. Different robustness levels are achieved by selecting the noise model for the observations $y_{n,g}$ from a set containing $K-1$ Student's t-distributions of different degrees of freedom, ν, and a Gaussian distribution (with $\nu = \infty$). For obtaining computationally tractable representations of Student's t-distributions with arbitrary degrees of freedom, we introduce the auxiliary variables $\varphi_{n,g}$, represent $p(y_{n,g}, \varphi_{n,g} | \beta_g, \tau, \nu)$ as a certain Gaussian-Gamma density and integrate over $\varphi_{n,g}$ (cf. Bernardo and Smith, 1994).

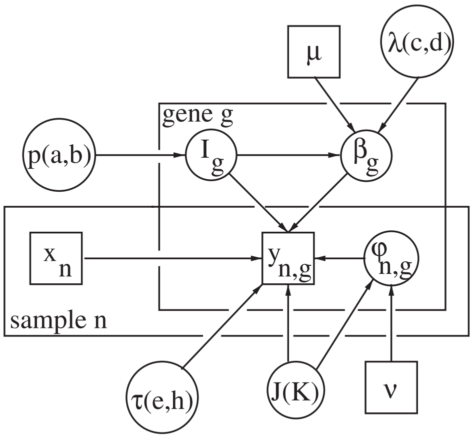

Fig. 1. We represent the proposed model as DAG with rectangular nodes denoting observed quantities and circular nodes denoting random variables. Hyperparameters associated with priors are shown in brackets. Sheets indicate replication. With n denoting the sample and g the gene index, we denote the measurements as $y_{n,g}$, the group indicators as x_n, the group specific means as β_g, the differential expression indicators as I_g, their prior probability as p, the noise precision as τ, the precision of the coefficients prior as λ, the auxiliary variables of the Student's-t density as $\varphi_{n,g}$, the degrees of freedom as ν and the corresponding model indicator as J. All hyperparameters are discussed in Section 2.1.

An essential aspect of robustness is adjusting the degrees of freedom ν to the level required by the data. We propose to this end selecting the best fitting degrees of freedom from a finite set of possible choices (cf. Berger, 1994), which includes the Gaussian ($\nu = \infty$) as the non-robust special case. The proposed model implements this selection via the multinomial-one distributed indicator variable J, which chooses a particular ν from the set $\nu := \{\nu \in \mathbb{R} | \nu := \nu_{\min} + j \cdot c_{\text{grid}}, \infty\}$[1] with $j \in [0,..,K-2]$ and $\nu_{\min} \geq 1$. This

[1] To simplify notation ν denotes the set and individual values of the degrees of freedom parameter.

formulation gives rise to K possible noise models. As we have no reason for preferring a particular choice, we use $1/K$ as uninformative prior probability for all J. The proposed model can be summarized by the joint density formulated in Equation (1)

$$p(I,p,\beta,\varphi,J,\lambda,\tau|X,Y,a,b,c,d,e,h,K) \propto p(p|a,b) \quad (1)$$
$$p(\lambda|c,d)p(\tau|g,h)p(J|K)\prod_g \Big(p(I_g|p)p(\beta_g|I_g,\mu,\lambda)$$
$$\prod_n \big(p(\varphi_{n,g}|\nu)p(y_{n,g}|\beta_g,\varphi_{n,g},I_g,\tau,x_n)\big)\Big),$$

where I, β, φ, X and Y are shortcuts for denoting all I_g, β_g, $\varphi_{n,g}$, x_n and $y_{n,g}$, respectively, and $p(p|a,b)$ denotes a Beta density, $p(\lambda|c,d)$, $p(\tau|g,h)$ and $p(\varphi_{n,g}|\nu/2,\nu/2)$ denote Gamma densities, $p(J|K)$ denotes a Multinomial-one density, $p(I_g|p)$ a Binomial-one density and $p(\beta_g|I_g,\mu,\lambda)$ and $p(y_{n,g}|\beta_g,\varphi_{n,g},I_g,\tau,x_n)$ denote Gaussian densities.

2.2 Algorithm

The complexity of the model requires approximate inference. Although closed form approximations (Liu *et al.*, 2006; Sykacek *et al.*, 2007) have computational advantages, we prefer here an unbiased approximation and follow (Bae and Mallick, 2004; Gottardo *et al.*, 2006; Huang *et al.*, 2002; Lewin *et al.*, 2007; Shahbaba and Neal, 2006; Tadesse *et al.*, 2003) who, among many others, have previously used Markov chain Monte Carlo (MCMC) in a bioinformatics context. MCMC is an application of the Law of Large Numbers and allows approximating expectations by averages of random draws from a given distribution. The random samples are realizations of a Markov chain that behave under certain conditions like draws from a single stationary distribution (cf. Gilks *et al.*, 1996; Robert and Casella, 2004). Denoting the sampling density as f and the random samples obtained from MCMC as ${\beta_g}^{(i)}$, MCMC allows us for example to approximate the expectation of the group-specific mean expression β_g as

$$\mathbb{E}_f[\beta_g] = \int_{\mathcal{X}} \beta_g f(\beta_g) d\beta_g \approx \overline{\beta_g} = \frac{1}{n}\sum_{i=1}^{n} {\beta_g}^{(i)}.$$

