BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 6 2012, pages 799–806
doi:10.1093/bioinformatics/bts022

Gene expression

Advance Access publication January 11, 2012

Surrogate variable analysis using partial least squares (SVA-PLS) in gene expression studies

Sutirtha Chakraborty*, Somnath Datta and Susmita Datta*

Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40202, USA

Associate Editor: Trey Ideker

ABSTRACT

Motivation: In a typical gene expression profiling study, our prime objective is to identify the genes that are differentially expressed between the samples from two different tissue types. Commonly, standard analysis of variance (ANOVA)/regression is implemented to identify the relative effects of these genes over the two types of samples from their respective arrays of expression levels. But, this technique becomes fundamentally flawed when there are unaccounted sources of variability in these arrays (latent variables attributable to different biological, environmental or other factors relevant in the context). These factors distort the true picture of differential gene expression between the two tissue types and introduce spurious signals of expression heterogeneity. As a result, many genes which are actually differentially expressed are not detected, whereas many others are falsely identified as positives. Moreover, these distortions can be different for different genes. Thus, it is also not possible to get rid of these variations by simple array normalizations. This both-way error can lead to a serious loss in sensitivity and specificity, thereby causing a severe inefficiency in the underlying multiple testing problem. In this work, we attempt to identify the hidden effects of the underlying latent factors in a gene expression profiling study by partial least squares (PLS) and apply ANCOVA technique with the PLS-identified signatures of these hidden effects as covariates, in order to identify the genes that are truly differentially expressed between the two concerned tissue types.

Results: We compare the performance of our method SVA-PLS with standard ANOVA and a relatively recent technique of surrogate variable analysis (SVA), on a wide variety of simulation settings (incorporating different effects of the hidden variable, under situations with varying signal intensities and gene groupings). In all settings, our method yields the highest sensitivity while maintaining relatively reasonable values for the specificity, false discovery rate and false non-discovery rate. Application of our method to gene expression profiling for acute megakaryoblastic leukemia shows that our method detects an additional six genes, that are missed by both the standard ANOVA method as well as SVA, but may be relevant to this disease, as can be seen from mining the existing literature.

Availability: The R code for our method, SVA-PLS, is freely available on the Supplementary website http://www.somnathdatta.org/Supp/SVPLS/

Contact: s0chak10@louisville.edu; susmita.datta@louisville.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

*To whom correspondence should be addressed.

Received on August 5, 2011; revised on November 30, 2011; accepted on January 8, 2012

1 INTRODUCTION

1.1 Background

Differential gene expression analyses in microarray studies typically overlook the important aspect of subject-specific heterogeneity. Subjects in a microarray study can have certain plausible biological profiles which are not known to be connected with the primary outcome of interest and therefore, the subjects may not be matched with respect those profiles in a case–control study. For example, in an expression profiling study with cancer/non-cancer patients the main objective is to identify the genes that are differentially expressed between these two varieties, which can lead to the discovery of potential biomarkers related to cancer. But, this true picture of differential expression can be blurred by several hidden biological effects specific to the subjects recruited in the study. It may happen that some genes are very highly expressed in the subjects with a certain biological, environmental or demographic profile (say, with high blood pressure, regular smoking habits, persons living in rural environments, persons sharing some hidden racial, familial or cryptic pattern pertaining to some inherent structure in the population, etc). On the other hand, some other genes may be repressed because of a similar reason. These factors distort the true signals of differential expression and introduce spurious effects of expression heterogeneity. Thus, many genes which are truly differentially expressed between the two varieties can get rendered as silent, whereas many others may be falsely detected as positives. To complicate things further, we can have a multitude of such hidden confounders in the study and their effects can also vary over different clusters of potentially correlated genes. Thus, it is also not possible to get rid of them by simply modifying the arrays of gene expression measures using a standard normalizing method. These difficulties pose serious problems in analyzing a gene expression data and can lead to erroneous conclusions along with a substantial reduction in the power of the testing procedure.

1.2 Related works

Only a limited number of studies are available in this area, which specifically address this issue of hidden variation in the context of gene expression profiling. With the exception of Leek and Storey (2007), Scheid and Spang (2007) and Listgarten *et al.*, (2010), most of the works in this area have considered specific types of confounding factors that can produce spurious signals of heterogeneity in the context of expression quantitative trait locus

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

(eQTL) mapping. Stegle *et al.* (2008; 2010) have devised methods to improve the power of eQTL studies under the presence of non-genetic confounders (unobserved cell culture conditions, batch effects, etc.). Yu *et al.* (2005), Kang *et al.* (2008a, b), Listgarten *et al.* (2010) discuss the use of linear mixed effect models to correct for confounders from some unknown experimental effects or some hidden population structure. Price *et al.* (2006) proposed the use of principle component analysis (PCA) to correct for some hidden stratification in genome-wide association studies. Scheid and Spang (2007) proposed a method using filtered permutations of the variety labels, which borrows information across the genes to identify and correct for unknown effects of the hidden confounders. Leek and Storey (2007) introduced the surrogate variable analysis (SVA) method and discussed its relevance in gene expression profiling analyses. This is treated as a benchmark technique in comparing the performance of our method. The method considers a singular value decomposition (SVD) of the residual matrix obtained after fitting a simple linear regression model to the log-transformed gene expression data. The significant eigenvectors from the SVD are then used to create a reduced residual matrix (containing statistically significant traces of residual expression heterogeneity). The eigenvectors of the original residual matrix that are maximally correlated with the eigenvectors of this reduced matrix are taken as the surrogate variables. These variables are then used in the original linear model to test for the truly differentially expressed genes. Overall, the method is fairly complex and uses a two step process for the construction of surrogate variables. Moreover, the method in its current form, is also limited in terms of model selection, as it uses a very simple regression framework without considering the effects of each gene and its possible interactions with the surrogate variable (containing effects of the hidden confounders on potentially correlated genes and the two sample varieties). This reduces its applicability to situations where the effect of the hidden confounders can be far more complicated. In essence, all existing techniques in the literature address certain specific patterns of residual expression heterogeneity and discuss relevant modeling techniques to compensate for their effects. In this article, we attempt to excavate the hidden sources of expression heterogeneity by the more generalized approach of partial least squares (PLS). Our method (SVA-PLS), due to its inherent principle, can perform the entire SVA from a more general perspective, by extracting the maximally correlated projections of the residual and original gene expression variables to two different latent factor spaces (connected by a linear relation), thereby ensuring an appropriate estimation of the hidden variables in terms of a set of orthogonal scores in the residual space. Also, our method considers a reasonably wide choice of models, which can potentially explain a large variety of confounding effects.

