BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 11 2011, pages 1496–1505
doi:10.1093/bioinformatics/btr171

Gene expression

Advance Access publication April 6, 2011

Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies

Andrew E. Teschendorff[1,*], Joanna Zhuang[1,2] and Martin Widschwendter[2]

[1]Statistical Genomics Group, Paul O'Gorman Building, UCL Cancer Institute, 72 Huntley Street and [2]Department of Gynecological Oncology, UCL Elizabeth Garrett Anderson Institute for Women's Health, University College London, London WC1E 6BT, UK

Associate Editor: David Rocke

ABSTRACT

Motivation: A common difficulty in large-scale microarray studies is the presence of confounding factors, which may significantly skew estimates of statistical significance, cause unreliable feature selection and high false negative rates. To deal with these difficulties, an algorithmic framework known as Surrogate Variable Analysis (SVA) was recently proposed.

Results: Based on the notion that data can be viewed as an interference pattern, reflecting the superposition of independent effects and random noise, we present a modified SVA, called Independent Surrogate Variable Analysis (ISVA), to identify features correlating with a phenotype of interest in the presence of potential confounding factors. Using simulated data, we show that ISVA performs well in identifying confounders as well as outperforming methods which do not adjust for confounding. Using four large-scale Illumina Infinium DNA methylation datasets subject to low signal to noise ratios and substantial confounding by beadchip effects and variable bisulfite conversion efficiency, we show that ISVA improves the identifiability of confounders and that this enables a framework for feature selection that is more robust to model misspecification and heterogeneous phenotypes. Finally, we demonstrate similar improvements of ISVA across four mRNA expression datasets. Thus, ISVA should be useful as a feature selection tool in studies that are subject to confounding.

Availability: An R-package *isva* is available from www.cran.r-project.org.

Contact: a.teschendorff@ucl.ac.uk

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

Received on August 17, 2010; revised on March 27, 2011; accepted on March 30, 2011

1 INTRODUCTION

Confounding is a common problem in large-scale microarray profiling studies (Leek and Storey, 2007). For example, a study profiling hundreds of samples over a relatively large number of chips is vulnerable to confounding by chip effects. Even if the study is well-designed and balanced in relation to the phenotype of interest, logistical or sample quality issues may induce confounding. Moreover, inter-chip variation cannot always be corrected for, even with powerful inter-array normalization techniques such as quantile normalization (Bolstad *et al.*, 2003; Teschendorff *et al.*, 2009). Global normalization methods such as quantile normalization may also not be desirable if they obscure biologically important global differences between samples, as it may happen for example in DNA methylation studies (Laird, 2010; Moore *et al.*, 2008). A further difficulty is that often the confounding factors (CFs) are unknown to the experimentalist (Leek and Storey, 2007) or, if known, may be subject to substantial uncertainty or measurement error. In these scenarios, techniques that have been used to explicitly adjust for known and error-free confounding factors (Johnson *et al.*, 2007) cannot be used.

CFs may affect estimates of statistical significance in various ways. Given a phenotype of interest, one of the most important measures of statistical significance is the false discovery rate (FDR), which can be thought of as the posterior probability that a feature declared to be positive is a true null (Storey and Tibshirani, 2003). A common assumption in estimating the FDR is the statistical independence of the multiple tests. This independence condition is, however, frequently violated as features often exhibit strong correlations. These correlations could be biological, or they could be due to experimental factors such as chip effects. Not taking these correlations into account can lead to instability of ranked gene lists, skew estimates of statistical significance and affect the power of a study (Leek and Storey, 2007). Confounding may also increase the variability within phenotypes, thus turning true positives into false negatives, and hence leading to unacceptably large miss rates.

To deal with these issues, an algorithm known as Surrogate Variable Analysis (SVA) was proposed (Leek and Storey, 2007), and shown to lead to improved estimates of statistical significance, biological accuracy and reproducibility. The main concept behind SVA is to model the potential confounding factors, which may or may not be known, as singular vectors (so called 'surrogate variables') derived from a singular value decomposition (SVD). By definition, these surrogate variables are linearly uncorrelated, leading to an SVA model in which the residual noise is effectively white (Leek and Storey, 2007, 2008).

We propose that CFs would be better modeled as statistically independent surrogate variables, as modeling them as such better reflects the way the confounding noise is generated. For example, given a gene expression study where the phenotype of interest is clinical outcome, it is clear that two confounders such as beadchip and age are statistically independent variables. It, therefore, seems natural to model the effect of these factors on the data as statistically

*To whom correspondence should be addressed.

independent random variables. This requires the variables to be uncorrelated in a non-linear fashion, a stronger condition than the linear uncorrelatedness imposed by an SVD. In order to model CFs as statistically independent variables, we therefore propose to use independent component analysis (ICA) (Comon, 1994; Hyvaerinen *et al.*, 2001) in a framework similar to the one used in SVA, which we call Independent Surrogate Variable Analysis (ISVA).

ICA has already been successfully applied to gene expression data to improve the identification of altered signaling pathways and prediction of clinical outcome in cancer (Carpentier *et al.*, 2004; Frigyesi *et al.*, 2006; Huang and Zheng, 2006; Lee and Batzoglou, 2003; Liebermeister, 2002; Martoglio *et al.*, 2002; Saidi *et al.*, 2004; Teschendorff *et al.*, 2007; Zhang *et al.*, 2005). As shown in several of these studies [see e.g. (Teschendorff *et al.*, 2007)], ICA outperforms Principal Component Analysis (PCA) (or SVD) in identifying alterations in biological pathways and regulatory modules. Based on this result, we therefore hypothesized that ICA would also allow us to better identify potential confounders. The ability to better identify confounding factors would thus provide a more flexible framework for robust downstream statistical inference. For instance, one may want to adjust feature profiles for only a subset of the confounding factors, as some of the confounders may be related to biological factors of interest. Indeed, as we shall see, model misspecification, which may be caused by a highly heterogeneous phenotype, can lead to true biological signal being misinterpreted as variation due to a CF. To address this problem, we propose a surrogate variable selection step within the ISVA framework. Thus, ISVA differs from SVA in two important aspects: (i) the use of ICA instead of PCA to identify variables that represent potential confounders and (ii) inclusion of a surrogate variable selection step.

In order to evaluate ISVA, we use both synthetic and real data, and benchmark it against SVA and the corresponding models with and without adjustment for confounding factors. We also consider a modified SVA algorithm (SVA*) which incorporates the same surrogate variable selection step implemented in ISVA. Motivated by a surge of interest in cancer epigenomics, we illustrate the application of ISVA to several large-scale DNA methylation datasets which have been profiled with Illumina's Infinium Methylation 27k arrays (Bibikova *et al.*, 2009). Despite the high reproducibility of these arrays, we have previously demonstrated that variation in bisulfite conversion efficiency and beadchip effects are prominent CFs in large-scale studies using these arrays (Teschendorff *et al.*, 2009, 2010). Thus, these data provide an ideal sandpit to explore surrogate variable methodologies. However, in order to demonstrate the general utility of ISVA, we also consider gene expression data.

