

# Bayesian Modelling of Shared Gene Function

P. Sykacek<sup>a,\*</sup>, R. Clarkson<sup>b</sup>, C. Print<sup>c</sup>, R. Furlong<sup>d</sup>, G. Micklem<sup>e,f</sup>

Dept. of Biotechnology, BOKU University, Vienna<sup>a</sup>, School of Biosciences, Cardiff University<sup>b</sup>,  
Dept. of Molecular Medicine & Pathology, University of Auckland<sup>c</sup>, Dept. of Pathology<sup>d</sup>, Dept. of  
Genetics<sup>e</sup> and Cambridge Computational Biology Institute, Dept. of Applied Mathematics and  
Theoretical Physics<sup>f</sup>, University of Cambridge

## ABSTRACT

**Motivation** Biological assays are often carried out on tissues that contain many cell lineages and active pathways. Microarray data produced using such material therefore reflect superimpositions of biological processes. Analysing such data for shared gene function by means of well matched assays may help to provide a better focus on specific cell types and processes. The identification of genes that behave similarly in different biological systems also has the potential to reveal new insights into preserved biological mechanisms.

**Results** In this paper we propose a hierarchical Bayesian model allowing integrated analysis of several microarray data sets for shared gene function. Each transcript is associated with an indicator variable that selects whether binary class labels are predicted from expression values or by a classifier which is common to all transcripts. Each indicator selects the component models for all involved data sets simultaneously. A quantitative measure of shared gene function is obtained by inferring a probability measure over these indicators.

Through experiments on synthetic data we illustrate potential advantages of this Bayesian approach over a standard method. A shared analysis of matched microarray experiments covering a) a cycle of mouse mammary gland development and b) the process of endothelial cell apoptosis is proposed as a biological gold standard. Several useful sanity checks are introduced during data analysis and we confirm the prior biological belief that shared apoptosis events occur in both systems. We conclude that a Bayesian analysis for shared gene function has the potential to reveal new biological insights, unobtainable by other means.

**Availability** An online supplement and MatLab code are available at <http://www.sykacek.net/research.html#mcbf>.

**Contact** peter@sykacek.net

## 1 INTRODUCTION

Robust methods for microarray data analysis are important for advancing biological research, for example allowing more focused drug design and better assessment of pathogenicity (Bild *et al.*, 2006; Dave *et al.*, 2006). Many approaches are available for the analysis of single experiments, including methods based on statistical testing (Tusher *et al.*, 2001; Pan, 2002; Reiner *et al.*, 2003; Wernisch *et al.*, 2003), Bayesian automatic relevance determination (ARD) (Li *et al.*, 2002), Bayesian variable selection (Lee *et al.*, 2003; Bae & Mallick, 2004) and model selection (Li & Yang, 2002). While experiments on relatively homogenous cells lines grown in

tissue culture can give simple results, analysis of *individual experiments* using multicellular tissues provides a blurred view of the molecular mechanisms operating in the tissue. This is caused by the superimposition of biological processes in, and between, multiple cell types. Here we show that a combined analysis of cell culture and whole tissue microarray experiments for *shared* gene function reveals informative common details in the different systems. If the assays are well-matched, correspondingly more focussed answers are made available. Previous approaches for shared analyses have been through meta-analysis of gene lists (Yang *et al.*, 2005; DeConde *et al.*, 2006). In addition, recently (Huttenhower *et al.*, 2006) have proposed a Bayesian network for integrated analysis of microarray data, which combines pairwise correlations of expression patterns.

Our approach to inferring *shared gene function* is a fully Bayesian assessment of whether we can establish, across experiments, a relationship between binary biological classifications (e.g. mutant vs. wild type) and gene expression measurements. Using gene expression values as regressors, we model individual predictors by probit link regression (Spang *et al.*, 2002; Lee *et al.*, 2003). Like (Li *et al.*, 2002; Lee *et al.*, 2003), we use a Bayesian generalised linear model (GLM). However we consider additional aspects: to provide a more precise focus on specific molecular biological processes, our analysis combines information from heterogeneous sources, such as whole tissues, with well-matched cultured cells. Such an approach also has the potential to allow data from a new assay to be combined in a principled way with pre-existing data. This increases both the statistical power and cost-effectiveness of experiments. To allow calculation of probability measures reliably, we compare individual genes against a reference model. We suggest use of a reference which predicts biological classifications according to prior probabilities which are estimated from the class frequencies in the training data. This reference model does not use information from microarrays and must clearly be outperformed by functionally important genes. The sensitivity of the results to subjectively chosen hyperparameters is minimised by following (Bae & Mallick, 2004) and using hierarchical priors. Our first experiment is a simulation using synthetic data and compares the Bayesian approach with a simple meta analysis. We then apply the model to the shared analysis of two timecourse experiments: 1) a cycle of growth and regression in mammary glands *in vivo* (Clarkson *et al.*, 2004) together with 2) an assay of programmed endothelial cell death investigated *in vitro* (Johnson *et al.*, 2004). Apoptosis of endothelial cells is known to occur during the mammary gland cycle and may play an important role in this process (Matsumoto *et al.*, 1992; Djonov *et al.*, 2001). Computer simulations support this prior biological belief, provide high predictive accuracy and confirm the strategy introduced here.

\*to whom correspondence should be addressed

## 2 METHODS

The following discussion assumes that biological samples are available with both microarray gene expression measurements and a known binary classification. In such situations, the importance of individual genes can be assessed by Bayesian model criticism (Bernardo & Smith, 1994). Similar approaches have previously been used by (Li *et al.*, 2002; Lee *et al.*, 2003), who applied Bayesian variable selection to obtain a measure of gene importance. Assuming an overall number of  $T$  genes, classical Bayesian variable selection attempts to infer a probability measure over a  $2^T$  dimensional space. To maintain feasibility, we follow previous strategies (Pan, 2002) and consider single gene models.

