BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 19 2014, pages 2693–2701
doi:10.1093/bioinformatics/btu361

Genome analysis

Advance Access publication June 11, 2014

Network-guided regression for detecting associations between DNA methylation and gene expression

Zi Wang[1], Edward Curry[2] and Giovanni Montana[1,3,*]

[1]Department of Mathematics, Imperial College London, London SW7 2AZ, [2]Division of Cancer, Imperial College London, Hammersmith Hospital, London W12 0NN and [3]Department of Biomedical Engineering, King's College London, St Thomas' Hospital, London SE1 7EH, UK

Associate Editor: John Hancock

ABSTRACT

Motivation: High-throughput profiling in biological research has resulted in the availability of a wealth of data cataloguing the genetic, epigenetic and transcriptional states of cells. These data could yield discoveries that may lead to breakthroughs in the diagnosis and treatment of human disease, but require statistical methods designed to find the most relevant patterns from millions of potential interactions. Aberrant DNA methylation is often a feature of cancer, and has been proposed as a therapeutic target. However, the relationship between DNA methylation and gene expression remains poorly understood.

Results: We propose Network-sparse Reduced-Rank Regression (NsRRR), a multivariate regression framework capable of using prior biological knowledge expressed as gene interaction networks to guide the search for associations between gene expression and DNA methylation signatures. We use simulations to show the advantage of our proposed model in terms of variable selection accuracy over alternative models that do not use prior network information. We discuss an application of NsRRR to The Cancer Genome Atlas datasets on primary ovarian tumours.

Availability and implementation: R code implementing the NsRRR model is available at http://www2.imperial.ac.uk/~gmontana

Contact: giovanni.montana@kcl.ac.uk

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

Received on March 11, 2014; revised on May 21, 2014; accepted on May 22, 2014

1 INTRODUCTION

Widespread adoption of high-throughput molecular profiling technology by the biological research community has resulted in a vast public resource of paired genomic data, where thousands of measurements of different types are available for the same biological samples. This in turn presents the opportunity to develop appropriate statistical tools to infer relationships between such paired datasets in different biological contexts (Minas *et al.*, 2013). Here we are particularly interested in identifying associations between paired DNA methylation profiles and expression levels of genes obtained from cancer samples.

DNA methylation involves addition of a methyl group to cytosines in DNA. Where this is concentrated in promoter regions, it tends to lead to silencing of expression of the downstream gene. The association between DNA methylation and gene expression is far from straightforward, and DNA methylation has also been shown to be associated with active transcription of genes (Suzuki and Bird, 2008). Aberrant DNA methylation has been observed in almost every type of cancer and patterns of DNA methylation have been linked to changes in gene expression that associate with the development of drug resistance. Obtaining insights into the links between DNA methylation and gene expression in cancer is expected to be particularly helpful in moving towards addressing the issue of acquired drug resistance, which remains one of the main unmet clinical needs in cancer medicine (Vaughan *et al.*, 2011).

A number of efforts have been made recently to address the problem of detecting associations between DNA methylation and gene expression profiles. By far, the most common scenario involves the mapping of loci with DNA methylation to genes, and using some measure of association between the levels of DNA methylation and the levels of expression of the same gene. Example measures of association include Pearson's correlation coefficient (Rhee *et al.*, 2013) and maximal information coefficient (Stone *et al.*, 2013). In case–control designs, a common approach consists in first identifying differentially methylated sites and differentially expressed traits, often using standard two-sample univariate statistical tests (e.g. a *t*-test), and then comparing the two resulting lists to detect potential overlaps (Gervin *et al.*, 2012)

In this article, we take a predictive modelling approach in which we aim to identify a subset of DNA methylation measurements that are highly predictive of a subset of gene expression traits across a set of heterogeneous biological samples that, nonetheless, share some characteristic of interest (e.g. patients with the same disease, and for which control samples may not be available). The precise mechanisms by which DNA methylation influence gene expression are far from fully understood, particularly when it comes to inter- and intra-genic methylation (Suzuki and Bird, 2008), and *trans*-acting regulation in which the 3D structure of chromatin enables interactions between non-adjacent regions of the genome (Lazarovici *et al.*, 2013). A recent study investigating genome-wide epigenetic and transcriptional changes following acquisition of platinum resistance in ovarian cancer cell lines showed that only a small proportion of genes with an increase in DNA methylation had a corresponding down-regulation of gene expression (Zeller *et al.*, 2012). For this reason, and to identify patterns of regulation where multiple

*To whom correspondence should be addressed.

sites of methylation can drive a systematic alteration of a functional expression program (akin to a higher eukaryotic equivalent of a bacterial operon), we do not wish to restrict our discovery to only those interactions for which each CpG locus is mapped to a gene and directly related to that same gene's expression level. Instead, we would like to be able to identify clear relationships between varying levels of DNA methylation at any loci and corresponding changes in expression of any genes, ideally with some relevance to some biological function. The simplest approach to this problem would consist of fitting all possible univariate linear regression models, in which each gene expression value is regressed on each DNA methylation level, and all the resulting pairs are ranked based on the magnitude and/or statistical significance of the regression coefficients. We will refer to this method as mass-univariate linear modelling (MULM). Such an approach, however, has two major drawbacks: first, because of the large number of hypothesis tests that are simultaneously carried out, the experiment-wide significance level would be far too stringent; second, all predictors and responses are modelled as statistically independent while also ignoring any functional relationship between genes.

To tackle these limitations, we present a multivariate regression model for the simultaneous selection of highly predictive DNA methylation predictors and the most predictable gene expression profiles. The high dimensionality and the small sample size pose the challenging problems of coefficient estimation and variable selection. Models involving dimension reduction techniques in conjunction with ℓ_1 penalization on the coefficients, such as sparse partial least squares (sPLS), have been explored in related applications, including the integrated analysis of paired 'omics' data (lé Cao *et al.*, 2008). In this work, we adopt a reduced-rank regression (RRR) approach, which assumes a linear association between DNA methylation and gene expression variables; the predictor vectors are projections of the methylation variables matrix into a lower dimensional space, and the response vectors are projections of the gene expression variables into a lower dimensional space.