2 METHODS

2.1 Notation and review of SVA

In what follows, we assume that we have a data matrix, X_{ij}, with i ($i=1,...,p$) labeling the features (genes,CpGs,...) and j ($j=1,...,n$) labeling the samples, with $p>>n$. Furthermore, we assume that each row of X has been mean centered, and that we have a phenotype of interest (POI) encoded by a vector $\vec{y}=\{y_1,...,y_n\}$. As in (Leek and Storey, 2007), we may allow for a general function of the phenotype vector, so that the starting model for SVA takes the form

$$X_{ij}=f_i(y_j)+\epsilon_{ij} \quad (1)$$

SVA proceeds by performing a SVD of the residual matrix

$$R=UDV^T \quad (2)$$

where the residual matrix is defined by $R_{ij}\equiv X_{ij}-\hat{f}_i(y_j)$. Thus, the singular vectors of the SVD capture variation which is orthogonal to the variation associated with the POI. This residual variation is likely to be associated with other biological factors not of direct interest, or experimental factors, all of which constitute potential confounders. SVA provides a prescription for the construction of surrogate variables in terms of the singular vectors of this SVD (Leek and Storey, 2007).

2.2 Review of ICA

Briefly, we review the ICA model (Comon, 1994; Hyvaerinen *et al.*, 2001). ICA produces an approximate decomposition of a data matrix X into the product of two matrices S (the 'source' matrix) and A (the 'mixing' matrix):

$$X_{ij}=\sum_{k=1}^{K}S_{ik}A_{kj}+\epsilon_{ij}, \quad (3)$$

where $K\leq\min\{p,n\}$ is the number of components to be computed. When K is strictly smaller than $\min\{p,n\}$, it is in general impossible to pick S and A such that the error matrix vanishes. Therefore, the ICA algorithm aims at making ϵ as small as possible, usually in the least squares sense. This condition on ϵ still leaves much leeway to select the matrices S and A. Given an initial source matrix estimate S' (for instance obtained by PCA), ICA amounts to finding a transformation W

$$S_{ik}=\sum_{m=1}^{K}S'_{im}W_{mk}, \quad (4)$$

such that the columns of S are as statistically independent as possible. Most ICA methods consider that the zero covariance property of S' is compatible with this goal, hence they preserve this property in S' by restricting W to the set of $K\times K$ orthogonal transformations. The ICA algorithms thus search for an orthogonal matrix W that maximizes the statistical independence of the columns of S'. The mixing matrix finally equals

$$A_{kj}=\sum_{m=1}^{K}W_{mk}A'_{mj} \quad (5)$$

A quantitative measure of independence between measurements of random variables is provided by a contrast function. There are many choices for the contrast function, leading to a variety of ICA algorithms, which may also differ in the numerical algorithm used for the optimization procedure. Here, we considered the 'FastICA' algorithm (Hyvaerinen, 1999; Hyvaerinen *et al.*, 2001) for which an R-package *fastICA* is readily available.

2.3 Dimensionality estimation using Random Matrix Theory

To estimate the number of components K to be used in ICA, we propose to use Random Matrix Theory (RMT) (Plerou *et al.*, 2002). RMT estimates the number of dimensions (components) of a data covariance matrix by comparing the statistics of the observed eigenvalues, i.e. the eigenvalues of the data covariance matrix (obtained from a PCA), to those of a random matrix counterpart. The density distribution of eigenvalues for such random matrices is known (Plerou *et al.*, 2002), and therefore comparison of the observed eigenvalues to this analytical 'null' distribution can be used to obtain an estimate for the number of components. Specifically, the number of observed eigenvalues larger than the analytical maximum provides an approximate estimate of the number of significant components (Plerou *et al.*, 2002). To validate the RMT estimate and to ensure that the theoretical null distribution does not deviate significantly from that of the empirical null (deviations might be expected because of the finite size of the matrix and because data may not be Gaussian) (Plerou *et al.*, 2002), we scrambled up the data matrix (for each column, a distinct permutation of the rows is performed) and verified that RMT predicted zero significant components. Thus, observed eigenvalues of the data covariance matrix larger than the

theoretical maximum provides a reasonable approximation to the number of dimensions to use. A considerable advantage of RMT over other algorithms based on explicit permutations (Buja and Eyuboglu, 1992) is its analytical nature, and thus the computational cost is minimal. To further justify the use of RMT in the present context, we note that in the *fastICA* implementation, a whitening of the data covariance matrix precedes the application of ICA. Since whitening consists of a PCA plus a rescaling (which turns the data covariance matrix into the identity matrix), it is natural to use RMT to estimate the number of principal components to retain. Subsequently, ICA is applied to the whitened data matrix over this K-dimensional subspace only.

2.4 ISVA

Variation in any dataset has three parts: one part reflects the biologically interesting variation, another reflects technical 'unwanted' variation caused by known or hidden confounders and the third part is the inherent stochastic (random) variation. The observed data can, therefore, be viewed as an interference pattern representing a superposition of these various parts. It is most natural to assume that confounders affect the data in ways that are statistically independent, or if not, it is natural to attempt to resolve the factors into components that are as statistically independent as possible. For example, if tumor samples are drawn from two related but distinct cohorts and are profiled on a large number of microarrays, the array and cohort of origin can be viewed as two independent and potential confounders. Even in the scenario where CFs may affect the data in a correlated fashion, deconvolution of confounding effects into statistically independent components should improve the interpretation of the data. Thus, the problem of identifying the sources of unwanted variation can also be viewed as a 'blind source separation problem' often encountered in other fields of science, and for which ICA provides the natural algorithmic framework in which to infer the sources (Hyvaerinen *et al.*, 2001). In the current context, the sources would correspond to the (potentially unknown) CFs. Motivated by this, we propose an extension to SVA, called ISVA, which we now describe.

As with SVA, there are two parts to the algorithm: (i) detection of confounding/unmodeled factors (steps 1–4) and (ii) construction of independent surrogate variables (ISVs) (steps 5–10):

(1) Construction of the residual variation matrix by removing the variation associated with the phenotype of interest: $R_{ij} \equiv X_{ij} - \hat{f}_i(y_j)$.

(2) We estimate the intrinsic dimensionality, K, of the residual variation matrix using Random Matrix Theory. This gives the number of components as input to the ICA algorithm.

(3) Perform ICA on R: $R = SA + \epsilon$, with S a $p \times K$ and A a $K \times n$ estimated data matrix.

(4) The significant independent components are given by the columns of S (S_k) and rows of A (A_k), respectively.

(5) We regress A_k to each X_i ($i = 1, ..., p$) and calculate P-values of association p_i.

(6) From this P-value distribution, we estimate the FDR using the q-value method (Storey and Tibshirani, 2003) and select the features with $q < 0.05$. If the number of selected features is less than 500, we select the top 500 features (based on P-values). Let r_k denote the number of selected features.

(7) We construct the reduced $r_k \times n$ data matrix X_r obtained by selecting the features in previous step.

(8) Perform ICA on X_r using K independent components: $X_r = S_r A_r + \epsilon_r$. Find the column k^* of A_r that best correlates (absolute correlation) with A_k.

(9) Set the ISV $v_k = (A_r)_{k^*}$.

(10) Repeat steps 5–9 for each significant independent component, A_k, obtained in step 4.