1.3 Summary of results

The rest of the article is organized as follows. The next section introduces our methodology in detail. Results from a performance study of standard analysis of variance (ANOVA), our method SVA-PLS and the benchmark SVA technique on a variety of simulated gene expression profiling analyses, are reported in Section 3. Finally, a comparative evaluation of the three methods is performed on a real-life dataset from the gene expression profiling study of acute megakaryoblastic leukemia (AMKL). The article ends with a discussion in Section 5.

2 METHODS

We consider a gene expression profiling analysis with g genes and n subjects, distributed over two tissue types/varieties (like, normal and cancer cell lines or two different biological conditions). Let the first n_1 subjects be under variety 1 and the rest n_2 be under variety 2. We start by applying the standard ANOVA technique on the log-transformed gene expression matrix Y (Kerr *et al.*, 2000, Kerr and Churchill, 2001, Wolfinger *et al.*, 2001 and Kerr *et al.*, 2002) and compute the fitted model residuals. Let Y_{ijk} denote the log-transformed gene expression value for the gene i in subject k under variety j, $i=1,2,\cdots,g$, $j=1,2$ and $k=1,2,\cdots,n_1$ for $j=1$ and $k=n_1+1,\cdots,n$ for $j=2$. We fit the following ANOVA model to the data and get the residuals.

$$Y_{ijk}=\mu+G_i+V_j+(GV)_{ij}+\epsilon_{ijk}, \tag{1}$$

where μ denotes the general mean effect in the model, G_i, V_j, respectively, stand for the main effects of gene i and variety j and $(GV)_{ij}$ defines their interaction (characterizing the expression effect of gene i on the subjects under variety j). ϵ_{ijk} denotes the random error term which is assumed to follow a $N(0,\sigma^2)$ distribution.

The fitted residuals from the above model are then given by $e_{ijk}=Y_{ijk}-\bar{Y}_{ij0}$, where $\bar{Y}_{ij0}=\frac{1}{n_j}\sum_{k\in A_j} Y_{ijk}$, A_j being the set of individuals corresponding to variety j. These residuals may contain the traces of subject-specific expression heterogeneity, which is independent of the primary variable signal from the sample types and can confound the true effect behind many potentially positive genes or can overestimate many silent genes as positives. In order to extract these spurious differential expression signals, we employ the PLS technique [Wold (1975, 1985); Helland (1999)]. We construct two $n\times g$ matrices Y and E, whose i-th column contains, respectively, the original log-transformed gene expression levels Y_{ijk} s and the residual gene expression levels e_{ijk} s for all the n individuals corresponding to gene i $(i=1,2,\ldots g)$. Thus, $E=((E_{rc}))_{n\times g}$ and $Y=((Y_{rc}))_{n\times g}$, $r=1,2,\ldots n$ and $c=1,2,\ldots g$. Conceptually, these matrices can be characterized as two sets of n observations on two g-dimensional random variables E and Y, where each dimension corresponds to a certain gene.

Our approach is to regress E on Y by PLS, in order to extract the hidden sources of gene expression heterogeneity. PLS, by virtue of its dimension reduction and covariance maximizing property extracts the additional signals from those groups of genes, whose expression levels, contained in the original gene expression matrix Y, are influenced by the hidden subject-specific effects, contained in the residual gene expression matrix E. Let the matrices now stand for their respective mean zero versions, obtained by subtracting the respective column means from their initial versions. We assume that E^TY is non-null. The statistical regression model for PLS can be written as

$$E = UQ^T+\epsilon_{\mathbf{1}}, \tag{2}$$

$$Y = TP^T+\epsilon_{\mathbf{2}}, \tag{3}$$

where $U=[u_1:u_2:\cdots:u_m]$ is an $n\times m$ matrix, containing the m latent factors $u_1,u_2,\cdots:u_m$ in the space of the response matrix E. Similarly, $T=[t_1:t_2:\cdots:t_m]$ is another $n\times m$ matrix, containing the m latent factors $t_1,t_2,\cdots t_m$ in the space of the covariate matrix Y. $Q=[q_1:q_2:\cdots:q_m]$ is an $g\times m$ matrix, consisting of the loadings $q_1,q_2,\cdots q_m$, which measure respectively, the importance of the latent factors $u_1,u_2,\cdots u_m$ in the response (E)'s space. Similarly, $P=[p_1:p_2:\cdots:p_m]$ is a $g\times m$ matrix, consisting of the loadings $p_1,p_2,\cdots p_m$, which measure, respectively, the importance of the latent factors $t_1,t_2,\cdots t_m$ in the covariate (Y)'s space. Further, for each $i=1,2,\cdots m$, $u_i=(u_{i1},u_{i2},\cdots,u_{in})^T$, $t_i=(t_{i1},t_{i2},\cdots,t_{in})^T$, $q_i=(q_{i1},q_{i2},\cdots,q_{ig})^T$ and $p_i=(p_{i1},p_{i2},\cdots,p_{ig})^T$. Here, $\epsilon_{\mathbf{1}}$ and $\epsilon_{\mathbf{2}}$ are the random error matrices characterizing the residual terms in the regression models for E and Y, respectively.

Now, the basic idea of PLS is to estimate the set of latent factor pairs $(u_1,t_1),(u_2,t_2)\cdots(u_m,t_m)$, one by one, along with the corresponding deflation

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

of the matrices E and Y at each step. This is executed by a process of alternating regression. For each latent factor pair (u_i, t_i), $i=1,2,\ldots m$, this procedure finds weight vectors c and w in such a way that the covariance of u_i and t_i is maximized. Specifically, c and w are such that $[cov(u_i,t_i)]^2 = [cov(Ec,Yw)]^2 = max_{\|r\|=\|s\|=1}[cov(Er,Ys)]^2$.

We initialize $E_1 = E$ and $Y_1 = Y$. Now for $i=1,2,\ldots,m$, we successively estimate the i-th latent factor pair (u_i,t_i), by the PLS algorithm presented below [see e.g. Abdi (2003); Rosipal and Krämer (2006); Mevik and Wehrens (2007)]. In this algorithm, we use $a \propto b$, to mean $a = b/\|b\|$, for any vector b.