Algorithm 1 illustrates MCMC sampling as pseudo-code. Inference requires a combination of Gibbs, Metropolis Hastings and Reversible Jump steps. Gibbs steps are used for updating the prior probability of differential expression, p, the prior precision λ, the error precision, τ, and, when keeping the differential expression indicator I_g fixed, for updating the group means β_g. A Metropolis Hastings step is used for updating J as long as we keep the Student's-t noise model. Updates of J that propose changing from a Student's-t to a Gaussian density and vice versa and updates of I_g rely on the reversible jump approach introduced in Green (1995). Further details about the model, the algorithm and a MatLab implementation are provided in http://bioinf.boku.ac.at/alexp/robmca.html.

Algorithm 1 Hybrid MCMC Sampler

```
Random initialization of parameters
$c_{\text{grid}} = 1$
for $n = 1$ to burnin do
  update parameters={
    update $\nu$, $J$ and $\varphi_{n,g}$ jointly
    update $p$
    update $\lambda$
    update $\beta_g$ and $I_g$ jointly
    update $\tau$ }
end for
$c_{\text{grid}} = 0.05$
for $n = 1$ to burnin+simulationlength do
  update parameters (see first burn-in)
end for
```

2.3 Data collection

For reliably inferring the optimal noise characteristics and evaluating the implications of potentially oversimplified Gaussian assumptions, we have to consider two aspects. A reliable assessment of different noise models requires calibrating the proposed inference scheme. Calibration makes sure that MCMC converges rapidly and that inference result are insensitive to the chosen hyperparameters. These aspects are best assessed when knowing the expected outcome by using synthetically generated data and dedicated spike-in experiments. Warranting that our findings are generally applicable requires analysing carefully selected microarray datasets, which cover a wide range of model organisms, experimental settings and measurement platforms and using several normalization and preprocessing methods.

Table 1. Depending on sample type which is either 1 or 2, genes from subset i are drawn from distributions with means equal to $\mu_{i,1}$ and $\mu_{i,2}$, respectively

Subset i	$\mu_{i,1}$	$\mu_{i,2}$	%
1	−12.0	12.0	20
2	−5.0	5.0	10
3	−1.0	1.0	30
4	−0.5	0.5	20
5	0.0	0.0	20

The proportion of genes in subset i is shown in column %.

Artificial data were generated with Gaussian and Student's-t noise distributions, the latter with 4 and 10 degrees of freedom. We simulated a two way comparison of 500 hypothetical genes with each gene assigned to one of five groups, the later defining the amount of hypothetical differential expression. The mean structure and fraction of occurrence of each group are reported in Table 1. Variances have been chosen in the range of 0.1–10, without altering the reported results. To mimic a realistic microarray scenario, we generated five replicates per group, resulting in 10 data points per gene. Some aspects of computer-generated data might deviate from real microarray measurements. We endorse therefore our respective conclusions by including the spike-in experiment of Choe *et al.* (2005) in our analysis.

For warranting far reaching validity of our results, we analysed 14 microarray experiments covering various organisms and measurement platforms. The data include investigations of plant soil responses, drosophila sleep deprivation, primate dietary comparisons and animal liver metabolism. The experiments, which are summarized in Table 2, are identified by the Gene Expression Omnibus (GEO) reference number (cf. Edgar *et al.*, 2002). Further details about each dataset can be found in the corresponding reference. The selection provided in Table 2 covers several different platforms and quantification algorithms (cf. column 'Prep.'). We used all data as provided by the owner and applied the conservative normalization method vsn (cf. Huber *et al.*, 2002).

2.4 Alternative normalization and analysis methods

It is well known that results from microarray data analysis may depend on the chosen normalization method (cf. Bolstad *et al.*, 2003). To ensure that our conclusions hold in general, we repeated the analysis for a subset of the data in Table 2 with additional normalization methods. Guided by their popularity in applied microarray papers, we chose loess (cf. Yang *et al.*, 2002) and quantile (cf. Bolstad *et al.*, 2003) normalization.

In the light of recent findings that intensities of highly expressed targets cross-talk to neighbouring probes due to scanner inadequacy (cf. Upton and Harrisson, 2010), we may expect that Affymetrix probe sets contain outlying measurements. Being designed to alleviate the effect of artefacts contaminating individual probes, the mmgMOS approach (cf. Liu *et al.*, 2005) and the PPLR method (cf. Liu *et al.*, 2006) could help improving

Table 2. Overview of the biological datasets describing the organism (Org.), the GEO ID (CAMDA 08 refers to the Endothelial Apoptosis contest datasets of the meeting), the preprocessing method (Prep.), the overall number of arrays (N), the average degrees of freedom ($\overline{\nu}$), the number of common genes (Comm.), the number of genes with noise model depending differential expression assessment (Diff.), the number of common GO terms (Comm.) and finally the number of noise model dependent GO terms (Diff.)