To finally identify features correlating with the phenotype of interest, we adjust for the ISVs using the general model:

$$X_{ij} = f_i(y_j) + \sum_{k=1}^{K} \lambda_{ki} v_{kj} + \epsilon_{ij} \quad (6)$$

2.5 Heterogeneous phenotypes and model misspecification: selection of (independent) surrogate variables

Misspecification of the phenotype model $f_i(y_j)$ is likely, specially if the POI is very heterogeneous. In this scenario, the residual orthogonal variation may contain biologically relevant variation associated with the POI. Thus, SVs (ISVs) may be interpreted as CFs when, in fact, they represent components of variation associated with the POI. Inclusion of these SVs (ISVs) in the model would, therefore, *remove* biological signal. Unfortunately, in the case where a confounder is unknown it would be very hard to tell if a potential association of a SV/ISV with the POI is genuine or caused by the confounder. On the other hand, if the most important CFs are known (but maybe subject to error/uncertainty) one may use the following criteria to select ISVs: only ISVs that significantly correlate ($P < 0.01$) with CFs are included in the model. If an ISV correlates with only the POI or if it correlates significantly and more strongly with the POI then it is excluded. Therefore, in this modified ISVA, only ISVs clearly associated with CFs are included and any ISV that correlates more strongly with the POI is excluded, thus residual biological signal associated with the POI would not be removed. It is important to realize that this procedure is not equivalent to using the CFs themselves as covariates as (i) the CFs may be subject to uncertainty or measurement error and (ii) as the effect of the CFs on the data may be highly complex.

2.6 Algorithm comparison

We benchmark ISVA against SVA (Leek and Storey, 2007, 2008) and the corresponding model without adjustment for confounding factors, denoted as LR (as we here consider simple linear regressions). To test the different models, we use both simulated as well as real biological data. When evaluating performance on real data, we also compare ISVA to a modified SVA algorithm (SVA*) that incorporates the same surrogate variable selection step used in ISVA, and to the explicit model that uses the CFs as covariates (LR + CFs). These additional comparisons allow us to (i) evaluate the separate contributions to overall performance from the ICA and ISV selection steps in ISVA and (ii) to evaluate if ISVA improves the identifiability of confounders and if so what the impact of such improved modeling is on downstream statistical inference.

2.7 Simulated data

Details of the simulated data can be found in the Appendix A. In the following, let H_A denote the number of features which follow the alternative hypothesis, and H_0 denote the number of true nulls. Furthermore, let H_{AA} denote the number of features that follow the alternative hypothesis but which are also affected by the confounders ($H_{AA} \leq H_A$). We compare SVA, ISVA and the simple linear regression model (LR, i.e. no adjustment for confounders) in their ability to detect positives (NP) at an estimated FDR threshold of 0.05 (Storey and Tibshirani, 2003), the positive predictive value PPV = TP/NP where TP is the number of true positives passing the same 0.05 FDR threshold and the false negative rate FNR = FN/H_A where FN is the number of false negatives at that threshold. We also compare the sensitivity of the algorithms to detect true positives which are affected by confounders, SE-A, defined as the ratio TP_A/H_{AA} where TP_A is the number of truly altered features which are affected by confounding and which pass the FDR threshold of 0.05 . Identifiability of inferred surrogate or independent surrogate variables with confounders is evaluated using the average of the best Pearson correlations between the known confounders and the surrogate or independent surrogate variables. To test that P-values from true nulls follow the uniform distribution between 0 and 1, we use the Kolmogorov–Smirnov (KS) test.

2.8 Biological data: DNA methylation

2.8.1 Illumina Infinium DNAm assay We consider four DNA methylation datasets which have been generated using Illumina's Infinium Human Methylation 27k Beadchips (Bibikova *et al.*, 2009) and which have already been presented elsewhere (Teschendorff *et al.*, 2010). The Beadchips interrogate the methylation status of ∼27 000 CpGs with an average of 2 CpGs per gene promoter region and with a mean separation of ∼600 bp. Each beadchip consists of 12 strips allowing 12 samples to be profiled per chip. In this work, we used the normalized data as described in (Teschendorff *et al.*, 2010). Let i denote the CpG and j the sample. The normalized methylation values of the CpGs follow an approximate β-valued distribution, with β constrained to lie between 0 (unmethylated locus) and 1 (methylated). This follows from the definition of β as the ratio of methylated to combined intensity values i.e.

$$\beta_{ij} = \frac{M_{ij}}{U_{ij} + M_{ij} + e} \tag{7}$$

where U_{ij} and M_{ij} are the unmethylated and methylated intensity values of the probe (averaged over bead replicates) and e is a small correction term to regularize probes of low total signal intensity (i.e. probes with $U_{ij} + M_{ij} \approx 0$ after background subtraction). Thus, our data matrices X_{ij} are such that $X_{ij} = \beta_{ij}$ where β_{ij} is the normalized methylation value as given above.

2.8.2 CFs In large-scale studies using the Illumina Infinium Human Methylation 27k array, we generally observe two main confounders: beadchip effects and variations in bisulfite conversion efficiency (BSCE). BSCE is assessed using the built-in chip controls. We observed that beadchip effects are not removed with the Illumina normalization protocol which is designed to ameliorate any spatial chip effects (i.e. outliers of the 20–30 bead replicates are removed when estimating averages). We also note that variation due to the beadchip and BSCE was not removed by quantile normalization procedures (Teschendorff *et al.*, 2009).

2.8.3 Datasets 1–4: DNAm of whole blood samples—In all datasets, age is the phenotype of interest. (i) Dataset 1: this DNAm dataset consists of 187 blood samples from patients (94 women and 93 men) with type-1 diabetes. This set served as validation for a DNAm signature for aging (Teschendorff *et al.*, 2010). We take BSCE, beadchip, cohort and sex as potential confounding factors. Samples were distributed over 17 beadchips. (ii) Dataset 2: this DNAm set consists of 108 blood samples from healthy post-menopausal women which served as controls for the UKOPS study (Teschendorff *et al.*, 2009). CFs in this study include BSCE, beadchip and DNA concentration (DNAc). Samples were distributed over 10 beadchips. (iii) Dataset 3: This is similar to Dataset 2 but consists of 145 blood samples from healthy post-menopausal women distributed over 36 beadchips (i.e. ∼4 healthy samples per chip, the other 8 blood samples per chip were from cancer cases) (Teschendorff *et al.*, 2009). (iv) Dataset 4: This data set consists of whole blood samples from a total of 84 women (49 healthy and 35 women with breast cancer). Samples were distributed over seven beadchips, and confounders are BSCE, status (cancer/healthy) and beadchip (Widschwendter,M., submitted for publication).