We start by setting $u_i = E_{i,.v}$, where $v = argmax_c \sum_{r=1}^{n} E_{i,rc}^2$, $t_{i,\text{old}} = (0,0,\cdots,0)$ and repeat Steps (i) to (iv) till convergence [as defined in Step (v)]:

(i) Regress Y_i on u_i to obtain $w_i \propto Y_i^T u_i$.

(ii) Compute the updated i-th, Y-space latent factor $t_i = Y_i w_i$.

(iii) Regress E_i on t_i to obtain $c_i \propto E_i^T t_i$.

(iv) Compute the updated i-th, E-space latent factor $u_i = E_i c_i$.

(v) If $\sum_{j=1}^{n} |t_{ij} - t_{ij,\text{old}}|/t_{ij} < \epsilon$, STOP; otherwise let $t_{i,\text{old}} = t_i$ and go back to Step (i). Throughout, we have used $\epsilon = 10^{-8}$.

Next deflate the matrices E_i and Y_i to obtain $E_{i+1} = E_i - t_i b_i^T$ and $Y_{i+1} = Y_i - t_i p_i^T$, where, $b_i = E_i^T t_i / t_i^T t_i$ and $p_i = Y_i^T t_i / t_i^T t_i$. E_{i+1} and Y_{i+1} are now used in place of E_i and Y_i to extract the $(i+1)$-th latent factor pair (u_{i+1}, t_{i+1}). In this way, we find the m latent factors from the E and Y spaces. The use of t in deflating both the response (E) as well as covariate (Y) matrices ensures orthogonality of the extracted latent factors $t_1, t_2, \ldots t_m$ in the Y-space, which in turn ensures their estimability in a linear model. From now onwards, we denote m by $p_{\max}$ to define the maximum number of hidden(latent) factors (scores) that are needed to be extracted from the two spaces. The $p_{\max}$ Y-space scores extracted by the above method can be characterized as a set of surrogate variables $Z^1, Z^2, \cdots, Z^{p_{\max}}$ that are optimally associated with the latent factors from the E-space, containing the hidden sources of expression heterogeneity in the original gene expression data. The mutual covariances between the extracted latent factors from the two spaces decrease gradually from the first pair (u_1, Z^1) to the $p_{\max}$-th pair $(u_m, Z^{p_{\max}})$. Thus, Z^1 contains maximum information on the residual gene expression heterogeneity compared the other factors. Now, we define a series of ANCOVA models M_p indexed by $p = 1,2,\cdots,p_{\max}$, where p denotes number the surrogate variables incorporated in the model, which capture effects of the residual gene expression heterogeneity.

$$M_p : Y_{ijk} = \mu + G_i + V_j + (GV)_{ij} + W_{ijk}^{\text{imp}} + \epsilon_{ijk}, \tag{4}$$

where μ denotes the general mean effect in the model, G_i, V_j and $(GV)_{ij}$ denote the main effects of gene i, variety j and their interaction effect. $W_{\text{ijk}}^{\text{imp}} = \sum_{l=1}^{p} \beta_l Z_{jk}^l + (GZ^1)_i Z_{jk}^1 + (VZ^1)_j Z_{jk}^1$, is incorporated in the model as the PLS-imputed estimate of the hidden residual expression heterogeneity in the data. Here, β_l is the regression coefficient for Z^l in the ANCOVA model (4). $(GZ^1)_i$ and $(VZ^1)_j$ define, respectively, the interaction effects of gene i and variety j with the first surrogate variable Z^1. These effects measure respectively, the variation in the impact of the hidden factors (captured by Z^1) over different groups of genes (which may be correlated) and over the two tissue types (which may affect the primary variable signal). As the first surrogate variable Z^1 contains maximum information on the residual expression heterogeneity compared with the other ones, we consider only its interactions with the gene and variety effects. The inclusion of these effects in the model ensures accurate estimation of the actual gene–variety interactions, capturing the true expression effects of a gene over the two varieties, if potential hidden variables are embedded in the data structure. ϵ_{ijk} denotes the random error term corresponding to Y_{ijk} in the model, which is assumed to follow a $N(0,\sigma^2)$ distribution. Here $p_{\max}$ can be specified by the user, considering the corresponding situation under study and affording a reasonable degree of complexity along with a manageable computational intensity. As for our purpose, we have selected $p_{\max} = 3$, since from several empirical studies (details reported in the Supplementary Material) we have found that the first three surrogate variables (Z^1, Z^2, Z^3) explain a substantial proportion of the dispersion for the variable E. Thus, overall we consider three different linear models from which the best is selected by the Akaike's Information Criterion (AIC) (Hirotsugu, 1974, 1980) and is then used to test for the equality of gene–variety interaction effects for identifying the truly differentially expressed genes. In the concerned multiple testing problem, we use the Benjamini–Hochberg method (Benjamini and Hochberg, 1995) to control the false discovery rate. The entire algorithm for our method SVA-PLS is presented below:

- *Step 1*: fit the standard ANOVA model (1) to the log-transformed gene expression data Y and calculate the fitted residual matrix E.
- *Step 2*: regress E on Y by PLS and extract $p_{\max}$ (user-specified) linear combinations (scores) from their respective latent factor spaces.
- *Step 3*: incorporate, one by one, the $p_{\max}$ scores in the surrogate variables along with the gene and variety interactions of the first PLS score in model (1) to develop a series of $p_{\max}$ new linear models (4).
- *Step 4:* compare AIC's for the models to select the best out of them (the model corresponding to the minimum AIC) and denote its corresponding number of surrogate variables by p_{opt}.
- *Step 5*: fit model $M_{p_{\text{opt}}}$ to estimate the actual gene–variety interaction effect $(GV)_{ij}$ for each gene i and variety j $(i=1,2,\cdots,g$ and $j=1,2)$. For each gene i, test the null hypothesis of no variety-specific differential expression $H_0 : (GV)_{i1} = (GV)_{i2}$ versus alternative hypothesis of differential expression $H_0 : (GV)_{i1} \neq (GV)_{i2}$, using the statistic t_i:

$$t_i = \frac{\widehat{(GV)}_{i1} - \widehat{(GV)}_{i2}}{\sqrt{\hat{\sigma}^2 * \widehat{Var}(\widehat{(GV)}_{i1} - \widehat{(GV)}_{i2})}} \tag{5}$$