In light of the complex interactions between nucleic acids, proteins and metabolites in all organisms, networks have become fundamental in the understanding of biological systems. In fact, these networks appear to underpin biological processes found throughout living things. For instance, in a recent methylation–gene expression study, while validating the findings of top-ranking gene pairs, a substantially larger proportion of them were found to be connected in the Protein–Protein Interaction (PPI) network compared with the expected number from a random distribution (Joung *et al.*, 2013). Multiple interacting components work together in functional modules to bring about disease phenotypes or other traits (Calvano *et al.*, 2005). From a statistical standpoint, the utilization of prior information encoded in biological networks provides a form of regularization, and has been shown to improve inference from high-dimensional datasets (Chuang *et al.*, 2007). We seek to improve variable selection accuracy and the interpretability of our regression model by guiding variable selection to highlight sets of co-methylated and co-expressed genes that are known to co-operate in enacting some biological processes. This is achieved by extending our previous work (Vounou *et al.*, 2010, 2012) on penalized regression through the use of an additional penalty function on the regression coefficients, which incorporates prior information regarding the functional relatedness between predictor variables and between response variables, respectively. We call the resulting model Network-sparse Reduced-Rank Regression (NsRRR). When the response variable is univariate, incorporating such structural information to encourage the selection of mutually interacting genes has been shown to give improved variable selection result in terms of both accuracy and interpretability (Li and Li, 2008; Azencott *et al.*, 2013). Here we extend this principle to multivariate responses arising when studying genome-wide expression levels. An illustration of the NsRRR principle is given in Figure 1, in which trivial networks of predictor variables and response variables are shown, with an example of possible sets of selected variables highlighted in a darker colour. For example, if variation in the levels of the darker response variables is explained by variation in the darker predictor variables, they are likely to be selected in the NsRRR model because each set is also connected within its respective network topology.

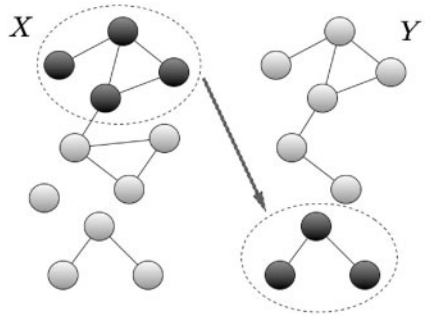

Fig. 1. Illustration of the NsRRR principle: predictor variables (X) and response variables (Y) are each organized into a network via an adjacency matrix encoding prior information on relatedness of the variables. A set of predictor variables (e.g. dark coloured *X*'s) will be selected by the NsRRR model if they show an association with a set of response variables (e.g. dark coloured *Y*'s). Within the sets of selected predictors and responses, variables are more likely to be selected together if they are also connected in the respective networks (as is the case with both the dark *X*'s and the dark *Y*'s)

The remainder of the article is organized as follows: In Section 2, we introduce the NsRRR model, the estimation algorithm and a procedure for parameter tuning. In Section 3, we report on extensive simulation studies set up to explore the properties of the model, and compare its performance on alternative approaches using both artificial and real datasets. In Section 4, we present an integrative analysis of ovarian cancer data containing genome-wide profiles of both DNA methylation and gene expressions.

2 METHODS

2.1 Problem formulation

We denote by X^* the $n \times p$ design matrix of DNA methylation predictors, and by Y^* the $n \times q$ response matrix of gene expression measurements. We normalize the columns of the data matrices to have zero mean and unit length, and denote the resulting matrices by X and Y, respectively. We write X_j for the n-dimensional column vector extracted from the j^{th} column of X, and likewise Y_k for the k^{th} column of Y.

The prior information to describe functional relatedness of the variables is expressed in the form of biological networks. These networks quantify the relatedness between any two genes, and can be built using a number of external sources. The application we present here was developed in the context of identifying systematic epigenetic regulation of pathway-level gene expression programs, and so chose to use a measure of the functional similarity based on shared pathway annotations between the genes mapped to each data entity (see Section 4). However, in other contexts there might be a number of alternative networks that could be advantageous for informed variable selection. For example, a regression task looking to use gene co-expression patterns to infer the functional roles of unknown genes or proteins might use the networks to encode sequence similarity or shared protein domains.

We denote by $G_x = (V_x, E_x)$ the DNA methylation network and by $G_y = (V_y, E_y)$ the gene expressions network. The vertex set V_x contains the p predictors in X, and the edge set E_x is given in terms of a weighted adjacency matrix W^x on V_x. The edge weight is a real number in [0,1]. A large weight indicates the corresponding genes pair is strongly functionally related.

2.2 Network-driven sRRR

The standard regression model links the predictors and responses by $Y = XC + E$, where C is the $p \times q$ coefficient matrix in which the (j, k) entry quantifies the association between the j^{th} DNA methylation and k^{th} gene expression, and E is the matrix of random errors with zero mean. An estimate of C, denoted by $\hat{C}$, is traditionally obtained by minimizing the loss function $||Y - XC||^2_{\mathcal{F}}$, where $||.||_{\mathcal{F}}$ denotes the Frobenius norm (Reinsel and Velu, 1988). When $p \leq n$ so that $X'X$ is invertible, a closed form solution exists, i.e. $\hat{C} = (X'X)^{-1}X'Y$, which is of full rank. However, minimizing the squared loss on the multivariate responses does not account for the dependence structure in the responses.

Our proposal is to impose a rank constraint on the coefficient matrix C, i.e. $\text{rank}(C) = R \leq min\{p, q\}$. For a given R, C can be expressed as the product of two matrices B and A, each having full rank,

$$C = BA = \sum_{r=1}^{R} b^{(r)} a^{(r)} = \sum_{r=1}^{R} C^{(r)} \tag{1}$$

where B is a $p \times R$ matrix whose r^{th} column is $b^{(r)}$, and A is an $R \times q$ matrix whose r^{th} row is $a^{(r)}$. The $p \times q$ matrix $C^{(r)}$ denotes the r^{th} layer of C, ranked in decreasing order of the strength of association. A desirable feature of the matrix decomposition $C = BA$ is that it allows separate evaluation on the effect of each predictor and each response: for instance, the j^{th} entry of $b^{(r)}$ quantifies the contribution of the j^{th} DNA methylation site and the k^{th} entry in of $a^{(r)}$ quantifies the effect on the k^{th} gene expression in the r^{th} rank. The rank constraint gives rise to a reduction in the number of parameters to estimate and accounts for the multivariate nature of the response variables.

Minimizing with respect to the empirical loss $||Y - XBA||^2_{\mathcal{F}}$, the estimates $\hat{B}$ and $\hat{A}$ can be obtained as, respectively,

$$\hat{B} = (X'X)^{-1}X'YH^{(R)}, \hat{A} = (H^{(R)})' \tag{2}$$

where $H^{(R)}$ is the $q \times R$ matrix in which the r^{th} column is the normalized eigenvector corresponding to the r^{th} largest eigenvalue of $Y'X(X'X)^{-1}X'Y$. The normalization condition ensures that (1) is uniquely defined because otherwise for any $R \times R$ non-singular square matrix T, we have matrices $\tilde{B}$ and $\tilde{A}$ such that: $C = BA = (BT)(T^{-1}A) = \tilde{B}\tilde{A}$.