2.9 Biological data: mRNA expression

2.9.1 Datasets 5–8: breast cancer mRNA expression—the mRNA expression profiles are all from primary breast cancers and three of the data sets (Datasets 5–7) were profiled on Affymetrix platforms, while Dataset 8 was profiled on an Illumina Beadchip. Normalized data were downloaded from GEO (*http://www.ncbi.nlm.nih.gov/geo/*) and probes mapping to the same Entrez ID identifier were averaged. Dataset 5: 14 223 genes and 101 samples (Sotiriou *et al.*, 2006); Dataset 6: 15 736 genes and 137 samples (Loi *et al.*, 2007); Dataset 7: 13 292 genes and 200 samples (Schmidt *et al.*, 2008); Dataset 8: 17 941 genes and 128 samples (Blenkiron *et al.*, 2007). In these datasets, we take histological grade as the phenotype of interest, and consider estrogen receptor status (ER) and tumor size as potential confounders. Cell cycle-related genes are known to discriminate low- and high-grade breast cancers irrespective of estrogen receptor status (Loi *et al.*, 2007; Sotiriou *et al.*, 2006). Therefore, we compare the algorithms in their ability to detect specifically cell cycle-related genes and not estrogen-regulated genes. To this end, we focused attention on four gene sets, two representing cell cycle-related genes from the Reactome *www.reactome.org* and gene ontoloy (GO) *www.geneontology.org*, and two sets representing estrogen receptor (*ESR1*) regulated genes (Doane *et al.*, 2006; van 't Veer *et al.*, 2002). The cell cycle sets showed negligible overlap with the *ESR1* gene sets; however, we removed the few overlapping genes to ensure mutual exclusivity of cell cycle and *ESR1* sets.

3 RESULTS

3.1 Parameter choices

In generating the synthetic data, the most important parameter is the relative effect size of the phenotype of interest to that of the CFs. The effect sizes of phenotype and CFs are often not known in advance of the experiment, but can be estimated from the generated data.

To obtain reasonable estimates for the relative effect size parameters to consider, we first used a real dataset (Dataset 1, Section 2). The dataset consists of DNAm profiles from whole blood samples of 187 individuals (94 women, 93 men) obtained using Illumina Infinium 27k arrays (Bibikova *et al.*, 2009). The distribution of methylation β values across all CpGs and samples exhibits the typical bimodality, reflecting the fact that most CpGs in the promoters of genes are unmethylated with a reduced number being hypermethylated (Supplementary Fig. S1). Because of this inherent bimodality, the top singular vector of an SVD merely captures this bimodality and if not removed would lead to biased estimates of the intrinsic dimensionality of the dataset. Therefore, to remove this trivial component of variation we mean-centered each CpG to zero, resulting in the more familar supergaussian distribution (Supplementary Fig. S1). SVD decomposition and RMT analysis of this data matrix revealed a total of seven significant singular values (Section 2, Fig. 1A), which can be subsequently correlated to the phenotype of interest (here age) and potential CFs (Fig. 1B). To validate RMT, we verified that RMT predicted zero significant components on randomised data matrices obtained from scrambling up the original data matrix (Section 2, Supplementary Fig. S1). From the heatmap of P-values of association between singular vectors and factors, we observed that most of the variation is correlating with factors other than age, notably sex, beadchip and bisulfite conversion efficiency. We point out that most of the variation associated with sex is due to the imbalance of the sex chromosomes. However, the variation associated with BSCE and beadchip cannot be corrected for easily (Teschendorff *et al.*, 2009) and associated components can be observed to account for 4–10 times more variation than the variation associated with age (Fig. 1B). Thus, motivated by this example we consider corresponding relative effect sizes of 2 and 4. For comparison, we also consider the case where the relative effect size is unity.

3.2 Simulated data

The synthetic datasets were generated across a total of 2000 features and 50 samples with a primary binary variable of interest and two CFs (also binary variables, Section 2). We assumed that 10% of features (i.e. 200) were discriminatory for each binary variable

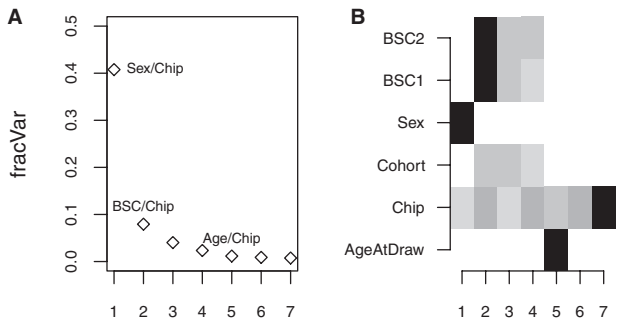

Fig. 1. **(A)** Relative fraction of variation carried by each of the seven significant singular vectors of an SVD, as measured relative to the total variation in the data. Number of significant singular vectors was estimated using Random Matrix Theory (RMT) (Section 2). Some of the singular values are labeled according to which confounders the corresponding singular vectors are correlated to, as shown in panel (B). **(B)** Heatmap of P-values of association between the seven significant singular vectors and the phenotype of interest (here age at sample draw) and CFs [Chip, cohort, sex and bisulphite conversion (BSC) efficiency controls 1 and 2]. P-values were estimated using linear ANOVA models in the case of chip, cohort and sex, while linear regressions were used for age and BSC efficiency. Color codes: $P < 1e-10$ (black), $P < 1e-5$ (dark grey), $P < 0.001$ (grey), $P < 0.05$ (light grey), $P > 0.05$ (white).

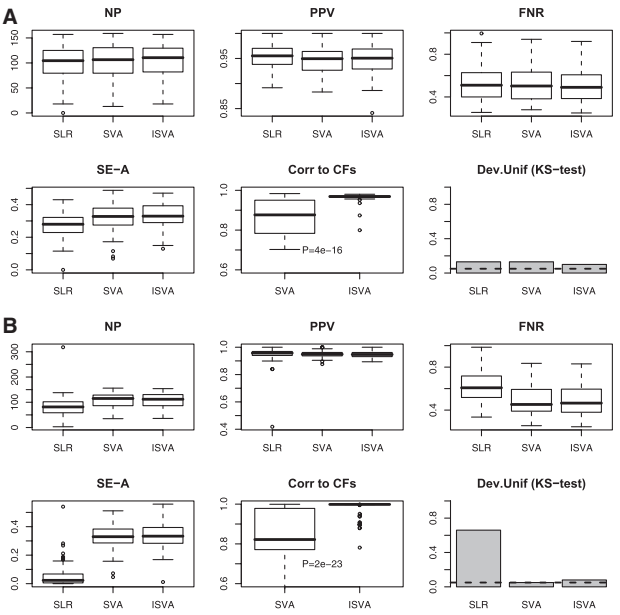

Fig. 2. Feature selection performance metrics of the different algorithms over 100 runs of synthetic data. The algorithms for feature selection are SVA, ISVA and simple linear regression without adjustment for confounders (SLR). For a given estimated FDR threshold of 0.05, we compare the number of positives (NP) that pass this threshold, the positive predictive value (PPV), the false negative rate (FNR), the sensitivity to detect true positives which are affected by confounders (SE-A), the average Pearson correlation between confounders and the best correlated surrogate or independent surrogate variable (Corr to CFs) (P-values shown are from a paired Wilcoxon rank sum test comparing these best correlation values between SVA and ISVA), and the fraction of runs with a null gene P-value distribution deviating from a uniform one according to Kolmogorov–Smirnov test (Dev.Unif KS-test). See Section 2 for further details. **(A)** Relative effect size = 1. **(B)** Effect size of confounders is assumed to be 4× the effect size of the phenotype of interest.

(phenotype and CFs) and allowed for overlap, so that on average ∼90 of the 200 true positives were also affected by confounding.