  which under H_0 follows a central t distribution with $\nu = g*n - 3*g - p_{\text{opt}}$ df and the corresponding P-value is $2*(1 - F_\nu^t(|t_i|))$, $F_\nu^t()$ being the distribution function for a central t distribution with ν df. $\widehat{(GV)}_{ij}$, $j=1,2$ is the least squares estimate of $(GV)_{ij}$, $\widehat{Var}(\widehat{(GV)}_{i1} - \widehat{(GV)}_{i2})$ is the estimated variance of $\widehat{(GV)}_{i1} - \widehat{(GV)}_{i2}$ and $\hat{\sigma}^2$ is the least squares estimate of σ^2, all computed from the model $M_{p_{\text{opt}}}$.
- *Step 6*: perform a multiple testing with these P-values for identifying the truly differentially expressed genes at a prespecified level of the false discovery rate (FDR), using the Benjamini–Hochberg method (Benjamini and Hochberg, 1995).

3 SIMULATION STUDIES

We envisage a gene expression profiling study with 500 genes and 20 subjects, distributed equally over 2 varieties. The entire simulation study is broadly divided into two settings: (i) assuming the genes to be independent of each other (Independent) and (ii) assuming dependence within different groups of genes (Clustered).

The log-transformed gene expression values (Y) are generated by using a linear model with the gene (G) and variety/tissue type (V) main effects, their interaction (GV) and a hidden variable (W). Thus, Y_{ijk}, corresponding to the i-th gene, j-th variety and k-th subject, is obtained as:

$$Y_{ijk} = \mu + G_i + V_j + (GV)_{ij} + \gamma W_{ijk} + \epsilon_{ijk}, \tag{6}$$

where $i = 1,2,\ldots,500$ denote the 500 genes, $j = 1,2$ denote the two varieties and $k \in A_j$, denote the 20 subjects in the study. Here $A_1 = \{1,2,\ldots 10\}$ and $A_2 = \{11,12,\ldots,20\}$ denote, respectively, the subsets of individuals corresponding to the two varieties 1 and 2. The error terms ϵ_{ijk} s are assumed to be independently distributed as $N(0,\sigma^2)$, where choice of σ^2 is described next.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Let X denote the design matrix corresponding to the above linear model, β denote the corresponding vector of regression coefficients and ϵ denote the vector of the corresponding random error terms. Then we have $Y=X\beta+\epsilon$. Define the noise to signal ratio (η) as $\eta=\sigma^2/\beta^T Var(X)\beta$ [σ^2 being the random error variance and $\beta^T Var(X)\beta$ being the variance of the signal $X\beta$ generating the actual gene expression levels]. This quantity measures the relative intensity of the noise coming from the random error and the confounded primary variable signal depicting the expression effect of the genes over the two varieties. We consider three different values 0.1, 0.5 and 1 for η to incorporate, respectively, the cases of strong, moderate and weak primary signal intensity. From these choices of the η, we compute the corresponding values of σ^2 and use them to simulate the values of ϵ_{ijk} s in the model (6).

For data generation, we assume the effects of all terms except the gene–variety interaction (GV) and the hidden confounder (W) to be zero. Overall we consider the first 70 genes to be truly differentially expressed among all the 500 genes. For $1\le i\le 20$, we take $(GV)_{i1}=-7.5$, $(GV)_{i2}=7.5$, for $21\le i\le 70$ $(GV)_{i1}=7.5$, $(GV)_{i2}=-7.5$ and for genes 71–500 we assume $(GV)_{i1}=(GV)_{i2}=0$.

For each gene i in 1–500, $j=1,2$ and subject $k\in A_j$, we generate a Bernoulli random variable s_{ijk} with success probability of 0.4. It is used to generate effects of W over the two varieties, under both the independent as well as clustered settings. Biologically, this accounts for hidden confounding effects from certain specific subjects under each of the two varieties, which is typically expected in a real-life gene expression analysis. In addition, we consider two separate scenarios, depending on whether the effect of the hidden variable W is same or different over the two varieties.

3.1 Independent tests

In this setting, we consider the genes to be independent of one another. We generate their log-transformed expression levels under two scenarios of similar and varying effects of the hidden variable over the two varieties.

The similarity in the effects of the missing variable over the two varieties is accomplished by simulating the latent variable W_{ijk} from the same normal distribution for $k\in A_1\cup A_2$ (covering subjects from both the varieties). This represents a case when the hidden variable W is orthogonal to the variety effect. The effect of W is varied over three different groups of genes by changing the mean parameter of its distribution. That is, we let $W_{ijk}=Z_{ijk}I(s_{ijk}=1)$, where Z_{ijk} is generated from $N(-3,0.01)$ or $N(2,0.01)$ or $N(10,0.01)$, depending on whether $1\le i\le 20$, $21\le i\le 70$ or $i>70$.

For generating different effects of the hidden variable over the two varieties, we simulate the latent variable W_{ijk} for the subjects $k\in A_1$ and $k\in A_2$, from two normal distributions with different means. This represents a case when the hidden variable W is confounded with the variety effect. Once again, the effect of the hidden variable is varied over the three gene groups. That is, we let $W_{ijk}=Z_{ijk}I(s_{ijk}=1)$, where for $k\in A_1$, Z_{i1k} is generated from $N(-3,0.01)$ or $N(2,0.01)$ or $N(10,0.01)$ and for $k\in A_2$, Z_{i2k} is generated from $N(3,0.01)$ or $N(15,0.01)$ or $N(20,0.01)$, depending on whether $1\le i\le 20$, $21\le i\le 70$ or $i>70$.