Org.	GEO ID	Reference	Prep.	N	$\overline{\nu}$	Comm. genes	Diff. genes	Comm. GO terms	Diff. GO terms
Arabidopsis thaliana	GDS3216	Dinneny *et al.* (2008)	MAS5.0	12	4.71	1176	150	111	78
Arabidopsis thaliana	GDS3225	Van Hoewyk *et al.* (2008)	MAS5.0	4	5.50	832	290	161	21
Danio rerio	GDS1404	Cameron *et al.* (2005)	PathStat	10	13.58	1776	136	11	14
Drosophila melanogaster	GDS1686 (I)	Zimmerman *et al.* (2006)	RMA	9	3.62	136	174	11	96
Homo sapiens	CAMDA 08	Affara *et al.* (2007)	CLSS4.1	24	4.04	400	304	26	67
Homo sapiens	GDS1375	Talantov *et al.* (2005)	MAS5.0	70	3.25	6861	3561	160	316
Homo sapiens	GDS810	Blalock *et al.* (2004)	MAS5.0	31	4.37	72	135	9	51
Homo sapiens	GDS2960	Yao *et al.* (2007)	RGP3.0	101	4.33	318	166	51	2
Mus musculus	GDS660	Small *et al.* (2005)	MAS5.0	22	10.48	584	126	20	26
Mus musculus	GDS3221	Somel *et al.* (2008)	RMA	24	4.21	180	119	108	52
Mus musculus	GDS3162	Someya *et al.* (2008)	MAS5.0	10	4.38	797	446	112	66
Mus musculus	GDS1555	MacLennan *et al.* (2006)	MAS5.0	8	3.90	131	183	24	110
Rattus norvegicus	GDS2946	Li *et al.* (2008)	MAS5.0	15	4.57	146	157	14	306
Rattus norvegicus	GDS972	Jin *et al.* (2003)	MAS5.0	44	4.98	369	163	94	71
Drosophila melanogaster	golden-spike	Choe *et al.* (2005)	MAS5.0	6	3.74	401	1748	–	–

The GEO entry GDS1686 (I) refers to the behavioural subset of the data (only the sleep-deprived flies). In column Prep., we use MAS5.0 to refer to the Affymetrix MAS 5.0 quantization method, RMA to refer to the 'Robust Multi-array Average' method by Irizarry *et al.* (2003) (both used for Affymetrix arrays), PathStat for referring to the package described in Middleton *et al.* (2004), CLSS4.1 to refer to the Codelink Software Suite 4.1 and RGP3.0 to refer to Research Genetics' Pathway software v. 3.0.

the Gaussianity of residuals. For testing whether such sophisticated representations of microarray expression can reduce the need for heavy-tailed noise models, we applied our algorithm to mmgMOS normalized data and the posterior expression estimates obtained by the PPLR method.

Our investigation relied so far on achieving robustness by representing the noise in microarray data with a suitably chosen parametric density. A different strategy for achieving robustness in microarray data analysis is obtained by abolishing distributional assumptions and using non-parametric methods (cf. de Haan *et al.*, 2009; Gao and Song, 2005; Lee *et al.*, 2005; Tusher *et al.*, 2001). To investigate whether non-parametric approaches are a viable alternative for robust analyses of microarray data, we compare gene rankings obtained with such approaches with gene rankings we obtain (i) with the Bayesian ANOVA when using the optimal (possibly heavy tailed) noise distribution and (ii) with the proposed model when assuming Gaussian distributed noise. Compatibility with our ANOVA model suggests applying a Kruskal–Wallis permutation test (cf. Lee *et al.*, 2005) and a robust ANOVA (cf. de Haan *et al.*, 2009). Unknown differences in scale which we have to expect when comparing P-values and Bayesian probabilities are overcome by using a P-value threshold of 0.01 for assigning differential expression in the statistical test and adjusting the probability threshold such that the number of differentially expressed genes match.

2.5 Biological implications

An important aspect in our assessment of different noise models for microarray data analysis is evaluating the biological implications of deviations from the appropriate noise model. The implication of choosing Gaussian noise instead of the optimal noise model can be quantified by comparing the number of genes, which are irrespective of the noise model assessed as differentially expressed with the number of genes which show a noise model-dependent assessment. For investigating the implications of unsuitable noise models at a higher level of biological abstraction, we propose inferring GO terms from the gene lists which we obtain with different noise models. We use to this end, GO term-specific Fishers exact tests (cf. Al-Shahrour *et al.*, 2004; Dennis. *et al.*, 2003) on the gene lists obtained with different noise models and compare the number of significant GO terms which are found irrespective of the chosen noise model with the number of GO terms with noise model-dependent assessment.