The constraint on $H^{(R)}$, which guarantees uniqueness of (1), can be reformulated through $\hat{B}$ and $\hat{A}$. Note that the real matrix $Y'X(X'X)^{-1}X'Y$ is symmetric and thus normal. By the spectral theorem, we have:

$$Y'X(X'X)^{-1}X'Y = H^{(R)}(\Theta^2)^{(R)}(H^{(R)})' \tag{3}$$

where $(\Theta^2)^{(R)}$ is the $R \times R$ diagonal matrix whose diagonal corresponds to the eigenvalues of $Y'X(X'X)^{-1}X'Y$. Pre-multiplying (3) by $(H^{(R)})'$ and post-multiplying by $H^{(R)}$, while noting $(H^{(R)})'H^{(R)} = I$, we have: $(H^{(R)})'Y'X(X'X)^{-1}X'YH^{(R)} = (\Theta^2)^{(R)}$. As such, $\hat{B}$ must satisfy:

$$\hat{B}'X'X\hat{B} = (H^{(R)})'Y'X(X'X)^{-1}X'YH^{(R)} = (\Theta^2)^{(R)} \tag{4}$$

Because the columns of $H^{(R)}$ are orthonormal vectors, $\hat{A}$ must satisfy:

$$\hat{A}\hat{A}' = (H^{(R)})'H^{(R)} = I. \tag{5}$$

Equations (4) and (5) imply that $(X\hat{B})\hat{A}$ corresponds to the singular value decomposition of $Y - E$, in which the R linear combinations of X via $(X\hat{B})$ are orthogonal and thus linearly independent, and so do the columns of $\hat{A}$.

The computation in (2) and (3) involve estimating the inverse covariance matrix, $(X'X)^{-1}$, which may not exist when $n < p$. A common option consists of taking $(X'X + \gamma I)^{-1}$, where I is the identity matrix and γ is a tuning scalar parameter (Chen *et al.*, 2012). Here we set $(X'X)^{-1}$ equal to I, which is also commonly done in high-dimensional settings (Tenenhaus *et al.*, 2014; Witten *et al.*, 2009), and does not require parameter tuning. Moreover, this parameterization can be interpreted as applying an extreme de-correlation. This can be seen by taking a large γ in the shrinkage estimator above while preserving the variances. Other alternatives include using a generalized inverse, but this would be too computationally demanding in our setting.

In our context, we expect that any existing association will only involve a small subset of DNA methylations and a small subset of gene expressions, which need to be identified. Variable selection can be carried out by adding a penalty on the ℓ_1 norm of the regression coefficients, which leads to a continuous shrinkage of the coefficient estimates towards zero (Tibshirani, 1996). We propose the use of an additional penalty that regularizes the solution by taking into account the functional relatedness among genes as encoded by the given networks. To our knowledge, this would be the first model, which gives network-regularized sparse association between DNA methylation and multiple-gene expression measurements. The RRR model grants us the flexibility of applying separate penalties for the coefficients associated to the DNA methylations variables and the gene expression variables. As such, we propose to minimize the following objective function:

$$||Y - X\sum_{r=1}^{R} b^{(r)}a^{(r)}||^2_{\mathcal{F}} + \sum_{r=1}^{R}(P_x(b^{(r)}) + P_y(a^{(r)})) \tag{6}$$

where $P_x(b^{(r)})$ is the penalty accounting for the pattern in DNA methylation, and $P_y(a^{(r)})$ accounts for the pattern in gene expression. Both penalty functions have a similar expression of form,

$$P_x(b^{(r)}) = 2\lambda_b||b^{(r)}||_1 + L(b^{(r)}) \tag{7}$$

involving an ℓ_1 penalty and a normalized Laplacian penalty. When the Laplacian penalty is not included, the model reduces to the sRRR model.

We write d^x_j for the node degree of X_j defined as $d^x_j = \sum_{t\sim j} w^x_{jt}$, where $t \sim j$ if and only if $w^x_{jt} \neq 0$. The vertex degree is a measure of centrality, so that nodes with large degree ('hub nodes') correspond to the active genes, which interact with many others. Similarly, we define W^y and d^y_i. The normalized Laplacian penalty regularizes the coefficients by prior information encoded in G_x and G_y. For the coefficients associated to DNA methylations, this penalty is defined as

$$L(b) = \mu\sum_{t\sim j} w^x_{jt}(\frac{b_j}{\sqrt{d^x_j}} - \frac{b_t}{\sqrt{d^x_t}})^2 \tag{8}$$

that is, it penalizes the square of normalized difference between b_j and b_t if the j^{th} and t^{th} predictors are connected in G_x, based on the assumption that functionally related variables exert similar effects. Here, we have dropped the superscript r for convenience.

The Laplacian penalty is in the form of the sum of squares, which is a strictly convex function. This enables (8) to smooth the estimates $\hat{b}_j$ and $\hat{b}_t$ such that $\hat{b}_j/\sqrt{d^x_j}$ and $\hat{b}_t/\sqrt{d^x_t}$ become close (Chung, 1997). In particular, if a sparse solution includes positive $\hat{b}_j$ and $\hat{b}_t = 0$, the normalized

Laplacian penalty (8) can drive $\hat{b}_t$ towards a non-zero value thus to encourage the selection of the j^{th} and t^{th} methylation altogether. As such, the functionally related pair is selected, and the prior structural knowledge has been accounted to guide the selection of a biologically plausible model. The coefficients $\hat{b}$ are rescaled according to the vertex degree to highlight the importance of high centrality nodes, while ensuring each variable is equally penalized regardless of its node degree in G_x (Chung, 1997). The regularization parameter μ, a non-negative real number, controls the strength of this smoothing effect. As μ increases from zero, the normalized $\hat{b}_j$ and $\hat{b}_t$ get closer to each other if the corresponding variables are connected via paths in G_X. As μ goes to infinity, the individual coefficient $\hat{b}_j$ tends to a weighted mean of the correlation coefficients between the responses and the predictors indexed $\{t : t \sim j\}$, where the weighted mean is computed with respect to the nodes degrees. Further details on the smoothing effects imposed by the graph penalty as μ goes to infinity, can be found in Supplementary Material, Section A.