### 2.1 Quantifying Shared Gene Function

A probabilistic quantification of shared gene function over several microarray experiments is obtained by generalising Bayesian model assessment for individual data sets. We assess gene function by quantifying the importance of a gene to a classifier which predicts a particular biological classification (e.g. mutant vs. wild type) from its expression values. We suggest in particular comparing two generalised linear regression (McCullagh & Nelder, 1989) models (GLMs) for every gene. One GLM predicts class labels from gene expression measurements and an intercept term. The other GLM predicts class labels only from an intercept. The latter GLM provides class priors and provides a base line accuracy which must be improved upon by a functionally relevant gene. Using  $t$  as gene index the binary indicator  $I_t$  denotes, for each gene, which of the two models is used within each of  $n$  predictions. Consequently we may express the posterior probability of the class label  $y_{n,t}$  by  $P(y_{n,t}|\mathbf{x}_{n,t}, \boldsymbol{\beta}_t, I_t)$ , where the regressor,  $\mathbf{x}_{n,t}$ , and the regression coefficient,  $\boldsymbol{\beta}_t$  depend on  $I_t$ . For  $I_t = 0$ ,  $\boldsymbol{\beta}_t$  is a scalar and  $\mathbf{x}_{n,t} = 1$  and for  $I_t = 1$ , both  $\boldsymbol{\beta}_t$  and  $\mathbf{x}_{n,t} = [\xi_{n,t}, 1]^T$  are two dimensional column vectors, with  $\xi_{n,t}$  denoting suitably transformed and normalised gene expression values. The prediction of  $y_{n,t}$  can thus be regarded as a two component mixture of GLMs. In a Bayesian context, (Spang *et al.*, 2002; Lee *et al.*, 2003) found that it is convenient to model binary classifications by probit link regression. The probit link GLM predicts  $P(y_{n,t} = 0|\mathbf{x}_{n,t}, \boldsymbol{\beta}_t, I_t)$  as the value of a Gaussian cumulative distribution function (cdf) with mean  $\mathbf{x}_{n,t}^T \boldsymbol{\beta}_t$  and unit standard deviation at zero. If we denote all classifications by  $D_t = [y_{1,t}, \dots, y_{n,t}]$  and all regressors by  $X_t = [\mathbf{x}_{1,t}, \dots, \mathbf{x}_{n,t}]$ , the probabilities  $P(y_{n,t}|\mathbf{x}_{n,t}, \boldsymbol{\beta}_t, I_t)$  give rise to the likelihood

$$p(D_t|X_t, \boldsymbol{\beta}_t, I_t) = \prod_n P(y_{n,t}|\mathbf{x}_{n,t}, \boldsymbol{\beta}_t, I_t). \quad (1)$$

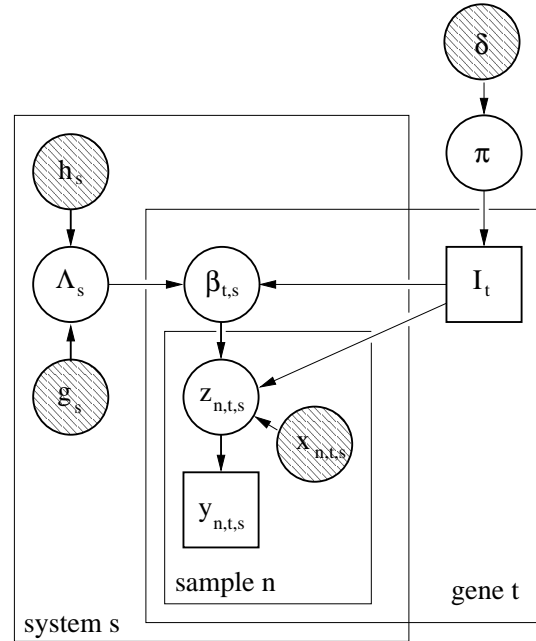
To obtain the joint distribution,  $p(D_t, \boldsymbol{\beta}_t|X_t, I_t)$ , Bayesian inference requires multiplying the likelihood with a prior over regression coefficients,  $p(\boldsymbol{\beta}_t|I_t)$ . The functional importance of genes is quantified by the posterior probability  $P(I_t = 1|D_t, X_t)$ , (Bernardo & Smith, 1994). Normalising the product of prior probability  $P(I_t)$  and marginal likelihood  $p(D_t|X_t, I_t) = \int \boldsymbol{\beta}_t p(D_t, \boldsymbol{\beta}_t|X_t, I_t) d\boldsymbol{\beta}_t$  we obtain the posterior

$$P(I_t|D_t, X_t) = \frac{P(I_t)p(D_t|I_t, X_t)}{\sum_{I_t} P(I_t)p(D_t|I_t, X_t)}. \quad (2)$$

This principle is extended to inferring genes that show a shared functional importance in several microarray experiments. After standardising the naming of transcripts across experiments e.g. by means of orthology mappings, index  $t$  represents the same gene in all  $s$  microarray experiments. Assuming that, given  $I_t$ , all  $s$  experiments are conditionally independent,

$$P(I_t|D_{1,t}, X_{1,t}, \dots, D_{t,S}, X_{t,S}) = \frac{P(I_t) \prod_s p(D_{t,s}|I_t, X_{t,s})}{\sum_{I_t} P(I_t) \prod_s p(D_{t,s}|I_t, X_{t,s})} \quad (3)$$

provides a measure of shared gene importance. Typically an analysis of shared gene function selects subsets of important genes from individual data sets and searches for genes shared by all subsets (see e.g. (Hockley *et al.*, 2006)). Unfortunately this discards rank information and in addition is sensitive to



**Fig. 1.** This directed acyclic graph (DAG) illustrates a probabilistic model for the analysis of shared gene function. Large rectangles indicate a replicated conditional independence relationship. Circles represent continuous and small squares discrete random variables. Shaded nodes represent observed variables. All clear nodes are subject to inference. Gene function is measured by the posterior probability over the binary indicator  $I_t$ .

censoring effects as a gene of very high importance in one experiment might be just below the threshold in another experiment and thus not appear in the final list. An analysis of shared gene function along the lines of Equation (3) does not suffer either drawback.

There are potential dangers if the proposed approach is implemented naïvely: Bayesian inference can suffer from sensitivity problems (Bernardo & Smith, 1994). As a result, the model probabilities in Equation (3) and the implied ranking might depend crucially on the hyper parameters used to parameterise the prior  $p(\boldsymbol{\beta}_t|I_t)$ . We avoid this problem by specifying a prior over all the hyperparameters that could contribute to such an adverse effect, and then inferring the hyperparameters as well.

### 2.2 Robust Modelling of Shared Gene Function

A model which takes these considerations into account is presented in Figure 1. The core of the model is a latent variable implementation (Andrieu *et al.*, 2002; Holmes & Denison, 2003; Lee *et al.*, 2003) of the binary classifier discussed above. We use index  $s$  as index over microarray experiments,  $t$  as transcript index and  $n$  as observation index within experiments. Variable  $z_{n,t,s}$  denotes a latent variable which, conditional on its parents, has a univariate Gaussian distribution with mean  $\mathbf{x}_{n,t,s}^T \boldsymbol{\beta}_{t,s}$  and unit standard deviation. For  $I_t = 1$  we use the log expressions as regressor  $\mathbf{x}_{n,t,s}$ . For  $I_t = 0$  the regression is based on a common reference model which uses only an intercept. The indicator  $I_t$  determines the component models for all  $s$  microarray experiments simultaneously and this allows inference of the probability measures of shared gene function,  $P(I_t|D_{1,t}, X_{1,t}, \dots, D_{t,S}, X_{t,S})$ .