2.6 Calibrating the algorithm

Calibration efforts are important for assuring unbiased and efficient inference with MCMC methods. Making sure that inference is unbiased requires considering the influence of all hyperparameters individually. We have a and b which are prior counts and thus easy to grasp with small values corresponding to small influence. A Jeffreys prior (cf. Jeffreys, 1961) is obtained when using $a\!=\!b\!=\!0.5$. The hyperparameters g and h of the Gamma prior over the noise precision τ also have no indirect consequences and can safely be set to 0 for obtaining the corresponding Jeffreys prior. Independent of whether we use a t-distribution or a Gaussian as noise model, the precision λ deserves more attention. Large values of λ indicate a strong preference for small β_g values. By entering the Bayes factors of the models represented by $I_g=0$ and $I_g=1$, the precision λ influences, however, also $P(I_g|X,Y,a,b,c,d,e,h,K)$ (cf. MacKay, 1992), with smaller λ making identifying differentially expressed genes harder. This problem can be solved by regarding λ as a random variable and providing a conjugate Jeffreys hyper-prior which is a Gamma density parameterized with $c=0,d=0$. Such hierarchical Bayesian models (cf. Lewin *et al.*, 2007; Shahbaba and Neal, 2006) are preferably used, because an indirect prior specification minimizes the dependency of inference results on hyper parameter settings. Jeffreys priors are theoretically well motivated in single variable cases, they can, however, exhibit strong indirect influence in multi-variable models (cf. Bernardo and Smith, 1994).

Having an indirect influence on decisions about differential expression, the precision λ deserves particular attention. We, therefore, propose investigating the influence of the hyperparameters c and d on the posterior probabilities of differential expression $P(I_g|X,Y,a,b,c,d,e,h,K)$. By representing the precision λ in the Gaussian prior over β_g as a random variable, the influence of c and d on λ is related to the prior variance $V[\lambda]_{p(\lambda|c,d)}=\frac{c}{d^2}$. A sensitivity analysis can therefore keep the prior expectation $E[\lambda]_{p(\lambda|c,d)}=\frac{c}{d}$ constant (e.g. 1, for other values please refer to the Supplementary Material) and change the prior variance. By displaying the ordered posterior probabilities, $P(I_g|X,Y,a,b,c,d,e,h,K)$, for several c, d combinations, the graphs in

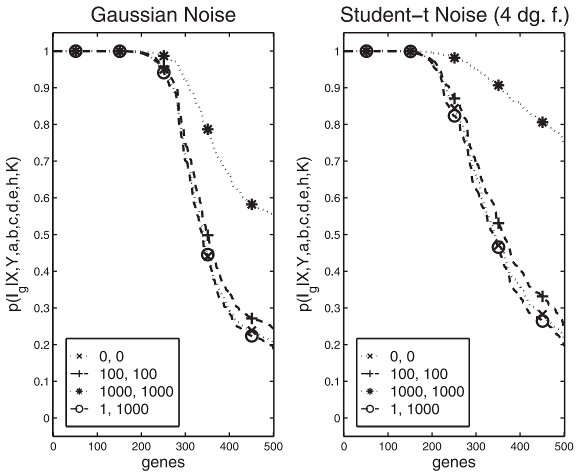

Fig. 2. The hyperparameters c and d in the prior $p(\lambda|c,d)$ have to be chosen carefully to avoid side effects. The graphs show the ordered posterior probabilities of differential expression $P(I_g|X,Y,a,b,c,d,e,h,K)$ with the legend denoting the corresponding c,d pair for a Gaussian and Student's-t noise model. Choices larger than around 100 increase the influence of these hyperparameters on the posterior of interest. This motivates our choice of an improper prior $(c=0,d=0)$.

Figure 2 illustrate this sensitivity analysis for synthetic data that was generated according to the description in Section 2.3. Choosing the Jeffreys prior $c=0,d=0$ is justified by observing that up to a prior variance of less than 1/100, the hyper-parameters c and d have, independently of the noise model, little influence on the posterior probabilities of differential expression.

Another important aspect of our inference scheme is providing accurate assessments of noise characteristics. This requires a clear distinction between Gaussian and Student's-t noise models and thus an appropriate choice for the upper limit for the degrees of freedom parameter ν, which marks the bound between Student's-t and Gaussian distributions. Simulations on synthetic data showed that taking $\nu_{max}=45$ as upper limit is a good choice because larger degrees of freedom parameters render Student's-t densities as indistinguishable from Gaussians and smaller values would unnecessarily misjudge Student's-t densities as Gaussians.

Further calibration efforts were concerned with assuring fast convergence of the sampling algorithm to the stationary distribution. Our simulations showed that convergence speed can be dramatically improved by adjusting the grid size c_{grid} between two burn-in phases. After starting with an initial value in the range of 1–5, we switch to a smaller value of about 0.05 which is then also used for sampling. A large initial grid size allows the algorithm to quickly determine the approximately correct error model with the smaller grid size improving the convergence properties of the Markov Chain and leading to better approximations of the true continuous degrees of freedom. Convergence towards the stationary distribution was assessed with the R package coda (cf. Plummer *et al.*, 2006). We found that 11000 draws were a suitable overall simulation length and that the first 500 draws should be considered as burn-in phase (cf. Algorithm 1).