We first considered the case where the relative effect size is 1. Over the 100 runs and using an FDR threshold of 0.05, we did not observe significant differences in the number of detected positives, PPV (1-FDR) or FNR between LR (no adjustment for confounding factors) and either SVA or ISVA (Fig. 2A). For all three methods, in about 10% of runs did the null gene P-values exhibit significant deviations (KS test $P < 0.05$) from a uniform distribution. However, we observed that SVA and ISVA outperformed LR in terms of the sensitivity of detecting true positives affected by confounding factors (SE-A) (Fig. 2A). Importantly, ISVA was much better than SVA in reconstructing the confounding factors achieving over 95% reconstruction accuracy (Fig. 2A). Next, we considered the case where the relative effect size is 4, representing the scenario depicted in Figure 1. As before, the estimated FDR (PPV) was close to the true FDR (PPV) (Fig. 2B). Most importantly, SVA and ISVA markedly reduced the FNR, significantly improved the SE-A and null gene P-values generally conformed to that of a uniform distribution (Fig. 2B). When comparing SVA to ISVA, we did not observe any appreciable differences in either FNR, SE-A or KS est. However, as in the previous case, ISVA achieved close to 100% reconstruction accuracy of CFs, which was significantly better than that achieved by SVA (Fig. 2B). Performance metrics for an intermediate case where the relative effect size is 2 were similar (Supplementary Fig. S2).

3.3 DNA methylation data

Based on the simulation results, we would also expect ISVA to outperform SVA in modeling the CFs when applied to real data. We considered four DNA methylation datasets, all subject to a low SNR and to confounding (Datasets 1–4, Fig. 1, Supplementary Figs S3–S5). Epigenetics and DNA methylation, in particular, have attracted considerable attention because of the now well-established role of DNA methylation in ageing and cancer (Teschendorff *et al.*, 2010). We, therefore, applied SVA and ISVA to these four datasets using age as the phenotype of interest. The number of inferred SVs/ISVs and their correlations to potential CFs can be found in Supplementary Table S1; Fig. S6). For each CF, we recorded the SV/ISV which best modeled the factor based on R^2 and P-value statistics (Table 1). In most cases, we observed that the chosen SVs/ISVs correlated significantly with the CFs. Based on this choice of best SV/ISV, we can see that ISVA compares favorably to SVA, specially in modeling beadchip effects. Since a single component may not capture all possible beadchip effects, we extended the analysis to consider the next best SV/ISV correlating with chip. In the larger datasets, the next best SV/ISV also correlated significantly with the beadchip with R^2 values larger than 0.7 in the case of ISVA, but not so for SVA (Figs 3 and 4). Corresponding F-statistics for ISVA are larger illustrating that the ISVs capture more of the inter-chip variability while keeping low variability within each chip (Figs 3 and 4).

Thus, these results extend the better modelling performance of ISVA seen in synthetic data to real biological data. Interestingly, the noticeable improvement of ISVA over SVA was specially true in

Table 1. ISVA versus SVA in DNA methylation Datasets 1–4

Dataset 1	SVA-P	SVA-R^2	ISVA-P	ISVA-R^2
BSCE	1e-36	0.58	4e-32	0.53
Sex	1e-214	0.99	8e-216	1.00
Chip	2e-48	0.77	1e-56	0.81
Dataset 2	**SVA-P**	**SVA-R^2**	**ISVA-P**	**ISVA-R^2**
DNAc	1e-5	0.16	2e-11	0.34
BSCE	4e-9	0.27	1e-9	0.28
Chip	3e-15	0.54	2e-56	0.94
Dataset 3	**SVA-P**	**SVA-R^2**	**ISVA-P**	**ISVA-R^2**
DNAc	0.04	0.02	0.04	0.02
BSCE	7e-18	0.39	1e-20	0.45
Chip	1e-21	0.68	7e-28	0.76
Dataset 4	**SVA-P**	**SVA-R^2**	**ISVA-P**	**ISVA-R^2**
Status	5e-4	0.13	3e-4	0.14
BSCE	1e-13	0.49	9e-14	0.49
Chip	9e-11	0.49	5e-27	0.81

For each confounding factor we give the R^2 and P-values of the best SV/ISV correlating to it.

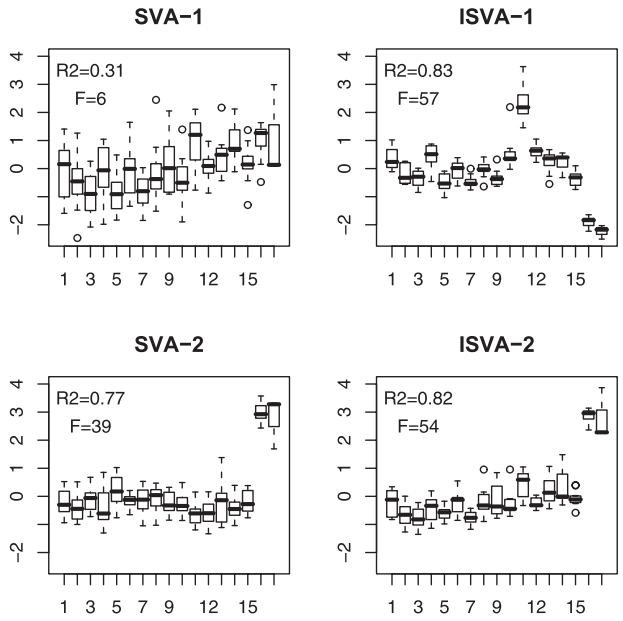

Fig. 3. Comparison of ISVA with SVA in identifying beadchip effects in Dataset 1. The weights (y-axis) in the two surrogate (SVA) and independent surrogate variables (ISVA) most significantly associated with beadchip effects are plotted against beadchip number (x-axis). To compare the identifiability of beadchip effects, we provide the R^2 and F-statistics of a linear ANOVA model with beadchip number as the independent variable.

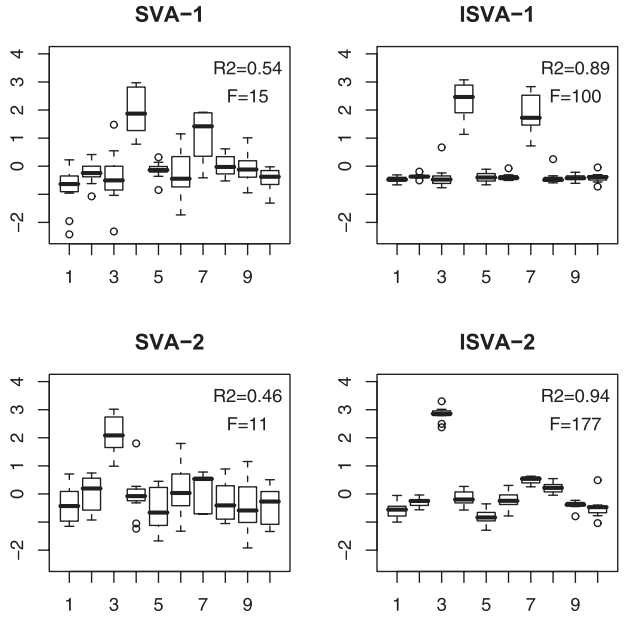

Fig. 4. Comparison of ISVA with SVA in identifying beadchip effects in Dataset 2. The weights (y-axis) in the two surrogate (SVA) and independent surrogate variables (ISVA) most significantly associated with beadchip effects are plotted against beadchip number (x-axis). To compare the identifiability of beadchip effects, we provide the R^2 and F-statistics of a linear ANOVA model with beadchip number as the independent variable.