Variable  $\boldsymbol{\beta}_{t,s}$  denotes the regression coefficients of the GLM. The prior over  $\boldsymbol{\beta}_{t,s}$  is a Gaussian distribution with zero mean and diagonal precision  $\Lambda_s$ . The robustness of inferring  $P(I_t|D_{1,t}, X_{1,t}, \dots, D_{t,S}, X_{t,S})$  is improved by specifying this prior hierarchically and using a product of Gamma distributions as prior over the diagonal elements of  $\Lambda_s$ . This

Gamma prior is parametrised by the coefficients  $h_s$  and  $g_s$ . The binomial prior over  $I_t$  is specified indirectly by giving parameters  $\pi$  a beta prior with counts  $\delta$ . Using these hierarchical priors reduces the sensitivity of  $P(I_t|D_{1,t}, X_{1,t}, \dots, D_{t,S}, X_{t,S})$  to the choice of hyperparameters. The posterior distributions over  $\Lambda_s$  and  $\pi$  depend on the prior and on all information the data provides about these variables. Since both variables depend on all transcripts, the influence of the hyperparameters is greatly reduced. This important aspect of the model is further investigated in the experiments section below. Concerning the relationship between the latent variable,  $z_{n,t,s}$ , and the biological classifications,  $y_{n,t,s}$ , if  $z_{n,t,s} < 0$ , the probability  $P(y_{n,t,s} = 0|z_{n,t,s})$  is 1, otherwise it is 0.  $P(y_{n,t,s} = 1|z_{n,t,s})$  is  $1 - P(y_{n,t,s} = 0|z_{n,t,s})$ . By integrating over  $z_{n,t,s}$ , (Denison *et al.*, 2002) show that this setting corresponds to a probit link GLM. Mathematical details of the joint distribution can be found in Equation (6) in the Appendix.

### 2.3 Variational Inference

Inferring shared gene function requires calculating the marginal posterior distributions over all  $I_t$  from the DAG in Figure 1. A Markov Chain Monte Carlo (MCMC) technique along the lines of (Green, 1995) could, at the expense of high computational cost, approximate  $P(I_t|D_{1,t}, X_{1,t}, \dots, D_{t,S}, X_{t,S})$  with arbitrary accuracy, but careful model checking would be essential and multiple MCMC runs would be required. With conventional computer infrastructure such an approach quickly becomes infeasible. A Laplace (MacKay, 1992; Chu *et al.*, 2005) or a variational approximation (Attias, 1999; Jordan *et al.*, 1999; Frey, 1998) is computationally simpler and better suited for our purposes. We decided to base model inference on the variational learning framework that was recently used in (Beal *et al.*, 2005). Variational learning implies approximating the joint distribution of the model (Equation (6) in the Appendix), by a factorising Ansatz. For brevity we use  $\theta$  for all random variables,  $D = \{D_1, \dots, D_S\}$  and  $X = \{X_1, \dots, X_S\}$  to obtain the approximation

$$p(\theta|G, H, \delta, D, X) \approx Q(\theta) = Q(\pi) \prod_s Q(\Lambda_s) \quad (4)$$

$$\times \prod_t Q(I_t) \prod_{I_t, t, s} \left( Q(\beta_{I_t, t, s}) \prod_n Q(z_{I_t, n, t, s}) \right).$$

We can now use Jensen’s inequality and obtain a lower bound on the log marginal likelihood of the DAG.

$$\log \left( \int_{\theta} p(\theta, D|G, H, \delta, X) d\theta \right) \geq \quad (5)$$

$$\int_{\theta} \left( \log(p(\theta, D|G, H, \delta, X)) - \log(Q(\theta)) \right) Q(\theta) d\theta$$

Variational learning requires maximising the lower bound (second line in Equation (5)). This is done iteratively by integrating the negative free energy with respect to all but one  $Q$ -distributions from Equation (4) and maximising the resulting functional with respect to the remaining  $Q$ -distribution. This provides, for every  $Q$ -function in Equation (4), a separate update rule. The essential difference between typical modelling approaches and the DAG in Figure 1 is the hierarchical priors that couple all individual gene models. Details of the  $Q$ -function updates are provided in the Appendix.

### 2.4 Computation of Shared Gene Function

An algorithm which infers shared gene function will iterate over all maximisation steps for the  $Q$ -distributions in Equation (4) and monitor the negative free energy. Although there is no guarantee that we will achieve the optimum result, it is very important to use all possible safeguards to detect potentially misleading conclusions. In this case, we have to ensure that the calculated measure of shared gene function does not overly depend on the chosen prior and so a sensitivity analysis is vital. Such an analysis examines the effect of chosen hyperparameters on the probability measure of shared gene function ( $Q(I_t)$  in Equation (12) and  $P(G \equiv t|D, X)$  in Equation (13)). An additional sanity check is obtained by cross validation. This provides an estimate of several gene measures and generalisation accuracy, each obtained from a

slightly perturbed data set. Poor generalisation accuracy or large variation in the gene measures would be warning signs. If one has prior biological knowledge about the gene expression assays, one can also check whether this is in agreement with an inference of GO categories (Dopazo, 2006). Therefore we have inferred active GO biological process categories by Fisher’s exact test (Al-Shahrour *et al.*, 2004). Details regarding implementation issues including pseudo code and how to predict the probabilities of unknown biological classifications can be found in the online supplement (Sykacek *et al.*, 2007).

## 3 EXPERIMENTS

Based on synthetic data, this section will first illustrate a comparison of the Bayesian approach with a standard method for shared gene analyses. A combined analysis of a microarray time course of mouse mammary gland development (Clarkson *et al.*, 2004) and an in-vitro culture of Endothelial cells under growth factor withdrawal. (Johnson *et al.*, 2004) is then used to infer genes that are related to cell number control in both processes. These data serve both as a biological gold standard and for demonstrating useful diagnostics.

### 3.1 Synthetic Data

We illustrate Bayesian modelling of shared gene function using two data sets, each containing three groups of variables representing simulated gene expression measurements. Within a group, the variables of both experiments were generated from the same distribution. Class separability is set to be different between groups and so genes from the different groups have varying degree of importance to the process of classification. The result in Table 1, column “Ranking” shows that the Bayesian approach ranks the different groups successfully, a property which is important for guiding follow-up experiments and not available by filtering gene lists for shared gene names as, for example, used in (Hockley *et al.*, 2006).