After calibration, we could confirm that the resulting algorithm infers the correct noise model in synthetic data. Data generated with Student's-t distributed noise with 4 and 10 degrees of freedom lead to little variation of the samples around the true value, whereas data generated with Gaussian distributed noise would assign all mass to the Gaussian density. We also tested whether the proposed algorithm infers differentially expressed genes reliably. We used for that purpose the golden-spike experiment from Choe *et al.* (2005). Resulting from a wet lab experiment, these data are both a realistic test case for microarray data and a gold standard with known ground truth. When using a cutoff probability threshold of 0.85, we find for Gaussian noise 72% and for the optimal Student's-t noise 78% of correctly assigned genes. These performance figures are in the top range of the results reported in Choe *et al.* (2005). The better performance of the Student's-t model is paired with a by far larger evidence in favour of this noise model. This observation allows the conclusion that already the technical noise component in microarray data, which is the only remaining source of variation in the golden-spike data, requires considering robust models.

Table 3. An assessment of robustness levels in dependence of normalization and preprocessing showing the expected degrees of freedom parameters

GEO ID	Loess	Quantile	mmgMOS	PPLR
GDS3216	2.02	1.13	2.23	1.17
GDS810	1.13	1.18	3.23	1.14
GDS3225	1.24	1.29	–	–
CAMDA 08	1.06	1.11	–	–
GDS1375	1.14	1.15	–	–
GDS2960	2.94	2.85	–	–
GDS1555	1.15	1.17	–	–
GDS972	1.38	1.4	3.67	1.15

$\overline{\nu}$ for various datasets. Dashes indicate unavailable results, which require for mmgMOS and PPLR to have Affymetrix cell files available. The results confirm that neither normalization nor sophisticated preprocessing compensate for the need of heavy-tailed noise models.

3 RESULTS

To highlight the importance of choosing valid noise models for microarray analysis, we applied the proposed inference scheme to 14 microarray datasets, which are summarized in Table 2. The arresting result of our evaluation is that a heavy-tailed Student's-t noise model is a better fit than a Gaussian noise model for every dataset we looked at (cf. column '$\overline{\nu}$'). For most datasets, a t-distribution with degrees of freedom between 3 and 5 got the highest posterior probability. This indicates the need of robust noise models, which can handle outlying data points well and suggests that Gaussian noise models are unsuitable for microarray data analysis, even if according to Novak *et al.* (2006) only about 5–15% of samples are non-Gaussian distributed.

Our assessments also revealed that biological inference depends considerably on the chosen noise model. For obtaining a quantitative statement, we inferred the differentially expressed genes set twice: once with a Gaussian noise model and once with the optimally inferred t-distribution. This approach provided for every dataset two gene lists with the intersect representing agreement and the symmetric difference representing different biological interpretations, which are solely caused by the different noise models (cf. Table 2, columns 'Comm. genes' and 'Diff. genes').

Microarray data analysis depends often critically on chosen preprocessing and normalization (cf. Bolstad *et al.*, 2003). To rule out being mislead by a particular choice, we repeated the assessment of optimal noise models using loess and quantile normalization and mmgMos and PPLR preprocessing. The expected degrees of freedom, $\overline{\nu}$ we report in Table 3 for these data allow concluding that our observations are independent of normalization and even sophisticated analysis methods do not compensate for the need of robust noise models. The robust model is in general less sensitive to outlying values. Models with t-distributed noise will therefore assign lower posterior probabilities of differential expression,

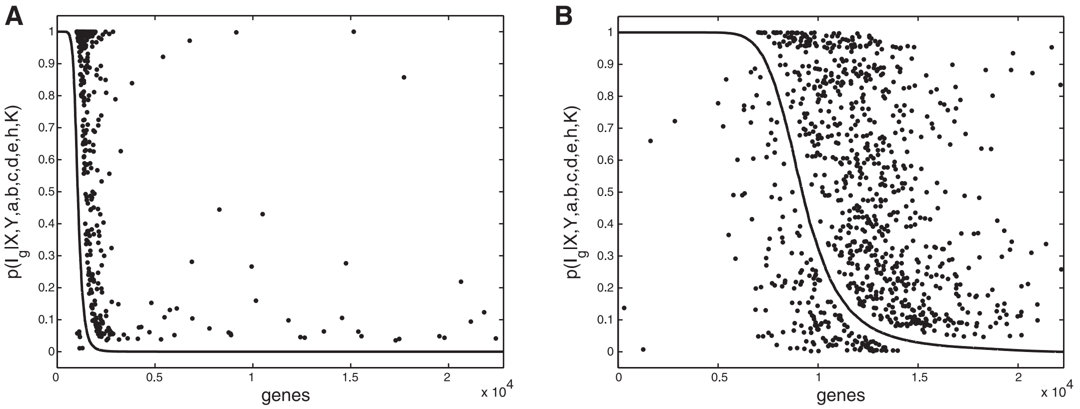

Fig. 3. Noise model dependencies of posterior probabilities. Subplot (**A**) illustrates the Arabidopsis data (GDS3216) ranked by the posterior probabilities of differential expression obtained with the most probable Student's-t distribution (probabilities shown as black line). The probabilities obtained for the same genes by a Gaussian-based analysis are shown as dots. Subplot (**B**) illustrates the Human melanoma data (GDS1375) ranked by the posterior probabilities of differential expression obtained by a Gaussian-based analysis (probabilities shown as black line). The probabilities obtained for the same genes from an optimally adjusted robust analysis are shown as dots.

when differential expression is caused by one or a few outlying measurements. In situations where outliers lead to an increase of variance or a decrease in average differential expression, the Gaussian noise model will overlook differentially expressed genes, which would be captured by the more appropriate t-distributed noise model. A wrongly chosen noise model will therefore lead to false positives and false negatives. Both error types are confirmed by the illustrations in Figure 3, which show many genes with noise model-dependent probabilities of differential expression.