Next, we consider yet another simulation setting, where the hidden variable results in a complex confounding pattern with the varieties. In this case, for variety 1, we simulate the latent variable W_{i1k} for the subjects $k=1,2,\ldots 5$ and $k=6,7,\ldots 10$, from two normal distributions with different means. Analogously, for variety 2, we simulate the latent variable W_{i2k} for the subjects $k=11,12,\ldots 15$ and $16,17,\ldots 20$, from two normal distributions with different means. Similar to the previous settings, we vary the effect of W over the three gene groups. Thus, under the variety 1, we let $W_{i1k}=Z_{i1k}I(s_{i1k}=1)$, where for $k=1,2,\ldots 5$, Z_{i1k} is generated from $N(-3,0.01)$ or $N(2,0.01)$ or $N(10,0.01)$ and for $k=6,7,\ldots 10$, Z_{i1k} is generated from $N(3,0.01)$ or $N(15,0.01)$ or $N(20,0.01)$, depending on whether $1\le i\le 20$, $21\le i\le 70$ or $i>70$. Similarly, under variety 2, we let $W_{i2k}=Z_{i2k}I(s_{i2k}=1)$, where for $k=11,12,\ldots 15$, Z_{i2k} is generated from $N(-6,0.01)$ or $N(2,0.01)$ or $N(10,0.01)$ and for $k=16,17,\ldots 20$, Z_{i2k} is generated from $N(5,0.01)$ or $N(15,0.01)$ or $N(15,0.01)$, depending on whether $1\le i\le 20$, $21\le i\le 70$ or $i>70$.

3.2 Cluster-dependent tests

Note that the usual statistical model for differential gene analysis by the ANOVA formulation assumes independent error terms. However, it is well known that in reality, certain groups of genes have correlated expressions. In this setting, we consider three clusters of correlated genes with the same gene–variety interaction effects as for the case of independently expressed genes, with the hidden variable (W) being generated according to the same set of simulation schemes as in Section 3.1.

The underlying dependence among the genes is incorporated by generating the random error term ϵ_{ijk} in model (6) as a weighted sum of two different errors ϵ^1 and ϵ^2, simulated independently of each other, with the values of ϵ^1 being same for all the genes in the same cluster. Let $C=(1,2,\ldots,20;51,52,\ldots,70;461,462,\ldots,500)$ denote the union of the three clusters. Then, mathematically the generation of e_{ijk} in the simulation model (6) is expressed as:

$$\epsilon_{ijk} = \frac{1}{\sqrt{2}}\epsilon^1_{I(i),j,k}+\frac{1}{\sqrt{2}}\epsilon^2_{i,j,k} \text{ if } i\in C, \quad (7)$$

$$= \epsilon^2_{i,j,k} \text{ if } i\notin C, \quad (8)$$

where, $I(i)$ denotes the cluster containing gene i. The random error terms $\epsilon^1_{I(i),j,k}$ and $\epsilon^2_{i,j,k}$ are generated from independent $N(0,\sigma^2)$ distributions (σ^2 being determined from the desired noise to signal ratio, as before). From a biological perspective, this simulation setting captures the idea that genes in the same cluster act cooperatively, resulting in correlated expression measurements.

The simulation study is concerned with a performance analysis of standard ANOVA, our method SVA-PLS and SVA (downloaded from www.bioconductor.org on November 29, 2011) with respect to four measures: sensitivity, specificity, FDR and false non-discovery rate (FNR).

(1) Sensitivity: proportion among differentially expressed genes that were declared significant.

(2) Specificity: proportion among non-differentially expressed genes that were declared non-significant.

(3) FDR: proportion among genes declared significant that were not differentially expressed.

(4) FNR: proportion among genes declared non-significant that were differentially expressed.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 1. Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of independently expressed genes with similar effects of the hidden variable over the two varieties

Method	Sensitivity	Specificity	FDR	FNR
	$\eta=0.1$			
Std. ANOVA	0.319	1	0	0.059
SVA-PLS	1	0.986	0.078	0
SVA	0.999	0.993	0.040	0.000
	$\eta=0.5$			
Std. ANOVA	0.154	1	0	0.078
SVA-PLS	0.813	0.993	0.047	0.029
SVA	0.640	0.995	0.045	0.055
	$\eta=1$			
Std. ANOVA	0.104	1	0.000	0.071
SVA-PLS	0.370	0.997	0.045	0.091
SVA	0.194	0.998	0.041	0.109

Table 2. Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of independently expressed genes with different effects of the hidden variable over the two varieties

Method	Sensitivity	Specificity	FDR	FNR
	$\eta=0.1$			
Std. ANOVA	0.176	1	0	0.031
SVA-PLS	0.926	0.990	0.061	0.012
SVA	0.290	0.997	0.045	0.035
	$\eta=0.5$			
Std. ANOVA	0.086	1	0	0.034
SVA-PLS	0.539	0.995	0.048	0.065
SVA	0.208	0.998	0.046	0.084
	$\eta=1$			
Std. ANOVA	0.026	1	0	0.027
SVA-PLS	0.251	0.997	0.052	0.095
SVA	0.080	0.999	0.058	0.074

The entire simulation study is performed 100 times under each scenario in order to compute the average values of the four performance measures. Under the clustered setting, the SVA software broke down at several iterations of the simulation study. Hence, for this setting we only report the performance of our method (SVA-PLS) and the standard ANOVA.

From the performance analysis of the three methods on the independent gene expression levels, with similar, varying and complex effects of the hidden variable (Tables 1, 2 and 3, respectively), we see that SVA-PLS achieves the highest sensitivity compared with standard ANOVA and SVA. Interestingly, the margin of sensitivity for our method is very high in the case of complex confounding (Table 3), followed by the case of varying effects of the hidden variable over the tissue types (Table 2). This observation demonstrates that our method is most useful for the relatively complicated situations, when the missing variable is in fact a statistical confounder affecting the primary variable signals from the two tissue types. In addition, our method produces a comparatively impressive performance with respect to the other two methods in

Table 3. Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of independently expressed genes, when the hidden variable has a complex differential pattern between the two varieties, resulting in a serious confounding

Method	Sensitivity	Specificity	FDR	FNR
	$\eta=0.1$			
Std. ANOVA	0.174	1	0	0.025
SVA-PLS	0.870	0.999	0.008	0.020
SVA	0.152	0.999	0.047	0.096
	$\eta=0.5$			
Std. ANOVA	0.047	1	0	0.028
SVA-PLS	0.269	0.999	0.023	0.095
SVA	0.027	0.999	0.067	0.051
	$\eta=1$			
Std. ANOVA	0.038	1	0	0.039
SVA-PLS	0.105	0.999	0.022	0.087
SVA	0.011	1	0.027	0.041

Table 4. Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of co-regulated genes with similar effects of the hidden variable over the two varieties

Method	Sensitivity	Specificity	FDR	FNR
	$\eta=0.5$			
Std. ANOVA	0.234	1	0	0.070
SVA-PLS	0.989	0.991	0.054	0.002