**Table 1.** Bayesian Rank Lists for Synthetic Data

	Test of Censoring		Ranking	
	gene nr.	$Q(I_t)$	gene nr.	$Q(I_t)$
group 1	gene 3	0.999	gene 4	0.999
	gene 1	0.999	gene 1	0.999
	gene 2	0.999	gene 3	0.999
	gene 4	0.999	gene 2	0.999
group 2	gene 6	0.998	gene 7	0.554
	gene 8	0.995	gene 8	0.499
	gene 7	0.989	gene 6	0.400
	gene 5	0.969	gene 5	0.194
group 3	gene 10	0.147	gene 12	0.049
	gene 11	0.088	gene 10	0.040
	gene 12	0.042	gene 11	0.039
	gene 9	0.033	gene 9	0.034

Analysing individually thresholded gene lists for shared gene function might censor important genes randomly. Thresholding converts a continuous measure into a binary decision about gene function and random effects like selecting particular biological replicates can alter the continuous measures. All borderline genes will thus, to some extent by chance, appear important or not. If one assay assesses particular genes as most relevant and a second assay finds

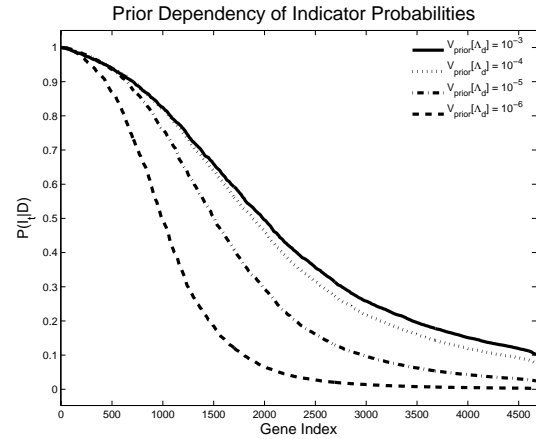
these genes just outside the threshold, combining the lists will censor promising candidates. Using synthetic data, we can illustrate this effect. The first and third group of genes were for both datasets generated from the same distribution. The genes in group one are relevant and those in group three irrelevant as predictors of class labels. Consequently, the Bayesian assessment and the conventional approach agree for these groups about shared gene function. The genes in the second group show for one dataset very high and for the other only moderate class separability. The gene measures obtained from the second dataset are in the range of the threshold and, as is shown in our online supplement (Sykacek *et al.*, 2007), all genes but one are censored. By calculating shared gene function directly, the Bayesian approach overcomes this problem and, as is illustrated in table 1, column “Test of Censoring”, censoring is avoided.

### 3.2 Searching for Shared Apoptosis Genes

The approach of searching for shared genes presented here can be applied in any situation where several microarray assays covering biological state transitions need be assessed for common genetic behaviour. Here we are interested in inferring which genes are of shared relevance during two transitions: 1) the transition from lactation to involution in a mouse mammary gland cycle *in vivo*, and 2) the transition from normal growth conditions to growth factor withdrawal in an *in vitro* culture of human Endothelial cells. The biological motivation for this investigation is that both tissues contain endothelial cells, and the transitions both involve cell death. Therefore a shared analysis has the potential to provide genetic markers involved in endothelial cell death within the mammary gland, even though no gene expression measurements were obtained from these cells in isolation.

The mouse mammary gland data was taken from (Clarkson *et al.*, 2004) and timepoints were labelled as being apoptotic on the basis that apoptosis is induced in involution. Expression values were obtained from two biological replicates measured using Affymetrix Murine U74 arrays, with six samples from lactation and ten samples during involution. From (Johnson *et al.*, 2004), who studied apoptosis in human endothelial cells, we took five samples under growth factor withdrawal and five control samples that were measured with Affymetrix human U95 arrays. Cross annotations were taken from the Affymetrix databases such that human and mouse genes that are orthologues were labelled the same. To increase consistency, we followed (Mecham *et al.*, 2004) and took only such probes that could be matched in sequence by a NCBI blast search. This left us with 4581 cross annotated transcripts. The expression values were extracted with MAS 5.0, converted to log scale and normalised by within-slide mean removal and scale adjustment. To ensure that all regressors are on a similar scale, we transformed the log expressions of each transcript to zero mean and unit standard deviation. Analysis for shared gene function compared the transcript specific GLMs with an intercept only GLM that models endothelial cell apoptosis and mammary tissue involution based on the prior frequencies of labels as observed in the training data.

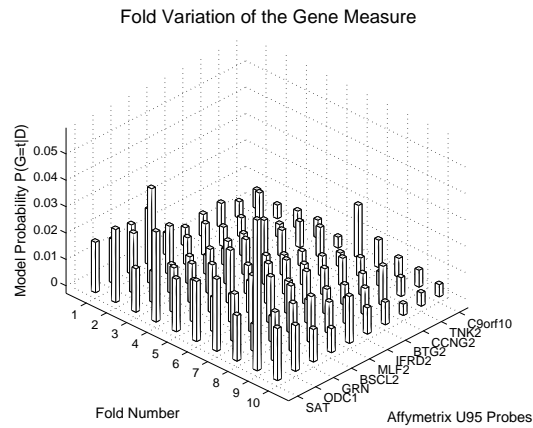
**3.2.1 Sensitivity to Prior Choices** To ensure that inferences about shared gene function do not depend crucially on the chosen hyperparameters, we specify the priors  $P(I_t|\pi)$  and  $p(\beta_{t,s}|\Lambda_s)$  indirectly. We interpret  $\pi$  as the a-priori fraction of genes we believe to be involved in the biological process. This requires us to be uninformative about  $\pi$  by using a small prior count like  $\delta = 1$ .



**Fig. 2.** Ordered probabilities of gene function plotted over all transcripts for different priors over  $\Lambda_s[d, d]$ . The expectation of  $\Lambda_s[d, d]$  is fixed at 0.01 and the prior variances change from  $10^{-3}$  to  $10^{-6}$ .

The situation with  $\Lambda_s$  is more subtle, since our choice will have an indirect effect on  $Q(I_t)$ . Therefore it is imperative to study the sensitivity of the model to choices of  $g_s$  and  $h_s$ . A sensitivity analysis will depend mainly on the variance,  $V[\Lambda_s[d, d]]$ , with respect to the prior  $p(\Lambda_s|g_s, h_s)$ . Thus we may fix the prior expectation  $E[\Lambda_s[d, d]]$  to 0.01 and vary  $V[\Lambda_s[d, d]]$  linear on a log scale from  $10^{-6}$  to 1 by using  $g_s = \{10^{-5}, 10^{-4}, \dots, 10^2\}$  and  $h_s = \{10^{-3}, 10^{-2}, \dots, 10^4\}$ . Computer simulations revealed that the prior over  $\Lambda_s[d, d]$  has no effect on the approximate posterior  $Q(\Lambda_s[d, d])$ , if we choose a variance that is larger than  $10^{-3}$ . See online supplement for a graph with details of this investigation. The ordered probability measures  $Q(I_t \equiv 1)$  in Figure 2 allow the same conclusion from the gene measure. Variances smaller than  $10^{-3}$  result in  $p(\Lambda_s|g_s, h_s)$  having an undesired effect on  $Q(I_t)$ . Curves for prior variances that are larger than  $10^{-3}$  are not shown as they are essentially indistinguishable from the curve obtained for that value. Therefore we concluded that sensitivity to the chosen prior is avoided if we set  $h_s \leq 10$  and  $g_s \leq 10^{-1}$ .