The smoothing effect triggered by μ has a decisive impact on the topological patterns of the selected variables by the NsRRR. When $\mu = 0$, no smoothing effect occurs, as the penalty terms acting on the networks are not taken into account; as μ increases, the selected variables tend to fall into several densely connected subgraphs, or 'communities' in the given networks. On the other hand, when μ is sufficiently large, the model is enforced to select variables with large node degrees, which usually lie in one or few components of the networks. Here, a component refers to a subgraph in which every pair of nodes is connected via paths and which is disconnected to the nodes not in the subgraph. The optimal amount of regularization that the prior knowledge contributes to the statistical model depends on graph density, component structure and how informative the networks are with respect to the data. The issue of parameter tuning will be addressed in Section 2.4.

2.3 Estimation algorithm

In this section, we outline the key steps in the estimation procedure to obtain $\hat{b}^{(r)}$ and $\hat{a}^{(r)}$, which minimize (6) subject to penalties (7) and (8) and under the constraint conditions (2) and (4). We begin with the rank-one NsRRR model. We introduce Lagrange multipliers δ_a and δ_b to solve the constraint optimization problem. After rearrangement, (6) becomes:

$$-2aY'Xb + aa'b'b + \delta_a aa' + \delta_b b'b + P_x(b) + P_y(a) \tag{9}$$

where δ_a and δ_b are constants such that $\check{b}'\hat{b} = \theta^2$ and $\hat{a}\hat{a}' = I$ and θ^2 is the largest eigenvalue of $Y'XX'Y$. Note that (9) is biconvex, and therefore, it can be solved by recursively fixing one of b and a and optimizing with respect to the other one alone, using a co-ordinate descent algorithm (Friedman *et al.*, 2007). A full derivation of the algorithm can be found in the Supplementary material, Section C, with the pseudocode given in Algorithm 1.

Algorithm 1 rank-one NsRRR

Input: data X, Y; parameters λ_a, λ_b, μ; initial values a^0, b^0; weighted adjacency matrices W^x, W^y; ϵ
Output: column vector $\hat{b}$, row vector $\hat{a}$
1: Define: $S_\lambda(x) = \text{sign}(x) \cdot (|x| - \lambda)$
2: $\theta^2 \leftarrow$ largest eigenvalue of $Y'XX'Y$
3: $\tilde{a} \leftarrow \frac{a^0}{||a^0||_2}$; $\tilde{b} \leftarrow \frac{b^0}{||b^0||_2}\theta$
4: **for** j in 1:p **do**
5: $\check{b}_j \leftarrow S_{\lambda_b}(\tilde{a}Y'X_j + \mu \sum_{t \neq j} \frac{w^x_{jt}}{\sqrt{d^x_j d^x_t}} \tilde{b}_t)$
6: **end for**
7: $\hat{b} = \frac{\check{b}}{||\check{b}||_2}\theta$
8: **for** k in 1:q **do**
9: $\check{a}_k \leftarrow S_{\lambda_a}(\check{b}'X'Y_k + \mu \sum_{s \neq k} \frac{w^y_{ks}}{\sqrt{d^y_k d^y_s}} \tilde{a}_s)$
10: **end for**
11: $\hat{a} = \frac{\check{a}}{||\check{a}||_2}$
12: **if** $||\hat{b} - \tilde{b}||_2 \leq \epsilon$ AND $||\hat{a} - \tilde{a}||_2 \leq \epsilon$ **then**
13: **Return**: $\hat{b}$, $\hat{a}$
14: **else**
15: $\tilde{a} \leftarrow \hat{a}$; $\tilde{b} \leftarrow \hat{b}$; go back to 4
16: **end if**

Once the estimates corresponding to the first rank have been obtained, the higher rank estimates can be extracted one at a time by first regressing out the fitted responses predicted by the selected predictors found at the current rank, and then re-fitting the model, as follows. Starting with $r = 1$, and $Y^{(r)} = Y^*$, suppose the r^{th} rank coefficients $\hat{b}^{(r)}$ and $\hat{a}^{(r)}$ have been obtained, and a decision has been made on which X's and Y's have been selected (see Section 2.4). We define the submatrices X_{sub} and Y_{sub} from X^* and $Y^{(r)}$, respectively, as consisting of the selected variables only. A rank-one RRR is then fit on (X_{sub}, Y_{sub}) and the fitted response $\hat{Y}_{sub}$ is obtained. The submatrix Y_{sub} in $Y^{(r)}$ is substituted by $Y_{sub} - \hat{Y}_{sub}$ and we call the resultant matrix $Y^{(r+1)}$. Finally, we normalize the columns of $Y^{(r+1)}$ and denote the output by Y. The $(r+1)^{th}$ rank estimates are obtained by applying the unit-rank NsRRR algorithm on the updated response matrix Y and the same X matrix as in the previous computations. This procedure is applied for each subsequent rank r.

2.4 Parameter tuning procedures

The rank-one NsRRR model has three regularization parameters, i.e. λ_a, λ_b and μ. In our setting, the first two parameters control the number of non-zero coefficients, whereas μ varies the degree of regularization that is imposed by the network structure. These regularization parameters are traditionally determined by a cross-validation procedure in which the model fit is assessed by its prediction error. Nonetheless, λ_a and λ_b optimized in such a way often result in many noise variables being selected, in particular, when $p, q >> n$ as in our case. When no prior connectivity knowledge is used, so that $\mu = 0$, a data resampling procedure called stability selection can be adopted to improve variable selection accuracy (Meinshausen and Bühlmann, 2010). Stability selection consists of fitting the sparse regression model to a large number of subsamples of the data matrices, where each subsample generally comprises half of the subjects. Variable selection results across all subsamples are integrated to compute empirical selection probabilities $\hat{p}^X$ and $\hat{p}^Y$. The sets of important variables are defined as $\hat{S}_X = \{j : \hat{p}^X_j \geq \pi_X\}$ and $\hat{S}_Y = \{k : \hat{p}^Y_k \geq \pi_Y\}$, where π_X and π_Y are real numbers in [0, 1]. One of the appealing features of this procedure is that the ranking of the variables, and in particular the high-ranking variables, is generally insensitive to the particular choice of the regularization parameters (Meinshausen and Bühlmann, 2010; Vounou *et al.*, 2012). For our model, we propose a hybrid cross-validation strategy to search for an optimal μ within a set Θ of possible candidate values, which uses an internal stability selection procedure for ranking the selected variables corresponding to each μ value.