Subgraph (A) in Figure 3 is ranked by the posterior probabilities obtained with the optimal t-distributed noise model (probabilities shown as black line). A subset of posteriors obtained with Gaussian noise is shown as dots. Subgraph (B) is ranked by the posterior probabilities obtained with a Gaussian noise model (probabilities shown as black line). A subset of the posteriors obtained with optimal Student's-t noise is shown as dots. We find in both subgraphs for many genes a substantial influence of the noise model on the posterior probability of differential expression. In the context that inference over degrees of freedom ν clearly favoured the Student's-t model, we can consider all genes which get only under t-distributed noise a large posterior probability of differential expression as potential false negatives under a standard Gaussian noise model. Genes that get only under a Gaussian density a large posterior probability of differential expression are likely to be false positives. Table 2 shows that the number of genes with noise model-dependent assessment of differential expression range from 119 to 3561. This is about one tenth to two times the number of genes, which are independently of the noise model assessed as differentially expressed. We can thus conclude that the choice of noise model can have a considerable influence on the inferred gene lists with a wrongly chosen noise model introducing both false positives and false negatives.

To investigate the biological significance of the noise model-dependent differences in gene lists, we applied a GO term inference (cf. Al-Shahrour *et al.*, 2004) twice: once using the gene list which we obtained with the Gaussian noise model and a second time using the gene list which we obtained when the noise is fixed to the most probable t-distribution. Table 2 lists the number of GO terms, which were found unambiguously and the number of GO terms with a noise model-dependent assessment (cf. columns 'Comm. GO terms' and 'Diff. GO terms'). Observing that the noise model-dependent GO term lists contain between one fifth and 22 times as many differences than common entries suggests that an unsuitably chosen noise model is likely to have a profound implication on biological conclusions drawn from an analysis.

Table 4. For comparing non-parametric robust methods with robust parametric methods, we provide the percentage agreement about differentially expressed genes

GEO ID	KW perm.		RANOVA	
	$\mathcal{T}$ (%)	$\mathcal{N}$ (%)	$\mathcal{T}$ (%)	$\mathcal{N}$ (%)
GDS3216	39	37	–	–
GDS1375	86	84	86	83
GDS2960	76	71	76	72

The columns under KW perm. illustrate the agreements of the Kruskal–Wallis permutation test with the robust parametric method (column '$\mathcal{T}$') and with a Gaussian-based analysis (column '$\mathcal{N}$'). The two columns under RANOVA show the same information for the robust ANOVA method. Dashes indicate that the non-parametric method did not find differentially expressed genes. These results allow the conclusion that non-parametric methods are viable for analysing microarray data robustly, as long as we have sufficiently many samples.

Having gathered substantial evidence that microarray data should be analysed by considering heavy-tailed noise, the question arises whether non-parametric approaches can help solving this issue. We compare to this end the agreement in gene lists obtained with two non-parametric tests with our robust Bayesian ANOVA and compare this with the agreement we observe between the same tests and the Gaussian version of our Bayesian ANOVA. The results in Table 4 show a better agreement of rankings between the robust methods, which suggests that non-parametric methods should be considered for analysing microarray data. Our results do, however, in agreement with Whitley and Ball (2002)) also reveal the loss in power inherent to non-parametric methods. In our analysis this manifests in finding no significant P-values with the robust ANOVA method for GEO ID GDS3216. For GEO ID GDS3225, GEO ID GDS1555 and the CAMDA 08 data, both non-parametric methods fail in finding significant P-values (data omitted from table). From Table 4, it is also obvious that small sample sizes (cf. Table 2, column 'N') lead in general to poor agreement. If sample sizes permit application, we can however recommend non-parametric methods for microarray data analysis.

4 DISCUSSION

This article provides an in-depth assessment of two competing assumptions about the noise characteristics in microarray data. Assuming Gaussian noise has the benefit of leading to highly efficient analysis methods. A considerable sensitivity to outlying observations is, however, an unfortunate weakness of Gaussian noise-based data analysis. This weakness may be overcome with non-parametric methods or by methods which assume heavy-tailed noise distributions. Applying robust analysis methods to microarray data has the disadvantage of introducing more involved computations. The application of non-parametric methods is in addition limited to problems with sufficiently many samples.