**3.2.2 Analysis of Gene Function** If we assume equal cost for both types of error in deciding about gene function, we should select all genes that have  $Q(I_t \equiv 1)$  larger than 0.5. For the chosen hyperparameters, this suggests that 2164 transcripts are potentially of interest: a ranked list in tab-delimited format is provided in the online supplement. This is a large though plausible number of genes. Given that we chose uninformative hyperparameters, we see that the hierarchical model allows the prior over regression coefficients to adjust to the data sets. Here, the data favours small regression coefficients. An illustration of this effect on synthetic data is provided in the online supplement. For the mammary gland data, the expected variances in the priors over intercept and regression coefficients are 0.13 and 0.93. The respective values for the apoptosis data are 0.11 and 1.77. The small variance of the effective prior over the regression parameter implies a small complexity penalty for the larger model (Jefferys & Berger, 1992). Therefore transcripts are favoured over the intercept-only model, even if they provide only a little information about the biological classification. To validate the result, we used the model for ten-fold cross testing. Figure 3 illustrates the fold



**Fig. 3.** This figure illustrates the fold variation of the gene measure  $P(G = t|D)$ , for those ten genes that were ranked highest using all data. We observe some deviations from the optimal ordering which are due to random deviations in the microarray data.

variation of the ten largest values of the gene measure,  $P(G = t|D)$ . We see that fold-based rankings and the overall ranking give similar results. However, there are some deviations which indicate that some slides are more influential than others. This effect should be reduced by using larger sample sizes. Cross testing is based on averaging predictions which are weighted according to Equation (13) (c.f. (Sykacek *et al.*, 2007) for further algorithmic details). Selecting the top-ranked transcript predictions (until the cumulative gene measure,  $\sum_t P(G = t|D)$ , reaches 0.8), produces on average 424 transcripts, and we obtain for both data sets a generalisation accuracy of 100%. High generalisation accuracy is reassuring since it suggests that the probability measure did favour informative genes.

To assess the biological plausibility of our shared gene measure, we followed (Lewin *et al.*, 2006), who inferred active GO categories by Fisher’s exact test. To do so, we regarded the top 30% of the genes from the rank list as active and the 30% genes at the lower end as inactive and inferred, for every GO category, the significance level of abundance of active over inactive genes. To increase the robustness of this assessment, we used the fold-based gene measures as they arose from estimating the generalization accuracies. A gene is counted as active, if its indicator probability  $Q(I_t)_{\text{fold}}$  is larger than 0.5 and the regression coefficients  $\beta_{t,s}$  have the same sign for all experiments  $s$ . After Bonferroni correction for multiple testing, we obtain 238 gene ontology categories with a significant abundance of active over inactive genes. Such analysis finds many GO categories that are indicative of shared metabolic changes. Our prior expectation that both assays share certain events related to cell death is confirmed since we find “programmed cell death” (GO:0012501), “regulation of apoptosis” (GO:0042981), “negative regulation of apoptosis” (GO:0043066), “anti-apoptosis” (GO:0006916), “caspase activation via cytochrome c” (GO:0008635), “induction of apoptosis via death domain receptors” (GO:0008625) and “apoptosis” (GO:0006915) are, with high significance, enriched by active genes. An XML file with all active GO categories is provided as an online supplement.

As a result of the hierarchical prior chosen in this work, we find in an equal cost scenario many genes with functional importance.

It is evident from Figure 2 that we will obtain fewer transcripts with probabilities larger than 0.5 by forcing small precisions in the prior over regression coefficients. This however meant to construct a convenient probability measure that has little support from the data and is thus not recommended. Instead we recommend a pragmatic approach of taking as many transcripts as one can afford in subsequent steps according to the ranking of shared functional importance, and possibly using additional criteria such as unknown GO or pathway annotation.

## 4 DISCUSSION

In this paper we propose a probabilistic model for a principled integrated analysis of several microarray experiments. The proposed model is a result of careful considerations of sensitivity issues. By specifying priors hierarchically, we reduce the effect of all hyperparameters of the algorithm and provide conclusions that are justified by the data. The proposed approach shares with meta analyses (Yang *et al.*, 2005; DeConde *et al.*, 2006; Hockley *et al.*, 2006) the advantage of combining data sets where the actual expression levels of different experiments need not be matched. A considerable advantage of a Bayesian analysis is that it provides rank information and does not suffer from the censoring effects of simple approaches that combine thresholded gene lists.

An application to shared analysis of gene function in mouse mammary gland tissue and an endothelial cell line illustrates how to diagnose and avoid potential sensitivity problems. Assessments of predictive accuracy and a confirmation of biological expectations reinforce the plausibility of the proposed approach. The results suggest that avoiding sensitivity is imperative in analysing microarray data, even if one can’t follow up a large number of positive genes. The proposed approach has two desirable properties. First it allows us to increase statistical power and cost efficiency by combining new assays with existing data. Even more important is the prospect of a successful search for molecular biological mechanisms that are shared by developmental processes or state transitions in different tissues or species. A fully Bayesian analysis for shared gene function therefore has the potential to lead to biological insights that might be unobtainable by other means.

## ACKNOWLEDGEMENTS

The authors thank David MacKay for his advice and the editor and the reviewers for their helpful comments. This work was supported by the BBSRC’s Exploiting Genomics initiative under ref. 8/EGH16106, “Shared Genetic Pathways in Cell Number Control”. Peter Sykacek is currently funded by the WWTF, ACBT, ARC Seibersdorf and Baxter AG and grateful for their support.

## REFERENCES

- Al-Shahrour, F., Díaz-Uriarte, R. & Dopazo, J. (2004) FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
- Andrieu, C., de Freitas, J. & Doucet, A. (2002) Rao-Blackwellised particle filtering via data augmentation. In *Advances in Neural Processing Systems 14*, (Dietterich, T., Becker, S. & Ghahramani, Z., eds), pp. 561–567 MIT Press.
- Atias, H. (1999) Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)* pp. 21–30 Morgan Kaufmann Publishers, San Francisco, CA.
- Bae, K. & Mallick, B. K. (2004) Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, **20** (18), 3423–3430.