Specifically, we use a 10-fold cross-validation. For each fold, we denote the training dataset as $\mathcal{D}_{\text{train}}$ and the testing dataset as $\mathcal{D}_{\text{test}}$. Within each fold, we firstly generate M subsamples extracted without replacement from $\mathcal{D}_{\text{train}}$, each consisting of half the subjects in $\mathcal{D}_{\text{train}}$. For each subsample, we fit a rank-one NsRRR model with fixed λ_a and λ_b and identify variables corresponding to the non-zero entries of $\hat{b}$ and $\hat{a}$. To assess the prediction error of a model fitted on the selected variables, we fit a rank-one RRR model using those variables and evaluate the prediction error in Frobenius norm. Ultimately, we select the value of the μ corresponding to the minimum prediction error. This procedure is repeated for all 10 folds, and during this process, we keep track of the number of times that each μ has been deemed optimal across the $10 \times M$ subsampled data. The candidate with the most counts is the chosen $\tilde{\mu}$. Once $\tilde{\mu}$ is obtained, we apply again a stability selection procedure by taking $\mu = \tilde{\mu}$ and selecting λ_a and λ_b to achieve the same sparsity previously obtained in $\hat{S}_X$ and $\hat{S}_Y$.

3 SIMULATION STUDIES

In this section, we present simulation studies to characterize the power of NsRRR model to detect the true predictors and responses, and compare its performance with the sparse sRRR model that does not take any prior knowledge into account, and to a MULM approach in which a linear model is fitted for all possible combinations of predictors and responses. The sRRR estimation algorithm is also related to the sPLS algorithms (lé Cao *et al.*, 2008; Vounou *et al.*, 2010), and they have similar performance in variable selection tasks. We present two types of simulation studies: (i) using randomly generated variables and (ii) using variables extracted from a real dataset.

In all simulations, we use $n = 100$, $p = 1000$ and $q = 1000$ so that $n << p$ and q. For each study type, we perform five sets of simulations, corresponding to different signal-to-noise ratios (SNRs) (0.01, 0.005, 0.0025, 0.001 and 0.0005). In each case, we randomly generate 2000 datasets, each consisting of data matrices X, and Y as well as networks G_X and G_Y. The signal-carriers variables are always known, so the performance of each model can be evaluated by comparing the number of correctly identified predictors and responses while controlling for the number of falsely detected variables. For the simulations of type (I), we generate the $n \times p$ predictor matrix X by simulating from independent uniform distributions in the unit interval. The associations between X and Y are introduced by assuming that exactly 80 predictors and 80 responses contribute to the non-random associations. For simplicity, we further assume the non-random associations have rank-one. By (1), we obtain the responses as $Y = Xba + \eta \cdot E$, where E is a $p \times q$ matrix of random errors, generated from independent standard normal distributions, and η is a positive real number set to achieve the desired SNRs. Because causal methylation sites tend to down-regulate gene expressions, we generate the non-zero entries of b and a such that the signal carrying X's and Y's are negatively correlated. For the simulations of type (II), we use real DNA methylation and associated gene expression data obtained from the The Cancer Genome Atlas (TCGA), as further described in Section 4. Starting from the full matrices, each artificial dataset is generated by randomly extracting an $n \times p$ submatrix X from the methylation matrix and an $n \times q$ submatrix Z from the gene expression matrix; we then permute the columns of Z to obtain $\tilde{Z}$ so as to remove any potential association; finally, we compute the response matrix Y as before by assuming a linear model of form $Y = Xba + \eta\tilde{Z}$. We normalize the columns of X and Y to have zero mean and unit Euclidean norm, as required by NsRRR and sRRR.

As for the network generation, we randomly partition the signal-carrying variables into clusters of 10 variables, and likewise partition the non–signal-carriers into clusters of 10, resulting in 100 clusters for X and the same number for Y. Assuming the probability that a pair of nodes are directly connected is independent to all other pairs, we generate the networks G_x and G_y using a set of three probability parameters: p_C is the probability of connection between variables belonging to the same cluster; p_{CC} is the probability of connection between signal-carriers or non–signal-carriers belonging to different clusters; p_{SN} is the probability of connection between a signal-carrier and a non–signal-carrier. We generate networks such that signal-carrying variables are relatively densely connected, whereas there are few links between these variables and the non–signal-carriers. Specifically, we set $p_C = 0.4$, $p_{CC} = 0.13$, $p_{SN} = 0.04$ to give an experimental average probability of connection of 0.12, matching that of the networks in the analysed data in Section 4.

We fit the rank-one NsRRR and sRRR models, as well as the MULM to each dataset, and we tune the parameters to ensure that exactly 160 X's and 160 Y's have non-zero coefficients for each model that is fitted. Note that the sparsity level does not affect the relative performance of the three models as long as the same sparsity is retained for all models so that the results are comparable. Initial test runs showed that the optimal μ for the simulated datasets, using the procedure described in Section 2.4, was rarely >5. Therefore, in the experiments presented here, we restrict the search of $\tilde{\mu}$ to a range of values in $\Theta = (0.5, 1, 2, 3, 5)$. Because in our experiments we generate a total number of 20 000 simulated datasets, the μ parameter is optimized only once for each simulation scenario, using the search procedure of Section 2.4, at a fixed SNR value of 0.0005. This approach reduces the computational burden without introducing any bias. However, as a result of this, the performance of NsRRR may be suboptimal in some cases. For type (I) simulations, we obtain $\tilde{\mu} = 3$ and type (II) simulations we obtain $\tilde{\mu} = 2$. We compute the empirical selection probabilities for all variables within each simulation set. By varying the threshold from 1 to 0 and defining the important variables as those with selection probabilities greater than the threshold, we can construct the receiver operating characteristic (ROC) curve for X and Y, respectively. In a ROC curve, the proportion of signal-carriers classified as important variables (true-positive rate) is plotted against the proportion of non–signal-carriers classified as important variables (false-positive rate). The area under curve (AUC) can be interpreted as the probability that a randomly chosen signal carrying variable having larger selection probability than a randomly chosen variable carrying no signal, which will be used as the evaluation criterion for model comparison.

The results of these studies are summarized in Figure 2, where the AUC is plotted against the SNR. The first column (plots A and C) summarizes the performance of variable selection in the space of predictors for simulation types (I) and (II), respectively, whereas the second column (plots B and D) corresponds to analogous performance levels in the space of responses. For type (I) simulations, we observe that the sRRR and NsRRR models, which are based on a global search for association, consistently outperform the MULM approach, which is based on pairwise testing. When the SNR is >0.005, the difference in performance between sRRR and NsRRR models is only marginal. However, as the SNR continues to decrease down to 0.0005, NsRRR model continues to maintain a much higher performance compared with the sRRR model. This demonstrates that incorporating prior knowledge in the form of a connectivity matrix can enhance the power of variable selection and outperform the models that ignore this additional information. Similar patterns are observed in plots C and D, which correspond to type (II) simulations.