Comparing robust analysis methods with Gaussian-based microarray data analysis has to provide conclusions, which are relevant for biological practise. Certain technical aspects can be tested by gold standards like the spike-in data from Choe *et al.* (2005). Other aspects like, for example, biological variation are only captured by data analysing real-world biology. Although certain facts about individual experiments are well known, complete knowledge of ground truth is not available for any biological microarray experiment. An assessment of biological implications has thus to resort to indirect strategies. The route chosen in this article first compares the technical suitability of Gaussian noise and heavy tailed t-distributions. This requires a mode of operation in data analysis, which allows comparing different noise models. Once we established which noise model is preferred for technical reasons, we can turn to investigating the biological implications caused by changing the noise model. This mode of operation relies in our analysis on counting the number of genes which show a noise model-dependent difference in differential expression. These gene counts are complemented by investigating which GO terms are significantly affected from the noise model-dependent gene lists. For providing conclusions of far-reaching validity, we analysed 14 carefully chosen microarray experiments, covering a wide range of model organisms and measurement platforms. To avoid reporting spurious results, our simulations included careful tuning of hyperparameters to minimize model sensitivity, steps for assessing convergence of the algorithm and applied different normalization and preprocessing methods.

The arresting result of our assessment is that we find highly decisive evidence in favour of t-distributions with high kurtosis for every experiment we looked at. The significance of this finding is backed up by the observation that the choice of error model considerably influences the biological conclusions drawn from the analyses. Gene lists differ in dependence of the noise model by between 119 and 3561 genes. These differences have a substantial influence on the conclusions we draw on a higher level of biological abstraction. The number of differences in the GO term lists we find in dependence of the chosen noise model ranges from 14 to 316. For many datasets, the number of GO terms with noise model-dependent equivocal assessment is larger than the number of GO terms we can unambiguously assign to these experiments irrespective of the chosen noise model. We may thus conclude that a substantial number of outlying measurements is present in many microarray studies. Relying on implicit Gaussian assumptions means ignoring the heavy tails of the residuals and that can have adverse effects on biological conclusions drawn from microarray data. Practitioners should thus apply robust approaches for microarray data analysis, which work reliably irrespective of whether noise is Gaussian or heavy tailed. We suggest for this purpouse considering non-parametric approaches (cf. de Haan *et al.*, 2009; Gao and Song, 2005; Lee *et al.*, 2005; Troyanskaya *et al.*, 2002; Tusher *et al.*, 2001; Zhao and Pan, 2003), or, for small sample sizes, apply Bayesian approaches like Gottardo *et al.* (2006) or the MatLab implementation which accompanies this paper at http://bioinf.boku.ac.at/alexp/robmca.html.

ACKNOWLEDGEMENTS

The authors would like to thank D. Kreil for his input on microarray normalization. We are also grateful to our anonymous reviewers who helped improving our article with their valuable comments.

Funding: A. Posekany and P. Sykacek are grateful to the Vienna Science and Technology Fund for funding this work within the project WWTF-LSC-2005-#35.

Conflict of Interest: none declared.

REFERENCES

Affara,M. *et al.* (2007) Understanding endothelial cell apoptosis: what can the transcriptome, glycome and proteome reveal? *Philos. Trans. R. Soc. B*, **362**, 1469–1487.

Al-Shahrour,F. *et al.* (2004) Fatigo: a web tool for finding significant association of gene ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Bae,K. and Mallick,B. (2004) Gene selection using a two-level hierarchical bayesian model. *Bioinformatics*, **20**, 3423–3430.

Baldi,P. and Long,A. (2001) A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, **17**, 509–519.

Berger,J.O. (1994) An overview of robust Bayesian analysis. *Test*, **3**, 5–124.

Bernardo,J. and Smith,A. (1994) *Bayesian Theory*. Wiley, Chichester.

Blalock,E. *et al.* (2004) Incipient alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. *Proc. Natl Acad. Sci.*, **101**, 2173–2178.

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **19**, 185–193.

Cameron,D. *et al.* (2005) Gene expression profiles of intact and regenerating zebrafish retina. *Mol. Vis.*, **11**, 775–791.

Choe,S. *et al.* (2005) Preferred analysis methods for affymetrix genechips revealed by a wholly defined control dataset. *Genome Biol.*, **6**, R16.

de Haan,J. *et al.* (2009) Robust anova for microarray data. *Chemometr. Intell. Lab. Syst.*, **98**, 38–44.

Dennis.,G. *et al.* (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, R60.

Dinneny,J. *et al.* (2008) Cell identity mediates the response of Arabidopsis roots to abiotic stress. *Science*, **320**, 942–945.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acid Res.*, **30**, 207–210.

Gao,X. and Song,P. (2005) Nonparametric tests for differential gene expression and interaction effects in multi-factorial microarray experiments. *BMC Bioinformatics*, **6**, 186.

Giles,P. and Kipling,D. (2003) Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, **19**, 2254–2262.

Gilks,W. *et al.* (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gottardo,R. *et al.* (2006) Bayesian robust inference for differential gene expression in microarrays with multiple samples. *Biometrics*, **62**, 10–18.

Green,P.J. (1995) Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Hardin,J. and Wilson,J. (2009) A note on oligonucleotide expression values not being normally distributed. *Biostatistics*, **10**, 446–450.

Holmes,C. and Held,L. (2006) Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Anal.*, **1**, 145–168.

Huang,E. *et al.* (2002) Gene expression profiling for prediction of clinical characteristics of breast cancer. *Hormone Res.*, **58**, 55–73.

Huber,W. *et al.* (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformaics*, **18** (Suppl. 1), S96–S104.