Beal, M. J., Falciani, F. L., Ghahramani, Z., Rangel, Z. & Wild, D. (2005) A Bayesian Approach to Reconstructing Genetic Regulatory Networks with Hidden Factors. *Bioinformatics*, **21**, 349–356.

Bernardo, J. M. & Smith, A. F. M. (1994) *Bayesian Theory*. Wiley, Chichester.

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., Chasse, D., Joshi, M. B., Harpole, D., Lancaster, J. M., Berchuck, A., Jr, J. A. O., Marks, J. R., Dressman, H. K., West, M. & Nevins, J. R. (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439** (7074), 353–357.

Chu, W., Ghahramani, Z., Falciani, F. & Wild, D. L. (2005) Biomarker Discovery with Gaussian Processes in Microarray Gene Expression Data. *Bioinformatics*, **21**, 3385–3393.

Clarkson, R. W. E., Wayland, M. T., Lee, J., Freeman, T. & Watson, C. J. (2004) Gene expression profiling of mammary gland development reveals putative roles for death receptors and immune mediators in post-lactational regression. *Breast Cancer Res.*, **6** (2), 92–109.

Dave, S. S., Fu, K., Wright, G. W., Lam, L. T., Kluin, P., Boerma, E.-J., Greiner, T. C., Weisenburger, D. D., Rosenwald, A., Ott, G., Müller-Hermelink, H.-K., Gascoyne, R. D., Delabie, J., Rimsza, L. M., Braziel, R. M., Grogan, T. M., Campo, E., Jaffe, E. S., Dave, B. J., Sanger, W., Bast, M., Vose, J. M., Armitage, J. O., Connors, J. M., Smeland, E. B., Kvaloy, S., Holte, H., Fisher, R. I., Miller, T. P., Montserrat, E., Wilson, W. H., Bahl, M., Zhao, H., Yang, L., Powell, J., Simon, R., Chan, W. C. & Staudt, L. M. (2006) Molecular diagnosis of burkitt's lymphoma. *NEJM*, **354** (23), 2431–2442.

DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., & Etzioni, R. (2006) Combining Results of Microarray Experiments: A Rank Aggregation Approach. *Statistical Applications in Genetics and Molecular Biology*, **5** (1), Article 15. Available at: <http://www.bepress.com/sagmb/vol5/iss1/art15>.

Denison, D. G. T., Holmes, C. C., Mallik, B. K. & Smith, A. F. M. (2002) *Bayesian methods for nonlinear classification and regression*. J. Wiley & Sons.

Djonov, V., Andres, A. C. & Ziemiecki, A. (2001) Vascular remodelling during the normal and malignant life cycle of the mammary gland. *Microscopy Research and Technique*, **15**, 182–189.

Dopazo, J. (2006) Functional Interpretation of Microarray Experiments. *OMICS: A Journal of Integrative Biology*, **10** (3), 398–410.

Frey, B. (1998) *Graphical models for machine learning and digital communication*. MIT Press, Cambridge Massachusetts.

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

Hockley, S. L., Arlt, V. M., Brewer, D., Giddings, I. & Phillips, D. H. (2006) Time- and concentration-dependent changes in gene expression induced by benzo(a)pyrene in two human cell lines, MCF-7 and HepG2. *BMC Genomics*, **7** (260).

Holmes, C. C. & Denison, D. G. T. (2003) Classification with Bayesian MARS. *Machine Learning*, **50**, 150–173.

Huttenhower, C., Hibbs, M., C.Meyers & Troyanskaya, O. G. (2006) A scalable method for integration and functional analysis of multiple microarray datasets. *Bioinformatics*, **22** (23), 2890–2897.

Jefferys, W. & Berger, J. (1992) Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64–72.

Johnson, N. A., Sengupta, S., Saidi, S. A., Lessan, K., Charnock-Jones, S. D., Scott, L., Stephens, R., Freeman, T. C., Tom, B. D., Harris, M., Denyer, G., Sundaram, M., ans S. K. Smith, R. S. & Print, C. G. (2004) Endothelial cells preparing to die by apoptosis initiate a program of transcriptome and glycome regulation. *The FASEB Journal*, **18**, 188–190.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999) An introduction to variational methods for graphical models. In *Learning in Graphical Models*, (Jordan, M. I., ed.), MIT Press Cambridge, MA pp. 105–161.

Lee, K. E., Sha, N., Dougherty, R. R., Vanucci, M. & Mallick, B. K. (2003) Gene selection: a Bayesian variable selection approach. *Bioinformatics*, **19** (1), 90–97.

Lewin, A., Richardson, S., Marshall, C., Glazier, A. & Aitman, T. (2006) Bayesian Modelling of Differential Gene Expression. *Biometrics*, **62** (1), 10–18.

Li, W. & Yang, Y. (2002) How many genes are needed for a discriminant microarray data analysis. In *Methods of Microarray Data Analysis*, (Lin, S. M. & Johnson, K. F., eds), Kluwer Academic pp. 137–150.

Li, Y., Campbell, C. & Tipping, M. (2002) Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, **18** (10), 1332–1339.

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

Matsumoto, M., Nishinakagawa, H., Kurohmaru, M., Y, Y. H. & Otsuka, J. (1992) Pregnancy and lactation affect the microvasculature of the mammary gland in mice. *The Journal of Veterinary Medical Science*, **54**, 937–943.

McCullagh, P. & Nelder, J. A. (1989) *Generalized Linear Models*. Second edition., Chapman & Hall, London.

Mecham, B. H., Klus, G. T., Strovel, J., Augustus, M., Byrne, D., Bozso, P., Wetmore, D. Z., Mariani, T. J., Kohane, I. S. & Szallasi, Z. (2004) Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Research*, **32**, e74.

Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Reiner, A., Yekutieli, D. & Benjamini, Y. (2003) Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, **19**, 368–375.

Spang, P., Blanchette, C., Zuzana, H., Marks, J. R., Nevins, J. & West, M. (2002) Prediction and uncertainty in the analysis of gene expression profiles. *In Silico Biol.*, **2** (3), 369–381.

Sykacek, P., Clarkson, R., Print, C., Furlong, R. & Micklem, G. (2007). Online Supplement to: Bayesian Modeling of Shared Gene Function. Technical report Department of Biotechnology, BOKU University, Vienna. [Available at <http://www.sykacek.net/pubs.html#TR072>].

Tusher, V., Tibshirani, R. & Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences, USA*, **98**, 5116–5121.