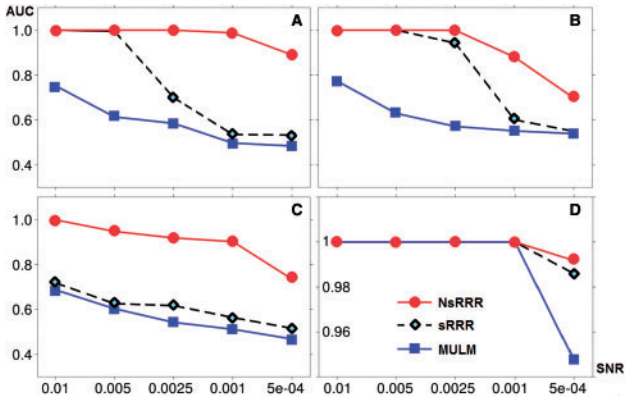

Fig. 2. AUC against SNR. The left column refers to variable selection in the predictor space, and the right column refers to variable selection in the response space. Each point on a curve is computed using 2000 randomly generated datasets. *A* and *B* refer to type (I) simulations that rely on artificially simulated data, whereas *C* and *D* refer to type (II) simulations obtained from a real TCGA dataset (see Section 4 for details) and simulated signals. Using the networks to guide the variable selection search enables NsRRR to gain additional power compared with alternative methods, especially when the SNR is low

4 AN APPLICATION TO OVARIAN CANCER

Aberrant DNA methylation in cancer has been proposed as a mechanism for facilitating the plasticity of gene expression states that enables tumour cells to adapt to chemotherapies, ultimately resulting in acquired drug resistance in cancer. Acquired drug resistance is the greatest unmet clinical need currently facing sufferers of a number of cancers, notably including ovarian cancer (Vaughan *et al.*, 2011). Identifying common associations between gene expression and DNA methylation in ovarian cancer could yield insight into the mechanisms by which tumour cells can exploit epigenetic dysregulation to adapt to changes in their surroundings. Therefore, this could indicate certain paths to acquired drug resistance.

We apply NsRRR to a paired dataset, containing genome-wide measurement profiles of both gene expression and DNA methylation for 349 primary ovarian tumours. DNA methylation levels are obtained from the ratios of background-corrected methylated and unmethylated probe intensities measured by Illumina HumanMethylation27k BeadArrays, downloaded from TCGA (2011). Gene expression levels are obtained by applying robust multi-array average (RMA) normalization (Irizarry *et al.*, 2003) to raw data CEL files from Affymetrix HT-HGU133A GeneChips, also downloaded from TCGA.

The functional relatedness between variables is encoded into networks based on shared pathway annotations. Pathway annotations are downloaded from Consensus Pathway DB (CPDB), a repository of pathway-level information from multiple databases, representing the consensus across the community regarding which genes are involved in which biological processes (Kamburov *et al.*, 2011). As the functional annotations are provided at the gene level, we first find which gene corresponds to each CpG site represented in the DNA methylation dataset, using the manufacturer's annotations (which are based on proximity). Similarly, we find which gene corresponds to each probeset in the gene expression dataset, again using the manufacturer's annotations. As such, weights between the methylation sites inherit the weights of the corresponding genes. We construct a weighted adjacency matrix describing the pairwise relationships between variables i and j based on Dice's coincidence index (Dice, 1945): $W_{ij} = 2|I \cap J|/(|I| + |J|)$, where I and J are the set of CPDB pathway annotations for the gene mapping to variable i and j, respectively. The number of elements in set A is denoted by $|A|$. Each W_{ij} score takes on a maximum value of 1 when the two annotation lists are identical, and a minimum value of 0 when there is no overlap between the two lists.

We apply the algorithm introduced in Section 2.4 to search for the optimal μ for our data from a broad range of candidates values, $\Theta = (0, 0.5, 5, 50, 500, 5000)$. The maximum value of μ was purposely chosen to be large to include the case of an extreme penalization. We set the number of subsamples in each of the 10 folds to $M = 200$ and fix λ_a and λ_b such that exactly 200 predictors and 200 responses are selected in each subsample. We present the full table of results in Supplementary Table ST1. The optimal μ parameter in this case was found to be $\tilde{\mu} = 50$. Next, we use stability selection procedure involving 5000 random subsampled datasets, fitting an NsRRR model to each one where $\mu = 50$ and exactly 200 X's and 200 Y's are selected. Estimates for the importance of each variable are given by the empirical selection probability throughout the subsamples. Here, we benefit from the robustness of the data resampling scheme in Section 2.4, which allows us to fix the number of selected variables in each subsampled data while not affecting much of the final ranking of the variables. Following the terminology of Section 2.4, we obtain a set $\hat{S}_X$ of 200 predictor variables with $P(X_j) > 0$ and a set $\hat{S}_Y$ of 4371 response variables with $P(Y_k) > 0$. For more manageable downstream analysis, a subset of the selected response variables $\hat{S}_Y' = Y_k, P(Y_k) > 0.5$ is chosen, which represents 116 gene expression probes whose levels are robustly and reliably predicted by the levels of $\hat{S}_X$. A list of the Affymetrix and Illumina probe identifiers and corresponding gene symbols for the selected variables $(\hat{S}_X, \hat{S}_Y)$ is given in Supplementary Table ST2. As the 200 top-ranking probes and 116 gene expressions all have large selection probabilities (all >0.5), and they constitute a manageable list of variables for further investigation, we decide not to proceed to the higher rank estimates of the NsRRR model.

Using the stability selection, it is also possible to extract a matrix of empirical association probabilities Φ, in which Φ_{jk} denotes the probability that the j^{th} methylation and the k^{th} gene expression are simultaneously selected across the subsamples. Interestingly, $\Phi_{jk} = \Phi_{j'k}, \forall j, j' \in \hat{S}_x, \forall k$, indicating that the set of selected predictor variables are always selected together, regardless of which response variables were selected. We illustrate this in Supplementary material, Section C (Figure 1), plotting the association scores Φ_{jk} across all response variables k, for the first three predictor variables in $\hat{S}_X$. These association profile plots show that each of the selected predictor variables show exactly the same pattern of selection probabilities across the response variables.

As pathway co-annotation scores are used in the variable selection process, it would be expected that the sets of predictor and response variables selected in the NsRRR models would fall into densely connected regions in the networks. In fact, the 200

selected predictors correspond to a densely connected subnetwork of G_X, where the average probability of connection is 0.9932, much higher than that for G_X, 0.116. For the 116 selected response variables, 107 of them comprise a large component in which every node is connected to all other nodes via paths. The average probability of connection for this large component is 0.143, higher than that for G_y, which is 0.125.