modeling beadchip effects and for large studies where the 12 samples represented on each beadchip constitute a small subset of the whole data. Indeed, we did not observe appreciable differences between SVA and ISVA for a smaller DNAm dataset of 48 cervical smear samples (four beadchips, data not shown). This is not unexpected as the ICA implementation used here works best for independent sources which are in some sense sparse, i.e. sources affecting a small fraction of the samples or features (Hyvaerinen *et al.*, 2001). We also note that ISVA (and SVA) could model fairly accurately (high R^2 values) CFs which are well-defined and subject to no measurement error or uncertainty (at least in the sense of how they vary across samples). Examples of such CFs are beadchip and sex. In contrast, for CFs which are subject to measurement error such as BSCE and DNAc, R^2 values were generally much lower. Put together this also suggests that the inferred ISVs may more accurately model the effect that these CFs have on the data than using the CFs themselves as covariates.

To investigate this further and to compare the power of the various algorithms on real data, we assessed their ability to identify CpGs differentially methylated with age (aDMCs), and specifically, their ability to detect among the age-hypermethylated aDMCs, CgGs that map to polycomb group targets (PCGTs), a family of genes which are known to undergo hypermethylation with age (Maegawa *et al.*, 2010; Rakyan *et al.*, 2010; Teschendorff *et al.*, 2010) (Table 2). Overall, the results across the four datasets demonstrate that ISVA is the most robust algorithm, conclusively capturing the age DNAm signature in all datasets examined (Table 2). The model using explicit CFs as covariates (LR + CFs) performed well but only obtained a more marginal association in Dataset 3. Importantly, ISVA compared

Table 2. Age-associated CpGs

Dataset 1	LR	LR + CFs(4)	SVA(4)	SVA*(4)	ISVA(6)
aDMCs	294	440	688	688	902
nPCGTs	75	96	110	110	148
P	3e-25	4e-32	7e-26	7e-26	2e-34
Dataset 2	LR	LR+CFs(3)	SVA(18)	SVA*(9)	ISVA(6)
aDMCs	225	267	4	8	232
nPCGTs	64	75	1	1	59
P	1e-20	4e-24	0.27	0.40	2e-19
Dataset 3	LR	LR+CFs(3)	SVA(21)	SVA*(8)	ISVA(8)
aDMCs	69	20	201	163	225
nPCGTs	5	4	15	13	29
P	0.08	0.001	0.01	0.007	3e-7
Dataset 4	LR	LR+CFs(3)	SVA(15)	SVA*(5)	ISVA(6)
aDMCs	969	564	185	479	469
nPCGTs	124	84	19	53	64
P	2e-30	7e-22	0.01	3e-11	8e-16

In each dataset and for each method (LR, LR + CFs, SVA, SVA*, ISVA), we give the number of CpGs differentially methylated with age (aDMCs) (FDR < 0.05 for Datasets 1–2, FDR < 0.3 for Datasets 3 and 4), the number of these that are hypermethylated with age and that map to polycomb group targets (PCGTs), and the *P*-value of PCGT enrichment among age-hypermethylated CpGs (Hypergeometric test). In brackets, we give the number of CFs, SVs, ISVs used as covariates in the regression analysis.

Table 3. Grade-associated expression differences

Dataset 5	LR	LR + CFs (2)	SVA (4)	SVA*(1)	ISVA (4)
nDEGs	5334	491	0	1998	607
Cell cycle(Reactome)	2e-4	6e-18	1	1e-27	5e-16
Cell cycle(GO)	5e-8	2e-19	1	4e-14	6e-16
ESR1-UP(Veer)	9e-18	9e-11	1	9e-12	8e-4
ESR1-UP(Doane)	2e-7	0.03	1	0.02	0.14
Dataset 6	LR	LR + CFs (2)	SVA (19)	SVA*(2)	ISVA (5)
nDEGs	3835	829	0	0	146
Cell cycle(Reactome)	7e-27	5e-37	1	1	7e-24
Cell cycle(GO)	3e-13	5e-18	1	1	7e-12
ESR1-UP(Veer)	5e-20	0.70	1	1	0.35
ESR1-UP(Doane)	1e-13	0.90	1	1	0.61
Dataset 7	LR	LR + CFs (2)	SVA (27)	SVA*(6)	ISVA (15)
nDEGs	4488	2364	0	0	451
Cell cycle(Reactome)	4e-18	3e-25	1	1	5e-19
Cell cycle(GO)	2e-12	8e-10	1	1	6e-12
ESR1-UP(Veer)	6e-33	1e-7	1	1	0.14
ESR1-UP(Doane)	1e-25	7e-4	1	1	0.14
Dataset 8	LR	LR + CFs (2)	SVA (20)	SVA*(3)	ISVA (8)
nDEGs	3837	1292	1	2756	829
Cell cycle(Reactome)	4e-27	2e-25	1	2e-28	7e-27
Cell cycle(GO)	5e-17	5e-16	1	3e-15	4e-16
ESR1-UP(Veer)	2e-26	6e-5	1	0.08	0.21
ESR1-UP(Doane)	2e-16	7e-4	1	0.25	0.31

In each mRNA expression dataset and for each method (LR, LR + CFs, SVA, SVA*, ISVA), we give the number of genes differentially expressed with histological grade (nDEGs) (FDR < 0.05), and the *P*-value of enrichment (Hypergeometric test) of cell cycle and estrogen-regulated gene categories (Section 2) among these differentially expressed genes. CFs here are ER status and tumor size. In brackets, we give the number of CFs, SVs, ISVs used as covariates in the regression analysis.

favorably to SVA in all four datasets. To understand why ISVA performed better, we compared ISVA to a modified SVA algorithm (SVA*) which incorporates the same surrogate variable selection step as implemented in ISVA (Section 2). This showed that while SVA* improved performance over SVA in Datasets 3 and 4, that it still failed to capture the age signature in Dataset 2. Thus, the improved performance of ISVA over SVA and SVA* is dependent on the details of each dataset and can be attributed to the combined use of ICA and the ISV feature selection step. In summary, the result in Dataset 3 demonstrates the pitfalls of not adjusting for CFs (LR), of adjusting using the error-prone CFs (LR + CF), while Datasets 2–4 expose the dangers of misspecifying the phenotypic model which results in true biological signal being misinterpreted as SVs (SVA). As the results in Datasets 2 and 3 demonstrate, the necessary SV/ISV feature selection step benefits from the use of a more powerful deconvolution algorithm (ICA) that is better at separating the effects of the different confounders from the phenotype of interest.