Table 5. Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of co-regulated genes with different effects of the hidden variable over the two varieties

Method	Sensitivity	Specificity	FDR	FNR
	$\eta=0.5$			
Std. ANOVA	0.100	1	0	0.030
SVA-PLS	0.747	0.993	0.052	0.039

terms of the high specificity and reasonably small FDR and FNR. Under the clustered setting with dependence inside several clusters of genes (Tables 4, 5 and 6 for the moderate case of $\eta=0.5$), SVA-PLS performs really well compared with standard ANOVA by detecting a larger number of truly positive genes with its high margin of sensitivity, at the cost of a slightly increased FDR, which is an obvious price to pay for achieving a higher performance in terms of detection power. For the other choices of the η too, SVA-PLS shows higher sensitivity compared with standard ANOVA (we refer to the Supplementary Material). Under this setting also, our method yields a reasonably high specificity in comparison to standard ANOVA along with an impressively small FDR and FNR. Specifically, the margin of sensitivity for SVA-PLS under both the simulation settings is the highest in the best case with very strong primary variable signal ($\eta=0.1$), closely followed by the moderate ($\eta=0.5$) and worst cases ($\eta=1$). Thus, overall the results demonstrate that our method, by

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Table 6. Performance analysis of standard ANOVA, SVA-PLS and SVA under the setting of co-regulated genes, when the hidden variable has a complex differential pattern between the two varieties, resulting in a serious confounding

Method	Sensitivity	Specificity	FDR	FNR
	$\eta = 0.5$			
Std. ANOVA	0.105	1	0	0.027
SVA-PLS	0.618	0.998	0.019	0.055

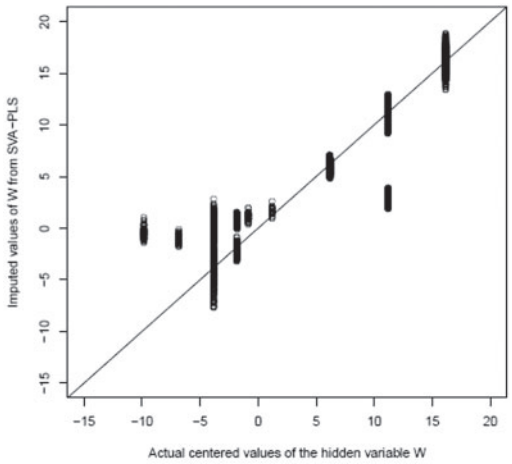

Fig. 1. Plot of the PLS-imputed values of W versus the actual centered values under the setting of independently expressed genes, when the hidden variable has a complex differential pattern between the two varieties, resulting in a serious confounding.

virtue of its high sensitivity in a wide variety of situations can potentially discover many truly differentially expressed genes that are masked by the effects of hidden factors and can simultaneously maintain acceptably small error rates.

We further illustrate the efficacy of our method by comparing the actual (mean centered) values of the hidden variable W_{ijk} (simulated in the model under the setting of independently expressed genes with serious confounding of the hidden variable W), with the PLS-imputed values W_{ijk}^{imp}, incorporated in the ANCOVA model (4). We observe a strongly linear relationship between the two sets of values with a very high positive correlation (0.95) (Fig. 1). We have noticed a similar effect in the other simulation settings as well. This demonstrates that our method SVA-PLS is effectively imputing the hidden variable (W) on the actual expression levels of the genes.

4 ANALYSIS OF LEUKEMIA DATA

We now explore the performance of our method on a dataset generated from a gene expression study of AMKL, which is a subtype of the disease acute myeloid leukemia (AML). The dataset was featured in Bourquin *et al*. (2006). It contains the expression levels of 22283 genes on two types of AMKL patients, 23 with down-syndrome and 38 without down-syndrome.

In general, down syndrome patients are more prone to AMKL compared those without it and treatment outcomes are also much more favorable for them. From an exploratory analysis of the dataset (Bourquin *et al*., 2006), it was found that the non-DS AMKL patients can be further subdivided into two groups using the expression profiles of the HOX/TALE family members. This latent grouping inside one tissue type can generate hidden confounders, which may in turn perturb the actual signals of variety-specific differential gene expression. Thus, it is important to search for the traces of residual gene expression heterogeneity in this dataset for ensuring a more accurate inference on the truly positive genes, which is built into our method. Indeed, we investigated whether the PLS-imputed values W^{imp} of the three genes, *HOXA9*, *HOXA10* and *MEIS1*, belonging to the HOX/TALE family, contain a subgroup signature. Figure 2 shows a heat map for the normalized values of the estimated PLS contributed part W^{imp}, corresponding to the 38 individuals in the non-DS AMKL group. This W^{imp} is free from the primary signal of variety-specific differential expression and is expected to contain the traces of residual expression heterogeneity corresponding to the hidden factors in the data. From Figure 2, we can observe a subgroup structure among these individuals. Clearly, the differential pattern is strongest for the *MEIS1* gene, followed by *HOXA9* and *HOXA10*.

The three methods SVA-PLS, standard ANOVA and SVA were applied to the log-transformed expression matrix of the 22 283 genes in the dataset. Overall, SVA-PLS detected 1585 genes followed by 1407 genes from standard ANOVA and 280 genes from SVA (Fig. 3). Our method detects a total of 427 genes, that are missed by others, of which at least six genes deserve special mention. These genes are *MLF1*, *BRCA2*, *TNF*, *c-MPL*, *CD44* and *MAGE-D4*.

The gene *MLF1* is actively involved in the development of AML. A chromosomal derangement associated with this gene is a cause of the myelodysplastic syndrome (MDS) (Block *et al*., 1953). Patients with this syndrome often develop acute anemia, which in most cases lead to low blood counts. In almost one-third of the patients, this syndrome causes progressive bone marrow failure, which in turn develops the disease into AML. Also, delayed bone marrow transplantation for patients with low risk of the MDS has been found to be connected with the improved outcome (Cutler *et al*., 2004).

The gene *BRCA2* is an important caretaker gene (Kinzler and Vogelstein, 1997), whose inactivation initiates a tumor and the resulting genetic instability causes accelerated mutation in all genes, which in turn, may lead to the rapid progression of the tumor. Germline mutations in this gene play a dominant role in the onset of breast and ovarian cancer, pancreatic cancer, prostate cancer, Fanconi anemia and pre-B-cell acute lymphoblastic leukemia (Lancaster *et al*., 1996, Murphy *et al*., 2002, Narod *et al*., 2008, Özcelik *et al*., 1997, Wagner *et al*., 2004).