To explore the biological implications of the NsRRR results, we tested the selected sets of variables for enrichment of CPDB pathway terms based on the hypergeometric distribution. Of particular note, we found that the 'mTOR signalling' pathway was enriched ($p = 0.059$), indicating greater membership among the selected response variables than would be expected by chance. This observation is interesting because mTOR signalling has been implicated in many cancers, and more specifically, has been shown to activate survival pathways that can lead to platinum resistance in ovarian cancer (Peng *et al.*, 2010).

One of our hypotheses relating to this study was that genes with expression predicted by DNA methylation at a range of loci are likely to be affected by pharmacological manipulation of the genome-wide state of DNA methylation. Such manipulation can occur through treatment with the DNA methyltransferase inhibitor 5-aza-dC, principally through the loss of methyl groups at CpG sites across the genome. Three studies were identified with published genome-wide gene expression profiling data in cancer cell lines both pre- and post-treatment with 5-aza-dC (Mueller *et al.*, 2007; Dannenberg and Edenberg, 2006; Khamas *et al.*, 2012), with Gene Expression Omnibus (GEO) accession numbers GSE4717, GSE5230 and GSE32323, respectively. Normalized data from each study were downloaded from GEO and analyzed individually through application of the statistical package linear models for microarray data (LIMMA) (Smyth, 2004), obtaining empirical Bayes-moderated t-statistics for the difference in expression of each gene (in each study) following treatment with 5-aza-dC. Systematic changes in the expression levels of the genes selected in our NsRRR model were evaluated in each of the 5-aza-dC treatment studies using the mean-rank gene-set enrichment test of (Michaud *et al.*, 2008) (as implemented in the 'geneSetTest' function of the Bioconductor package *limma*). This test evaluates the statistical significance of the tendency for genes in the list to be highly ranked among the study's differentially expressed genes. Interestingly, the results indicated that the set of genes coming out of our analysis was systematically up-regulated on treatment with 5-aza-dC in all three separate studies ($P = 0.01$, 0.02 and 0.002, respectively). The combination of these results is highly significant ($P = 2.7 \times 10^{-4}$), incorporating the three sets of evidence using Fisher's method. This result is particularly important in that it illustrates that the set of genes identified through our application of NsRRR as having expression levels predicted by DNA methylation profiles in ovarian cancer, display consistent alteration in expression following alteration of the genome-wide state of DNA methylation in three different cancer cell lines. Therefore, it is highly likely that the expression of this set of genes is controlled by the state of DNA methylation, and the results obtained from our application of NsRRR reflect a real biological phenomenon.

Finally, we investigated aberrant expression of the genes corresponding to our response variables in a range of cancer types. Using the CancerMA database it was found that most of the genes with expression levels predicted by the selected DNA methylation probes were over- or under-expressed in some cancer types, with a number repeatedly displaying the same effect in many different cancer types. For example, the gene GATM was selected as a response variable, and this gene shows consistent down-regulation of expression in many cancer types, which is illustrated in Supplementary Material, Section C (Figure 2), with a Forest Plot showing the distributions of base-2 log fold-changes of measured expression in cancer samples compared with normal tissue, for a large number of independent studies spanning nine different cancer types. This information suggests that the genes we have identified as having expression levels predicted by methylation of a set of loci in a cancer dataset, and that are likely to show increased expression following treatment with a DNA methyltransferase inhibitor, may have functional roles in malignancy.

5 DISCUSSION

The components of biological systems can be organized into functional modules in which the constituent members work together to enact one process, such that alteration of a number of individual members, as may occur in disease, can result in a similar functional consequence (Zanzoni *et al.*, 2009). Using network topologies for incorporating prior knowledge about biological systems into large-scale data analysis can highlight important functional patterns hidden in the data. This more systems-level approach to multivariate data analysis, incorporating prior knowledge, offers an improvement in robustness and reliability over single-gene models (Efroni *et al.*, 2007) There are also advantages in terms of interpretation of the results, as it is easier to plan laboratory experiments targeting a specific function than it is a long list of individual genes. In our context, it is being able to identify modules in which the methylation status of functionally related genes could predict the expression level of another set of functionally related genes would be particularly useful, and might be indicative of a mechanistic link between the two functional units.

One example of an intuitive network would be a binary encoding of PPIs: that is, two genes are connected in the network if and only if their protein products interact (Li *et al.*, 2012). Such a network may be particularly suitable when the NsRRR model is being applied to datasets involving a subset of genes that is enriched for interaction, but for genome-wide analysis tasks the network may be too sparse to be sufficiently informative for the regression model. We propose a means of encoding functional relatedness of the variables in the form of networks of pathway annotation similarity.

We applied NsRRR to multivariate analysis of paired gene expression and DNA methylation datasets from primary ovarian tumours. We discovered a set of genes for which the expression levels are predicted by the levels of methylation at a set of CpG loci, mapping to a different set of genes. An encouraging result came in the form of confirmation in three independent gene expression profiling datasets that expression of the predicted genes was significantly up-regulated following treatment with the DNA methyltransferase inhibitor 5-aza-dC. Thus, pharmacological alteration of the state of DNA methylation was found to alter the state of expression of the predicted genes. A number of the genes

in question were found to show aberrant expression profiles in a number of cancers. So, although further clinical consequences of this observation are yet to be established, these results may shed light on the mechanism of action of epigenetic therapies that aim to undo the aberrant DNA methylation patterns acquired by cancer cells and give an indication of potential consequences of these treatments. This is further suggested by the observation that the set of genes with expression predicted by DNA methylation profiles was enriched for mTOR signalling, which has been shown to promote survival of platinum-resistant ovarian cancer cells. A further point to note is that the genomic loci for which methylation levels predicted expression of these genes were found to map to a different set of genes: therefore, this regulatory behaviour is unlikely to reflect a direct mechanism of regulating expression via DNA methylation, and it is perhaps more likely that there is a downstream functional link between these two sets of genes. That such a pattern was identified further justifies the decision to take a multivariate approach to the analysis of genome-wide levels of gene expression and DNA methylation.

In our application, we extracted only variables associated to the first rank of the NsRRR model. Further rank estimates could have been extracted following the procedure in Section 2.3. For the unpenalized RRR model, there are well-established approaches to guide the selection of the rank including graphical procedures such as the rank trace plot (Izenman, 2008). In the NsRRR model, an additional rank will generally be considered as long as the current selection probabilities are sufficiently high, and no further ranks would be explored as soon as the selection probabilities fall below a given threshold.