3.4 mRNA expression data

To further demonstrate the added value of ISVA, we compared it to SVA/SVA* and the other two algorithms on four gene expression datasets from primary breast cancers (Datasets 5–8, Section 2). Genes implicated in cell proliferation and cell cycle are known to exhibit increased expression in high-grade breast cancers relative to the lower grade counterparts, irrespective of ER status (Loi *et al.*, 2007; Sotiriou *et al.*, 2006). ER status and tumour size are known to also correlate with tumor grade and therefore can be viewed as potential confounders. In particular, estrogen-regulated genes (*ESR1* targets) may be viewed as confounding features. Therefore, we compared the algorithms in their ability to detect cell cycle specific gene expression differences between high- and low-grade breast cancers without enrichment of confounding *ESR1* targets (Table 3). As expected, no adjustment (LR) led to enrichment of both cell cycle and *ESR1* gene categories. Importantly, explicit adjustment for the confounders (LR + CFs) still failed to detect specific cell cycle enrichment in three of the four datasets. This result further demonstrates the need to model confounders from the data instead of using error-prone confounders. However, SVA failed to capture the cell cycle signature in all datasets examined. Close inspection of heatmaps of association between the POI and confounders with the SVs revealed that this was caused by the presence of SVs correlating significantly with grade (Fig S7, Table S2). Incorporating the same SV selection step (SVA*) as in ISVA led to substantial improvement, but in only two of the four datasets. In contrast, ISVA detected cell cycle-specific enrichment in essentially all four data sets, once again demonstrating that the

ISV selection step benefits significantly from the use of ICA instead of SVD/PCA.

4 DISCUSSION AND CONCLUSIONS

We have proposed an extension of the SVA framework which relies on a blind source separation technique (ICA) to model potential confounders and to subsequently use these as covariates for feature selection. Our results across eight datasets encompassing two different data types (Illumina Infinium DNAm and mRNA expression) show that ISVA (ICA) improves the identifiability of CFs over the SVD/PCA framework used in SVA, and that this leads to a more robust algorithm for feature selection and downstream statistical inference. In particular, we have shown that ISVA is the only algorithm of the five considered to have clearly captured the known *specific* biological signature in each of the eight datasets examined. In this context, it is important to observe however that there was substantial interstudy variability and that ISVA may not outperform the other methods (SVA*, SVA, LR + CF) in every single dataset. We also observed that while explicit adjustment for confounders (LR + CF) worked well in some datasets, that inferred signatures were not specific enough (Table 3). This, therefore, lends further support as to why LR + CF may not be desirable (Leek and Storey, 2007).

It is also important to note that improved modeling of confounders by itself does not necessarily mean that ISVA would lead to improved statistical inference over SVA. Indeed, SVA was never intended as an algorithm to identify confounders but rather only as a tool for finding a basis for the relevant subspace (Leek and Storey, 2007). Thus, while ISVs may improve the identifiability of confounders, the subspace spanned by these might be the same as that spanned by SVs. In this scenario, SVA and ISVA would therefore lead to very similar feature selection. There is, however, an important exception to this, which happens when the data model for the phenotype of interest is misspecified. Since model misspecification is likely to occur if the phenotype of interest is highly heterogeneus, this is not an uncommon scenario. Indeed, we have provided clear examples where surrogate variables correlating with the phenotype of interest are present and that inclusion of these as covariates in the feature selection model removes biological signal (Tables 2 and 3, Figs S6–S8). In these cases, identifiability of confounders becomes important because not all SVs/ISVs should necessarily be included as covariates. This in turn means that potentially different surrogate variable subspaces would be included depending on the algorithm used to model the confounders. Thus, ISVA provides a more flexible framework as the improved identifiability of confounders allows it to better capture the potential heterogeneity of the phenotype of interest, thereby allowing a more careful choice of which variables to include as covariates.

Our results indicate that ISVA's robustness stems from the combined use of ICA and the ISV selection step. While another difference between SVA/SVA* and ISVA is in the dimensionality estimation algorithm used, we have verified that the use of RMT in the dimensionality estimation step only contributes marginally to the improved performance. Indeed, a modified ISVA which uses the same permutation-based dimensionality algorithm implemented in SVA still identified confounders better and led to similar performance as ISVA, further demonstrating the added value of ICA over PCA (Supplementary Table S3).

The proposed ISV/SV selection step is problematic if ISVs/SVs that correlate with the phenotype of interest also correlate with unknown confounders. In this scenario, it will be challenging to distinguish genuine biological variation from that caused by confounders. Another potential difficulty with ISVA (and SVA) is the sequential estimation of the two models describing the phenotype of interest and (hidden) confounders, as this two-stage approach may lead to biases in the estimated regression coefficients (Bartholomew, 1981). Therefore, it will be interesting to investigate surrogate variable methods which perform joint modeling of the phenotype and confounders and which are robust to model misspecification (Sanchez *et al.*, 2009). Similarly, it will also be interesting to investigate surrogate variable methods in the context of other popular factor decomposition methods such as non-negative matrix factorisation (Zheng *et al.*, 2009).

In summary, our results demonstrate that ISVA does well in identifying confounders and that it provides a flexible and robust framework for downstream statistical inference. It should, therefore, be useful as an alternative tool for deriving lists of differentially altered features in the presence of known or hidden confounding factors.

Funding: Heller Research Fellowship to A.E.T.; J.Z. was supported by a Comprehensive Biomedical Research Centre grant.

Conflict of Interest: none declared.

REFERENCES

Bartholomew,D.J. (1981) Posterior analysis of the factor model. *Br. J. Math. Stat. Psych.*, **34**, 93–99.

Bibikova,M. *et al.* (2009) Genome-wide DNA methylation profiling using Infinium assay. *Epigenomics*, **1**, 177–200.

Blenkiron,C. *et al.* (2007)MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.*, **8**, R214.

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

Buja,A. and Eyuboglu,N. (1992) Remarks on parallel analysis. *Multivar. Behav. Res.*, **27**, 509–540.

Carpentier,A.S. *et al.* (2004) The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. *Comput. Biol. Chem.*, **28**, 3–10.

Comon,P. (1994) Independent component analysis, a new concept? *Signal Process.*, **36**, 287–314.

Doane,A.S. *et al.* (2006) An estrogen receptor-negative breast cancer subset characterized by a hormonally regulated transcriptional program and response to androgen. *Oncogene*, **25**, 3994–4008.

Frigyesi,A. *et al.* (2006) Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics*, **7**, 290.

Huang,D.S. and Zheng,C.H. (2006) Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics*, **22**, 1855–1862.

Hyvaerinen,A. *et al.* (2001) *Independent Component Analysis*. Wiley, New York.

Hyvaerinen,A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, **10**, 626–634.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Laird,P.W. (2010) Principles and challenges of genome-wide dna methylation analysis. *Nat. Rev. Genet.*, **11**, 191–203.

Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.*, **3**, 1724–1735.

Leek,J.T. and Storey,J.D. (2008) A general framework for multiple testing dependence. *Proc. Natl Acad. Sci. USA*, **105**(48), 18718–18723.

Lee,S.I. and Batzoglou,S. (2003) Application of independent component analysis to microarrays. *Genome Biol.*, **4**, R76.