The apoptosis-inducing ligand TRAIL related to the gene *TNF*, plays an active role in the development of different types of cancers. Downregulation of *TRAIL-R2* inhibits the TRAIL-mediated apoptosis in AML (Riccioni *et al*., 2005). Monoallelic deletion of the tumor-suppressing genes *TRAIL-R1* and *TRAIL-R2* can inactivate the TRAIL-induced apoptosis in B-cell lymphoma (Rubio-Moscardo *et al*., 2005). In the development of colorectal cancer, there is a substantial increase in sensitivity to TRAIL-induced apoptosis, with the progression from benign to malignant tumors (Haque *et al*., 2005).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

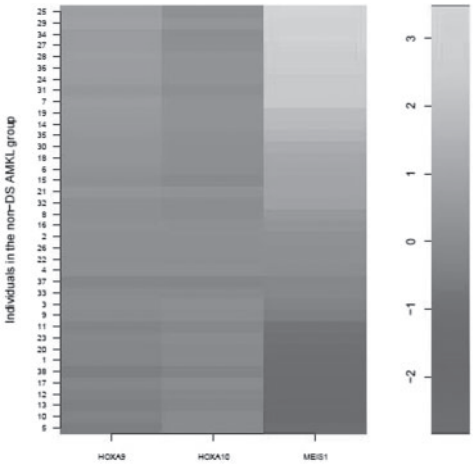

Fig. 2. Heat map of the PLS imputed W^{imp} for the three HOX/TALE family genes in the individuals under the non-DS AMKL variety showing a subgroup structure.

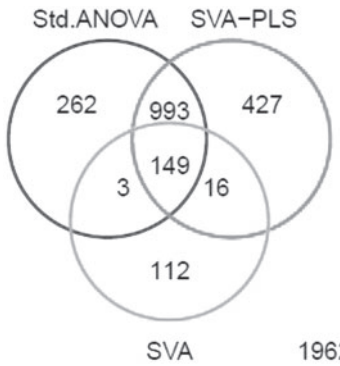

Fig. 3. Venn Diagram showing the number of significant genes detected from the AMKL data by Standard ANOVA, SVA-PLS and SVA.

Expression of the gene *c-MPL* has been found to be involved in the progression of CD34+ and M2FAB subtypes of AML (Ayala *et al*., 2009).

Ligation of the gene *CD44* with specific anti-*CD44* antibodies (or with its natural ligand hyaluronan) can reverse the blockage in the differentiation of several subtypes of AML, thereby improving the survival of patients using differentiating agents (e.g. retinoic acid) (Charrad *et al*., 1999). The 8:21 chromosomal translocation is commonly observed in AML. Acute myeloid leukemia-1 transcription factor AML1-ETO and its splice variant AML1-ETO9a are capable of modulating the expression of *CD44*, thereby connecting the abnormal translocation 8:21 to the regulation of a cell adhesion molecule, that is involved in the nurturing of AML blast/stem cells (Peterson *et al*., 2007). In the acute promyelocytic leukemia cell line NB4, overexpression of the gene *CD44* receptor results in apoptosis (Abecassis *et al*., 2008). In addition, downregulation of this gene has been found to be conducive to keratoacanthoma and squamous cell carcinoma (Tataroglu *et al*., 2007).

Upregulation of the gene *MAGE-D4* results in the proliferation of tumor cells in non-small cell lung cancer (NSCLC) (Ito *et al*., 2006).

Thus, we find that a number of the additional genes selected by our method are connected to AML or some other related type of carcinoma. These genes being found to be differentially expressed between the subjects with and without down syndrome, can serve as important candidates for research on leukemia and down syndrome.

5 DISCUSSION

Hidden array-specific (subject-specific) factors in microarray analyses may constitute a substantial source of gene expression heterogeneity. The effects of these factors are not detectable from outside and also cannot be removed by any standard normalizing method. But they can perturb the primary signals of differential gene expression and lead to erroneous conclusions on the detection of differentially expressed genes.

This problem is relatively unexplored in gene expression studies. In this article, we have developed a novel technique for identifying these latent factors by using PLS and applied it to a wide variety of simulation settings characterizing different patterns of viable gene expression profiling studies. We have shown that the technique of PLS, by virtue of its basic principle of projecting to latent structures, can produce precise estimates of the hidden factors causing the spurious signal heterogeneity. These estimates (surrogate variables) when incorporated in the ANOVA model enhances detection of the gene–variety interaction effects thereby leading to a large gain in sensitivity of the underlying bioinformatics screening procedure. The resulting method, SVA-PLS, also yields a reasonably high specificity for a wide range of data structures, thereby ensuring an efficient control over the incorrect detection of many silent genes. The FDR is marginally higher for our method, but is sufficiently well compensated by a substantially large gain in the margin of sensitivity. Overall, SVA-PLS emerges as the winner when compared with two other competing methods in a range of controlled settings. The utility of our method in detecting potentially interesting genes missed by other methods is also demonstrated by an analysis of a real dataset on AMKL patients.

Unaccounted sources of variation (hidden variables) in a model can adversely affect the outcomes of statistical tests. This is particularly true if the unmeasured variables are confounders, i.e. correlated with the variables in the model whose effects on the outcomes are being tested. In a simple two group comparison, a standard assumption for the validity of the commonly used two sample pooled *t*-test is that the error variances in the two groups (populations) are equal. A departure from this model assumption is known as the Behrens–Fisher problem and has received a great deal of assumption in the statistics literature [see e.g. Lehmann (1986)]. A common solution to this problem is to use separate variance estimates for the error distribution in two groups and resort either to an approximate *t*-distribution (Welch, 1938) or to a large sample normal approximation of the distribution of the test statistics. Indeed, if one assumes (as in our formulation) the existence of an unmeasured factor contributing to the outcome in a linear model formulation of the two sample problem that is equated with the model errors, one gets a model with unequal variances in the two groups. Thus, our method may provide an alternative solution to

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

the Behrens–Fisher problem. We plan to explore this connection in greater details elsewhere.

ACKNOWLEDGEMENTS

We sincerely thank three anonymous reviewers for their constructive comments which lead to an improved manuscript.

Conflict of Interest: none declared.