The statistical methods presented in this article build on our previous work on variable selection for linear regression models with high-dimensional responses within the framework of penalized least squares (Vounou *et al.*, 2010, 2012). An analogous Bayesian approach could also be developed, for instance, by extending the Bayesian RRR model proposed by Geweke (1996). The original model was not suitable for settings in which the sample size is much smaller than the number of variables, and the regression coefficients were assigned improper and flat priors, while the prior distribution of the covariance of random errors was assumed to be an inverse-Wishart. Variable selection could be obtained through double-exponential priors (Park and Casella, 2008), and prior information regarding the pairwise relationships between predictors and responses could be leveraged by means of Markov random fields priors (Stingo *et al.*, 2011). A Bayesian counterpart of the NsRRR would constitute future work.

Finally, it should be noted that the proposed model can be applied to a broader range of applied problems involving the search for association between two high-dimensional measurements observed on the same random sample, including the detection of expression quantitative trait loci.

Funding: GM acknowledges funding from Imperial College London Healthcare NHS Trust, Biological Research Council, award number DCIM_P31665; ZW is the recipient of the Beit Fellowship for Scientific Research scholarship. EC is supported by the National Institute for Health Research (NIHR) Imperial Biomedical Research Centre based at Imperial College Healthcare NHS Trust and Imperial College London.

Conflicts of Interest: none declared.

REFERENCES

Azencott,C.A. *et al.* (2013) Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, **29**, i171–i179.

Calvano,S. *et al.* (2005) A network-based analysis of systemic inflammation in humans. *Nature*, **437**, 1037–1032.

Chen,K. *et al.* (2012) Reduced-rank stochastic regression with a sparse singular value decomposition. *J. R. Stat. Soc, B*, **74**, 203–221.

Chuang,H.Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Chung,F. (1997) *Spectral Graph Theory. CBMS Regional Conference Series 92*. American Mathematical Society, Providence, RI, MR1421568.

Dannenberg,L. and Edenberg,H. (2006) Epigenetics of gene expression in human hepatoma cells: Expression profiling the response to inhibition of dna methylation and histone deacetylation. *BMC Genomics*, **7**, 181.

Dice,L. (1945) Measures of the amount of ecologic association between species. *Ecology*, **26**, 297–302.

Efroni,S. *et al.* (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS One*, **2**, e425.

Friedman,J. *et al.* (2007) Pathwise coordinate optimization. *Ann. Appl. Stat.*, **1**, 302–332.

Gervin,K. *et al.* (2012) Dna methylation and gene expression changes in monozygotic twins discordant for psoriasis: Identification of epigenetically dysregulated genes. *PLoS Genet.*, **8**, e1002454.

Geweke,J. (1996) Bayesian reduced rank regression in econometrics. *J. Econom.*, **75**, 121–146.

Irizarry,R. *et al.* (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Izenman,A. (2008) *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*. Springer, New York. ISBN 978-0-387-78188-4.

Joung,J. *et al.* (2013) Extracting coordinated patterns of dna methylation and gene expression in ovarian cancer. *J. Am. Med. Inform. Assoc.*, **20**, 637–642.

Kamburov,A. *et al.* (2011) Consensuspathdb: Toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.

Khamas,A. *et al.* (2012) Screening for epigenetically masked genes in colorectal cancer using 5-aza-2-deoxycytidine, microarray and gene expression profile. *Cancer Genomics Proteomics*, **9**, 67–75.

Lazarovici,A. *et al.* (2013) Probing dna shape and methylation state on a genomic scale with dnase i. *PNAS*, **110**, 6376–6381.

lé Cao,K. *et al.* (2008) A sparse pls for variable selection when integrating omics data. *Stat. Appl. Genet. Mol. Biol.*, **7**, 35.

Li,B.Q. *et al.* (2012) Identification of colorectal cancer related genes with mrmr and shortest path in protein-protein interaction network. *PLoS One*, **7**, e33393.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Meinshausen,N. and Bühlmann,P. (2010) Stability selection. *J. R. Stat. Soc. B*, **72**, 417–473.

Michaud,J. *et al.* (2008) Integrative analysis of runx1 downstream pathways and target genes. *BMC Genomics*, **9**, 363.

Minas,C. *et al.* (2013) A distance-based test of association between paired heterogeneous genomic data. *Bioinformatics*, **29**, 2555–2563.

Mueller,W. *et al.* (2007) Downregulation of runx3 and tes by hypermethylation in glioblastoma. *Oncogene*, **26**, 583–593.

Park,T. and Casella,G. (2008) The bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Peng,D.J. *et al.* (2010) Role of the akt/mtor survival pathway in cisplatin resistance in ovarian cancer cells. *Biochem. Biophys. Res. Commun.*, **394**, 600–605.

Reinsel,G. and Velu,R. (1988) *Multivariate Reduced-rank Regression: Theory and Applications*. New York: Springer.

Rhee,J.K. *et al.* (2013) Integrated analysis of genome-wide dna methylation and gene expression profiles in molecular subtypes of breast cancer. *Nucleic Acids Res.*, **41**, 8464–8474.

Smyth,G. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, 3.

Stingo,F. *et al.* (2011) Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes. *Ann. Appl. Stat.*, **5**, 1978–2002.

Stone,A. *et al.* (2013) Bcl-2 hypermethylation is a potential biomarker of sensitivity to antimitotic chemotherapy in endocrine-resistant breast cancer. *Mol. Cancer Ther.*, **12**, 1874–1885.

Suzuki,M. and Bird,A. (2008) Dna methylation landscapes: Provocative insights from epigenomics. *Nat. Rev. Genet.*, **9**, 465–476.

TCGA. (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.

Tenenhaus,A. *et al.* (2014) Variable selection for generalized canonical correlation analysis. *Biostatistics*, **15**, 569–583.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat., Soc.B*, **58**, 267–288.

Vaughan,S. *et al.* (2011) Rethinking ovarian cancer: Recommendations for improving outcomes. *Nat. Rev. Cancer*, **11**, 719–725.

Vounou,M. *et al.* (2010) Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage*, **53**, 1147–1159.

Vounou,M. *et al.* (2012) Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in alzheimer's disease. *Neuroimage*, **60**, 700–716.

Witten,D. *et al.* (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, **10**, 515–534.

Zanzoni,A. *et al.* (2009) A network medicine approach to human disease. *FEBS Lett.*, **583**, 1759–1765.

Zeller,C. *et al.* (2012) Candidate dna methylation drivers of acquired cisplatin resistance in ovarian cancer identified by methylome and expression profiling. *Oncogene*, **31**, 4567–4576.