Ibrahim,J. *et al.* (2002) Bayesian models for gene expression with dna microarray data. *J. Am. Stat. Assoc.*, **97**, 88–99.

Irizarry,R. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **31**, 249–264.

Ishwaran,H. and Rao,J. (2003) Detecting differentially expressed gene in microarrays using Bayesian model selection. *J. Am. Stat. Assoc.*, **98**, 438–455.

Jeffreys,H. (1961) *Theory of Probability*, 3rd edn. Clarendon Press, Oxford.

Jin,J. *et al.* (2003) Modeling of corticosteroid pharmacogenomics in rat liver using gene microarrays. *J. Pharmalcol. Exp. Ther.*, **307**, 93–109.

Lee,M. *et al.* (2005) Nonparametric methods for microarray data based on exchangeability and borrowed power. *J. Biopharm. Stat.*, **15**, 783–797.

Lewin,A. *et al.* (2007) Fully Bayesian mixture model for differential gene expression: simulations and model checks. *Stat. Appl. Genet. Mol. Biol.*, **6**, doi:10.2202/1544-6115.1314.

Li,S. *et al.* (2008) Assessment of diet-induced obese rats as an obesity model by comparative functional genomics. *Obesity* , **16**, 811–818.

Liu,X. *et al.* (2005) A tractable probabilistic model for affymetrix probe-level analysis across multiple chips. *Bioinformatics*, **21**, 3637–3644.

Liu,X. *et al.* (2006) Probe-level measurement error improves accuracy in detecting differential gene expression. *Bioinformatics*, **22**, 2107–2113.

MacKay,D.J.C. (1992) Bayesian interpolation. *Neural Comput.*, **4**, 415–447.

MacLennan,N. *et al.* (2006) Targeted disruption of glycerol kinase gene in mice: expression analysis in liver shows alterations in network partners related to glycerol kinase activity. *Hum. Mol. Genet.*, **15**, 405–415.

Middleton,F. *et al.* (2004) Application of genomic technologies: DNA microarrays and metabolic profiling of obesity in the hypothalamus and in subcutaneous fat. *Nutrition*, **20**, 14–25.

Novak,J. *et al.* (2006) Generalization of DNA microarray dispersion properties: microarray equivalent of t-distribution. *Biol. Direct*, **1**, 27.

Plummer,M. *et al.* (2006) CODA: convergence diagnosis and output analysis for MCMC. *R. News*, **6**, 7–11.

Robert,C.P. and Casella,R. (2004) *Monte Carlo Statistical Methods*. Springer, New York.

Shahbaba,B. and Neal,R.M. (2006) Gene function classification using Bayesian models with hierarchy-based priors. *BMC Bioinformatics*, **7**, 448.

Small,C. *et al.* (2005) Profiling gene expression during the differentiation and development of the murine embryonic gonad. *Biol. Reprod.*, **72**, 492–501.

Smyth,G.K. (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and BioConductor* . Springer, New York, pp. 397–420.

Somel,M. *et al.* (2008) Human and chimpanzee gene expression differences replicated in mice fed different diets. *PLoS One*, **3**, e1504.

Someya,S. *et al.* (2008) The role of mtdna mutations in the pathogenesis of age-related hearing loss in mice carrying a mutator dna polymerase gamma. *Neurobiol. Aging*, **29**, 1080–1092.

Sykacek,P. *et al.* (2007) Bayesian modelling of shared gene function. *Bioinformatics*, **23**, 1936–1944.

Tadesse,M. *et al.* (2003) Identification of differentially expressed genes in high-density oligonucleotide arrays accounting for the quantification limits of the technology. *Biometrics*, **59**, 542–554.

Talantov,D. *et al.* (2005) Novel genes associated with malignant melanoma but not benign melanocytic lesions. *Clin. Cancer Res.*, **11**, 7234–7242.

Troyanskaya,O. *et al.* (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461.

Tusher,V. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci.*, **98**, 5116–5121.

Upton,G.J.G. and Harrisson,A.P. (2010) The detection of blur in Affymetrix GeneChips. *Stat. Appl. Genet. Mol. Biol.*, **9**, doi:10.2202/1544-6115.1590.

Van Hoewyk,D. *et al.* (2008) Transcriptome analyses give insights into selenium-stress responses and selenium tolerance mechanisms in arabidopsis. *Physiol. Plant.*, **132**, 236–253.

Whitley,E. and Ball,J. (2002) Statistics review 6: nonparametric methods. *Crit. Care*, **6**, 509–513.

Yang,Y. *et al.* (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acid Res.*, **30**, e15.

Yao,Z. *et al.* (2007) A Marfan syndrome gene expression phenotype in cultured skin fibroblasts. *BMC Genomics*, **8**, 319.

Zhao,H. *et al.* (2008) Multivariate hierarchical Bayesian model for differential gene expression analysis in microarray experiments. *BMC Bioinformatics*, **9** (Suppl. 1), S9.

Zhao,Y. and Pan,W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **19**, 1046–1054.

Zimmerman,J. *et al.* (2006) Multiple mechanisms limit the duration of wakefulness in Drosophila brain. *Physiol. Genomics*, **27**, 337–350.