REFERENCES

Abdi,H. (2003) Partial least squares regression (PLS-regression). In Lewis-Beck,M. *et al.* (ed.) *Encyclopedia for Research Methods for the Social Sciences*. Sage, Thousand Oaks, CA.

Abecassis,I. *et al*. (2008) Re-expression of DNA methylation-silenced CD44 gene in a resistant NB4 cell line: rescue of CD44-dependent cell death by cAMP. *Leukemia*, **22**, 511–520.

Ayala,R.M. *et al*. (2009) Clinical significance of Gata-1, Gata-2, EKLF, and c-MPL expression in acute myeloid leukemia. *Am. J. Hematol.*, **84**, 79–86.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, **57**, 289–300.

Block,M. *et al*. (1953) Preleukemic acute human leukemia. *JAMA*, **152**, 1018–1028.

Bourquin,J.-P. *et al*. (2006) Identification of distinct molecular phenotypes in acute megakaryoblastic leukemia by gene expression profiling. *Proc. Natl Acad. Sci.* USA, **103**, 3339–3344.

Charrad,R.-S. *et al*. (1999) Ligation of the CD44 adhesion molecule reverses blockage of differentiation in human acute myeloid leukemia. *Nat. Med.*, **5**, 669–676.

Cutler,C.S. *et al*. (2004) A decision analysis of allogeneic bone marrow transplantation for the myelodysplastic syndromes: delayed transplantation for low-risk myelodysplasia is associated with improved outcome. *Blood*, **104**, 579–585.

Haque,A. *et al*. (2005) Increased sensitivity to TRAIL-induced apoptosis occurs during the adenoma to carcinoma transition of colorectal carcinogenesis. *Br. J. Cancer*, **92**, 736–742.

Helland,I.S. (1999) Some theoretical aspects of partial least squares regression. *Chemometr. Intell. Lab. Syst.*, **58**, 97–107.

Hirotsugu,A. (1974) A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **19**, 716–723.

Hirotsugu,A. (1980) Likelihood and the Bayes procedure. *Bayesian Stat.*, **166**, 143–166.

Ito,S. *et al*. (2006) Expression of MAGE-D4, a novel MAGE family antigen, is correlated with tumor-cell proliferation of non-small cell lung cancer. *Lung Cancer*, **51**, 79–88.

Kang,H.M. *et al*. (2008a) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709–1723.

Kang,H.M. *et al*. (2008b) Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, **180**, 1909–1925.

Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961–8965.

Kerr,M.K. *et al*. (2002) Statistical analysis of a gene expression microarray experiment with replication. *Stat. Sin.*, **12**, 203–217.

Kerr,M.K. *et al*. (2000) Analysis of variance for gene expression microarray data. *J. Comput. Biol.*, **7**, 819–837.

Kinzler,K.W. and Vogelstein,B. (1997) Gatekeepers and caretakers. *Nature*, **386**, 761–763.

Lancaster,J.M. *et al*. (1996) BRCA2 mutations in primary breast and ovarian cancers. *Nat. Genet.*, **13**, 238–240.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, e161.

Lehmann,E.L. (1986) *Testing Statistical Hypotheses*, 2nd edn. Wiley, New York.

Listgarten,J. *et al*. (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl Acad. Sci. USA*, **107**, 16465–16470.

Mevik,B.-H. and Wehrens,R. (2007) The pls package: principal component and partial least squares regression in R. *J. Stat. Softwr.*, **18**, 1–24.

Murphy,K.M. *et al*. (2002) Evaluation of candidate genes MAP2K4, MADH4, ACVR1B, and BRCA2 in familial pancreatic cancer: deleterious BRCA2 mutations in 17. *Cancer Res.*, **62**, 3789–3793.

Narod,S.A. *et al*. (2008) Rapid progression of prostate cancer in men with a BRCA2 mutation. *Br. J. Cancer*, **99**, 371–374.

Özcelik,H. *et al*. (1997) Germline BRCA2 6174delT mutations in Ashkenazi Jewish pancreatic cancer patients. *Nat. Genet.*, **16**, 17–18.

Peterson,L.F. *et al*. (2007) The multi-functional cellular adhesion molecule CD44 is regulated by the 8;21 chromosomal translocation. *Leukemia*, **21**, 2010–2019.

Price, *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Riccioni,R. *et al*. (2005) TRAIL decoy receptors mediate resistance of acute myeloid leukemia cells to TRAIL. *Haematologica*, **90**, 621–624.

Rosipal,R. and Krämer,N. (2006) Overview and recent advances in partial least squares. In *Subspace, Latent Structure and Feature Selection Techniques, Lecture Notes in Computer Science*. Springer, pp. 34-51.

Rubio-Moscardo,F. *et al*. (2005) Characterization of 8p21.3 chromosomal deletions in B-cell lymphoma: TRAIL-R1 and TRAIL-R2 as candidate dosage-dependent tumor suppressor genes. *Blood*, **106**, 3214–3222.

Scheid,S. and Spang,R. (2007) Compensating for unknown confounders in microarray data analysis using filtered permutations. *J. Comput. Biol.*, **14**, 669–681.

Stegle,O. *et al*. (2008) Accounting for non-genetic factors improves the power of eQTL studies. In *Proceedings of the 12th International Conference on Research in Computational Molecular Biology*, pp. 411–422.

Stegle,O. *et al*. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, **6**, e1000770.

Tataroglu,C. *et al*. (2007) Beta-catenin and CD44 expression in keratoacanthoma and squamous cell carcinoma of the skin. *Tumori.*, **93**, 284–289.

Wagner,J.E. *et al*. (2004) Germline mutations in BRCA2: shared genetic susceptibility to breast cancer, early onset leukemia, and Fanconi anemia. *Blood*, **103**, 3226–3229.

Welch,B.L. (1938) The significance of the difference between two means when the population variances are unequal. *Biometrika*, **29**, 350–362.

Wold,H. (1975) Path models with latent variables: the NIPALS approach. In *Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building*, pp. 307–357.

Wold,H. (1985) Partial least squares. In Kotz,S. and Johnson,N.L. (eds), *Encyclopedia of the Statistical Sciences*, vol. 6. Wiley, New York, pp. 581–591.

Wolfinger,R.D. *et al*. (2001) Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.*, **8**, 625–637.

Yu,J.M. *et al*. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.*, **38**, 203–208.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016