Liebermeister,W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.
Loi,S. *et al.* (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.*, **25**, 1239–1246.
Maegawa,S. *et al.* (2010) Widespread and tissue specific age-related DNA methylation changes in mice. *Genome Res.*, **20**, 332–340.
Martoglio,A.M. *et al.* (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, **18**, 1617–1624.
Moore,L.E. *et al.* (2008) Genomic DNA hypomethylation as a biomarker for bladder cancer susceptibility in the spanish bladder cancer study: a case-control study. *Lancet Oncol.*, **9**, 359–366.
Plerou,V. *et al.* (2002) Random matrix approach to cross correlations in financial data. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **65**, 066126.
Rakyan,V.K. *et al.* (2010) Human aging-associated dna hypermethylation occurs preferentially at bivalent chromatin domains. *Genome Res.*, **20**, 434–439.
Saidi,S.A. *et al.* (2004) Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene*, **23**, 6677–6683.
Sanchez,B.N. *et al.* (2009) An estimating equations approach to fitting latent exposure models with longitudinal health outcomes. *Ann. Appl. Stat.*, **3**, 830–856.
Schmidt,M. *et al.* (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.*, **68**, 5405–5413.
Sotiriou,C. *et al.* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl Cancer Inst.*, **98**, 262–272.
Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
Teschendorff,A.E. *et al.* (2007) Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.*, **3**, e161.
Teschendorff,A.E. *et al.* (2009) An epigenetic signature in peripheral blood predicts active ovarian cancer. *PLoS One*, **4**, e8274.
Teschendorff,A.E. *et al.* (2010) Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. *Genome Res.*, **20**, 440–446.
van 't Veer,L.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
Zhang,X.W. *et al.* (2005) Molecular diagnosis of human cancer type by gene expression profiles and independent component analysis. *Eur. J. Hum. Genet.*, **13**, 1303–1311.
Zheng,C.H. *et al.* (2009) Tumor clustering using nonnegative matrix factorization with gene selection. *IEEE Trans. Inf. Technol. Biomed.*, **13**, 599–607.

APPENDIX A

We simulated data matrices with 2000 features and 50 samples and considered the case of two CFs in addition to the primary phenotype of interest. The primary phenotype is a binary variable I_1 with 25 samples in one class ($I_1=0$) and the other half with $I_1=1$. Similarly, each CF is assumed to be a binary variable affecting one half of the samples (randomly selected). For a given sample s, we thus have a 3-tuple of indicator variables $I_s=(I_{1s},I_{2s},I_{3s})$ where I_2 and I_3 are the indicators for the two CFs. Thus, samples fall into eight classes. For instance, if $I_s=(0,0,0)$ then this sample belongs to phenotype class 1 and is not affected by the two CFs. Similarly, $I_s=(0,1,0)$ means that the sample belongs to class 1 and is affected by the first CF but not the second.

We assume 10% of features (200 features) to be TPs discriminating between the two phenotypic classes. We model the CFs as follows: each confounding factor is assumed to affect 10% of features with a 25% overlap with the TPs (i.e. 50 of the 200 TPs are confounded by each factor). Let J_g denote the indicator variable of feature g, so J_g is a 3-tuple (J_{1g},J_{2g},J_{3g}) with J_{1g} an indicator for the feature to be a true positive and J_{2g} (J_{3g}) an indicator for the feature to be affected by the first (second) CF. Thus, the space of features is also divided into eight groups. Furthermore, let (e_1,e_2,e_3) denote the effect sizes of the primary variable and the two CFs, respectively, where we assume for simplicity that $e_2=e_3$. Without loss of generality, we further assume that noise is modeled by a Gaussian of mean zero and unit variance $N(0,1)$. Thus, for a given sample s we draw data values for the various feature groups as follows:

(1) $J_g=(0,0,0)$: null unaffected features

$$p(x|I_s) \sim \delta_{J_g,000}N(0,1)$$

(2) $J_g=(0,1,0)$ or $(0,0,1)$: null features affected by only one CF

$$\begin{aligned}p(x|I_s) \sim \delta_{J_g,010}\{&\delta_{I_s,x1z}N(e_2,1)\\&+\delta_{I_s,x0z}N(0,1)\}\\&+\delta_{J_g,001}\{\delta_{I_s,xy1}N(e_3,1)\\&+\delta_{I_s,xy0}N(0,1)\}\end{aligned}$$

(3) $J_g=(0,1,1)$: null features affected by the two CFs

$$\begin{aligned}p(x|I_s) \sim \delta_{J_g,011}\{&\delta_{I_s,x11}N(e_2+e_3,1)\\&+\delta_{I_s,x01}N(e_3,1)\\&+\delta_{I_s,x10}N(e_2,1)\\&+\delta_{I_s,x00}N(0,1)\}\end{aligned}$$

(4) $J_g=(1,0,0)$: true positives not affected by CFs

$$\begin{aligned}p(x|I_s) \sim \delta_{J_g,100}\{&\delta_{I_s,0yz}N(0,1)\\&+\delta_{I_s,1yz}(\pi_{-1}N(-e_1,1)+\pi_1N(e_1,1))\}\end{aligned}$$

(5) $J_g=(1,0,1)$ or $(1,1,0)$: true positives affected by one CF

$$\begin{aligned}p(x|I_s) &\sim \delta_{J_g,101}\{\delta_{I_s,0y0}N(0,1)+\delta_{I_s,0y1}N(e_3,1)\\&\quad+\delta_{I_s,1y0}(\pi_{-1}N(-e_1,1)+\pi_1N(e_1,1))\\&\quad+\delta_{I_s,1y1}(\pi_{-1}N(-e_1+e_3,1)\\&\quad+\pi_1N(e_1+e_3,1))\}\\&\sim \delta_{J_g,110}\{\delta_{I_s,00z}N(0,1)+\delta_{I_s,01z}N(e_2,1)\\&\quad+\delta_{I_s,10z}(\pi_{-1}N(-e_1,1)+\pi_1N(e_1,1))\\&\quad+\delta_{I_s,11z}(\pi_{-1}N(-e_1+e_2,1)\\&\quad+\pi_1N(e_1+e_2,1))\}\end{aligned}$$

(6) $J_g=(1,1,1)$: true positives affected by all CFs

$$\begin{aligned}p(x|I_s) \sim \delta_{J_g,111}\{&\delta_{I_s,000}N(0,1)\\&+\delta_{I_s,010}N(e_2,1)+\delta_{I_s,001}N(e_3,1)\\&+\delta_{I_s,011}N(e_2+e_3,1)\\&+\delta_{I_s,101}(\pi_{-1}N(-e_1+e_3,1)\\&+\pi_1N(e_1+e_3,1))\end{aligned}$$

$$+\delta_{I_s,110}(\pi_{-1}N(-e_1+e_2,1)$$
$$+\pi_1 N(e_1+e_2,1))$$
$$+\delta_{I_s,111}(\pi_{-1}N(-e_1+e_2+e_3,1)$$
$$+\pi_1 N(e_1+e_2+e_3,1))\Big\}$$

where in the above $\delta_{x'y'z',xyz}$ denotes the triple Kronecker delta: $\delta_{x'y'z',xyz}=1$ if and only if $x'=x$, $y'=y$ and $z'=z$, otherwise $\delta_{x'y'z',xyz}=0$, and (π_{-1},π_1) are weights satisfying $\pi_{-1}+\pi_1=1$. In our case, we used $\pi_1=\pi_{-1}=0.5$.