Wernisch, L., Kendall, S. L., Soneji, S., Wietzorrek, A., Parish, T., Hinds, J., Butcher, P. G. & Stoker, N. G. (2003) Analysis of whole-genome microarray replicates using mixed models. *Bioinformatics*, **19** (1), 53–61.

Yang, X., Bentink, S. & Spang, R. (2005) Detecting common gene expression patterns in multiple cancer outcome entities. *Biomedical Microdevices*, **7** (3), 247–251.

## APPENDIX

### Joint Distribution

If we abbreviate the regression coefficients of all  $T$  transcripts and  $S$  experiments as  $B = \{\beta_{1,1}, \dots, \beta_{T,S}\}$ , all indicators as  $J = \{I_1, \dots, I_T\}$ , all latent variables as  $Z = \{z_{1,1,1}, \dots, z_{N_S, T, S}\}$ , the precision matrices as  $L = \{\Lambda_1, \dots, \Lambda_S\}$ , the hyperparameters as  $G = \{g_1, \dots, g_S\}$  and  $H = \{h_1, \dots, h_S\}$ , we may express the joint distribution as

$$p(\pi, L, B, J, Z, D | G, H, \delta) = p(\pi | \delta) \quad (6)$$

$$\times \prod_{s=1}^S p(\Lambda_s | g_s, h_s) \prod_t p(I_t | \pi)$$

$$\times \prod_{t=1}^T \prod_{s=1}^S \left( p(\beta_{t,s} | \Lambda_s, I_t) \right.$$

$$\left. \times \prod_{n=1}^{N_s} \left( p(z_{n,t,s} | \beta_{t,s}, I_t) p(y_{n,t,s} | z_{n,t,s}) \right) \right).$$

The conditional probability of  $y_{n,t,s}$  given  $z_{n,t,s}$  is

$$P(y_{n,t,s} = 1 | z_{n,t,s}) = \begin{cases} 1, & \text{if } z_{n,t,s} > 0 \\ 0, & \text{if } z_{n,t,s} \leq 0 \end{cases}$$

$$P(y_{n,t,s} = 0 | z_{n,t,s}) = 1 - P(y_{n,t,s} = 1 | z_{n,t,s})$$

The conditional probability of  $z_{n,t,s}$  given  $\beta_{t,s}$  and  $I_t$  is a univariate Gaussian

$$p(z_{n,t,s} | \beta_{t,s}, I_t) = (2\pi)^{-0.5}$$

$$\times \exp \left( -0.5(z_{n,t,s} - \mathbf{x}_{n,t,s, I_t}^T \beta_{t,s, I_t})^2 \right).$$

The prior over  $\beta_{t,s}$  is a multivariate Gaussian

$$p(\beta_{t,s} | \Lambda_s, I_t) = (2\pi)^{-\frac{d_{I_t}}{2}} |\Lambda_{s, I_t}|^{0.5}$$

$$\times \exp \left( -0.5 \beta_{t,s, I_t}^T \Lambda_{s, I_t} \beta_{t,s, I_t} \right).$$

The prior over  $I_t$  is Bernoulli distributed

$$p(I_t | \pi) = \Theta^{I_t} (1 - \Theta)^{1 - I_t}.$$

The prior over  $\Lambda_s$  is a product of Gamma distributions

$$p(\Lambda_s | g_s, h_s) = \prod_d \left( \frac{h_s^{g_s}}{\Gamma(g_s)} \Lambda_s[d, d]^{g_s-1} \right. \\ \left. \times \exp(-\Lambda_s[d, d] h_s) \right)$$

and the prior over  $\pi$  a Beta distribution

$$p(\pi | \delta) = \frac{\Gamma(\delta_1 + \delta_2)}{\Gamma(\delta_1)\Gamma(\delta_2)} \Theta^{(\delta_1-1)} (1 - \Theta)^{(\delta_2-1)}.$$

To indicate a conditional dependency on  $I_t$ , we use the latter as index. The equations link to the graph in Figure 1 by  $\pi = [\Theta, (1 - \Theta)]$  (i.e. a two state probability) and  $\delta = [\delta_1, \delta_2]$  specifying the prior counts in the distribution over  $\pi$ . The  $d$ -th diagonal element of the matrix  $\Lambda_s$  is denoted as  $\Lambda_s[d, d]$ .

## Variational Maximisation

Variational learning follows the generic approach sketched in section 2.3. We iterate over integrating the negative free energy from Equation (5) with respect to all but one  $Q$ -distributions and maximising the resulting functional. For the  $Q$ -distributions in Equation (4) we get the following update equations.

*Maximising with respect to  $Q(z_{I_t, n, t, s})$*  results in a truncated Normal distribution

$$Q(z_{I_t, n, t, s}) = (2\pi)^{-0.5} \frac{1}{\Phi(b_{n, t, s}) - \Phi(a_{n, t, s})} \\ \times \exp(-0.5(z_{I_t, n, t, s} - \hat{z}_{I_t, n, t, s})^2) \quad (7)$$

$$\text{where } \hat{z}_{I_t, n, t, s} = \mathbf{x}_{I_t, n, t, s}^T \hat{\beta}_{I_t, t, s}.$$

We use  $\hat{\beta}_{I_t, t, s}$  for the mode of the  $Q$ -distribution over  $\beta_{I_t, t, s}$ . The expressions  $\Phi(a_{n, t, s})$  and  $\Phi(b_{n, t, s})$  denote Gaussian cdfs with mean  $\hat{z}_{I_t, n, t, s}$  and unit standard deviation at  $a_{n, t, s}$  and  $b_{n, t, s}$ . The latter are implied by the definition of  $P(y_{n, t, s} | z_{n, t, s})$ : for  $y_{n, t, s} = 1$ , we get  $a_{n, t, s} = 0$  and  $b_{n, t, s} = \infty$ ; for  $y_{n, t, s} = 0$ , we get  $a_{n, t, s} = -\infty$  and  $b_{n, t, s} = 0$ .

*Maximising with respect to  $Q(\beta_{I_t, n, t, s})$*  results in a Gaussian distribution

$$Q(\beta_{I_t, n, t, s}) = (2\pi)^{-0.5 d_{I_t, t, s}} |\hat{\Lambda}_{I_t, t, s}|^{0.5} \\ \times \exp\left(-\frac{1}{2}(\beta_{I_t, t, s} - \hat{\beta}_{I_t, t, s})^T \hat{\Lambda}_{I_t, t, s} \right. \\ \left. \times (\beta_{I_t, t, s} - \hat{\beta}_{I_t, t, s})\right) \quad (8)$$

where

$$\hat{\Lambda}_{I_t, t, s} = \langle \Lambda_{I_t, s} \rangle + \sum_{n=1}^{N_s} \mathbf{x}_{I_t, n, t, s} \mathbf{x}_{I_t, n, t, s}^T$$

$$\hat{\beta}_{I_t, t, s} = \hat{\Lambda}_{I_t, t, s}^{-1} \sum_{n=1}^{N_s} \mathbf{x}_{I_t, n, t, s} \langle z_{I_t, n, t, s} \rangle$$

$$\langle z_{I_t, n, t, s} \rangle = \hat{z}_{I_t, n, t, s} - \frac{f(b_{n, t, s}) - f(a_{n, t, s})}{\Phi(b_{n, t, s}) - \Phi(a_{n, t, s})},$$

with  $\langle \Lambda_{I_t, s} \rangle = \text{diag} \left( \frac{\hat{g}_{1, s}}{\hat{h}_{1, s}}, \dots, \frac{\hat{g}_{d_{I_t, s}}}{\hat{h}_{d_{I_t, s}}} \right)$  denoting the expectation under

the  $Q$ -distribution. In addition to previously defined symbols,  $f(b_{n, t, s})$  and  $f(a_{n, t, s})$  denote Gaussian density functions with mean  $\hat{z}_{I_t, n, t, s}$  and unit standard deviation. We use  $\Lambda_{I_t, s}$ , to indicate that  $I_t$  will select a sub matrix of  $\Lambda_s$ .

*Maximising with respect to  $Q(\Lambda_s)$*  results in a product of Gamma distributions over the diagonal terms of the prior precision matrix  $\Lambda_s$ .

$$Q(\Lambda_s[d, d]) = \frac{\hat{h}_{d, s}}{\Gamma(\hat{g}_{d, s})} \Lambda_s[d, d]^{\hat{g}_{d, s}-1} \\ \times \exp(-\Lambda_s[d, d] \hat{h}_{d, s}) \quad (9)$$

where

$$\hat{g}_{d, s} = g_s + \frac{1}{2} \sum_t \sum_{I_t | d_{I_t, t, s} \geq d} Q(I_t)$$

$$\hat{h}_{d, s} = h_s + \frac{1}{2} \sum_t \sum_{I_t | d_{I_t, t, s} \geq d} \left( Q(I_t) \right.$$

$$\left. \times (\hat{\beta}_{I_t, t, s}[d]^2 + \hat{\Lambda}_{I_t, t, s}[d, d]^{-1}) \right)$$

*Maximising with respect to  $Q(\pi)$*  results in a Beta distribution over the binary probability  $\pi$

$$Q(\pi) = \frac{\Gamma(\hat{\delta}_1 + \hat{\delta}_2)}{\Gamma(\hat{\delta}_1)\Gamma(\hat{\delta}_2)} \Theta^{(\hat{\delta}_1-1)} (1 - \Theta)^{(1-\hat{\delta}_2)} \quad (10)$$

where

$$\hat{\delta}_1 = \delta_1 + \sum_t Q(I_t = 1) \text{ and } \hat{\delta}_2 = \delta_2 + \sum_t Q(I_t = 0)$$

*Inferring probabilities of shared gene function* requires maximising the lower bound of the log marginal likelihood in Equation (5) with respect to  $Q(I_t)$  which results in Bernoulli distributions over  $I_t$ .

$$Q(I_t) = \prod_i P_i^{\delta(I_t \equiv i)}, \text{ where } P_i = \frac{\exp(f_{I_t})}{\sum_i \exp(f_i)} \text{ and} \quad (11)$$

$$f_{I_t} = \psi(\hat{\delta}_{I_t}) - \psi(\hat{\delta}) + \sum_s \left( -\frac{1}{2} \log |\hat{\Lambda}_{I_t, t, s}| \right. \quad (12)$$

$$\left. + \frac{1}{2} d_{I_t, t, s} + \frac{1}{2} \sum_{d=1}^{d_{I_t, s}} (\psi(\hat{g}_{d, s}) - \log(\hat{h}_{d, s})) \right.$$

$$\left. - \frac{1}{2} (\hat{\beta}_{I_t, t, s}^T \langle \Lambda_{I_t, s} \rangle \hat{\beta}_{I_t, t, s} + \text{tr} \hat{\Lambda}_{I_t, t, s}^{-1} \langle \Lambda_{I_t, s} \rangle) \right.$$

$$\left. + \sum_n (\log(\Phi(b_{n, t, s}) - \Phi(a_{n, t, s}))) \right.$$

$$\left. - \frac{1}{2} \mathbf{x}_{I_t, n, t, s}^T \hat{\Lambda}_{I_t, t, s}^{-1} \mathbf{x}_{I_t, n, t, s} \right).$$

We use  $I_t$  as index to indicate the conditional dependency of the variables on the model indicator,  $\hat{\beta}_{I_t, t, s}$  for the mode of the  $Q$ -distribution over  $\beta_{I_t, t, s}$

and  $\langle \Lambda_{I_t, s} \rangle = \text{diag} \left( \frac{\hat{g}_{1, s}}{\hat{h}_{1, s}}, \dots, \frac{\hat{g}_{d_{I_t, s}}}{\hat{h}_{d_{I_t, s}}} \right)$  as the expectation of  $\Lambda_{I_t, s}$  under

the  $Q$ -distribution. The expressions  $\Phi(a_{n, t, s})$  and  $\Phi(b_{n, t, s})$  denote Gaussian cdfs with mean  $\hat{z}_{I_t, n, t, s}$  and unit standard deviation at  $a_{n, t, s}$  and  $b_{n, t, s}$ . The latter are implied by the definition of  $P(y_{n, t, s} | z_{n, t, s})$ : for  $y_{n, t, s} = 1$ , we get  $a_{n, t, s} = 0$  and  $b_{n, t, s} = \infty$ ; for  $y_{n, t, s} = 0$ , we get  $a_{n, t, s} = -\infty$  and  $b_{n, t, s} = 0$ . In addition we have  $\psi(x)$  as the digamma function,  $\hat{\delta} = \sum_{I_t} \hat{\delta}_{I_t}$  and  $\text{tr}$  to denote the matrix trace operation. Shared gene function of the  $t$ -th transcript is captured by  $Q(I_t = 1)$ . We provide with  $Q(I_t)$  an approximate measure of shared gene function, which compares a transcript specific model against a common alternative. We can thus assess transcripts relative to each other by transforming all  $Q(I_t)$  into one measure over a  $T$ -dimensional ordinal variable  $G$ .

$$P(G \equiv t | D, X) \approx \frac{Q(I_t \equiv 1)}{Q(I_t \equiv 0)} / \left( \sum_{k=1}^T \frac{Q(I_k \equiv 1)}{Q(I_k \equiv 0)} \right) \quad (13)$$