Identification of supervised and sparse functional genomic pathways

Fan Zhang[1,*], Jeffrey C. Miecznikowski[1], and David L. Tritchler[1,2]

[1]*Department of Biostatistics, SUNY University at Buffalo, Buffalo NY 14214, USA*
[2]*Division of Biostatistics, University of Toronto, ON M5T 3M7, Toronto, Canada*
[*]*Corresponding author, fzhang8@buffalo.edu*

Abstract: Functional pathways involve a series of biological alterations that may result in the occurrence of many diseases including cancer. With the availability of various "omics" technologies it becomes feasible to integrate information from a hierarchy of biological layers to provide a more comprehensive understanding to the disease. In many diseases, it is believed that only a small number of networks, each relatively small in size, drive the disease. Our goal in this study is to develop methods to discover these functional networks across biological layers correlated with the phenotype. We derive a novel Network Summary Matrix (NSM) that highlights potential pathways conforming to least squares regression relationships. An algorithm called Decomposition of Network Summary Matrix via Instability (DNSMI) involving decomposition of NSM using instability regularization is proposed. Simulations and real data analysis from The Cancer Genome Atlas (TCGA) program will be shown to demonstrate the performance of the algorithm.

Keywords: supervised network, sparse, instability, pathway analysis

1 Introduction

In biology, identifying mediating processes or pathways is important for understanding the causation of an outcome of interest such as disease. An example is genome-wide association study (GWAS) where considering gene expression as an intermediate step to disease may explain the association of disease with a host's genotype. It has been observed that statistically significant single-nucleotide polymorphisms (SNPs) often have no immediate causal interpretation, and a possible explanation for this is their effect on the expression of (possibly distal) genes. To address this issue in GWAS, Gamazon et al. (2015) have exploited

recent genotype-tissue gene expression databases to determine linear combinations of SNPs that can be used to predict gene expression. The association of a predicted expression level for a gene with a trait then can provide mechanistic insights regarding SNPs by considering the SNPs that predict the expression of a gene found to associate with disease. In addition to the interpretability of SNP effects, this approach also enables a number of SNPs each with modest effect to combine and contribute to a larger gene-level effect.

A second example is the investigation of possible mechanisms by which microbial communities affect human host physiological status. Morgan et al. (2015) studied gene expression as a possible mechanism for the influence of the host microbiome on the presence of an irritable bowel syndrome (IBS) like affliction. One strategy they employed was dimension reduction. Transcripts were condensed into groups based on prior evidence of association with disease, or were organized into principle components using unsupervised methods. The microbial operational taxonomic units (OTUs) were also condensed into principle components. Then the analysis of associations among disease, microbiome, and transcription could be studied effectively.

The two examples mentioned demonstrate the use of synthetic composite variates to explain disease causation. Those approaches summarize the two processes measured independently. For example, Morgan et al. (2015) form transcript clusters, independently of the analysis used to form microbial OTU clusters.

Recently, methods have been developed for the identification of modules for high-dimensional data using sparse formulations of canonical correlation (Parkhomenko et al. (2009), Witten et al. (2009), Witten and Tibshirani (2009), Lê Cao et al. (2008), and Waaijenborg et al. (2008)). These methods provide sparse methods for canonical correlation analysis which infer genetic modules and transcript modules which maximally correlate with each other. This study extends those approaches by requiring that these modules are also associated with the outcome in a causal pathway. Also, this study is an expansion of the previous work done in Miecznikowski et al. (2016) where unsupervised methods are used to discover gene networks. Here we present an exploratory method which extends sparse canonical correlation to determine the two correlated variable sets such that they conform to a clear causal pathway for the outcome. This method can be applied to applications like the two mentioned above, genotype $\rightarrow$ disease and transcript $\rightarrow$ disease as causal pathways. In Section 2, the proposed method will be described as well as the latent model and the associated assumptions. Sections 3 and 4 will focus on simulation and application on a The Cancer Genome Atlas (TCGA) (Weinstein et al. (2013)) project. The implementation is described in Section 5. Some discussion and conclusions are given in Sections 6 and 7.

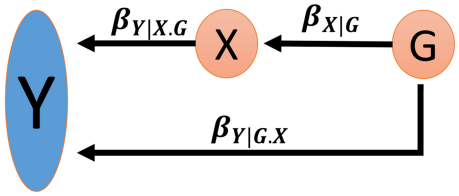


Figure 1: The graph representing the pathway through which the expression of gene X modulates the effect of a variable G on outcome Y. G may be SNP genotype or microbial composition, $\beta_{Y|X.G}$ is the regression coefficient for X in the model including the additional regressor G and $\beta_{Y|G.X}$ is the regression coefficient for G in that equation. Likewise, $\beta_{X|G}$ is the coefficient of G in the simple regression of X on G. Tracing the path $G \to X \to Y$ calculates this effect as $\beta_{X|G} \cdot \beta_{Y|X.G}$.

2 Method

2.1 Latent models and statistical assumptions for one pathway network

Figure 1 provides a graphical description of the pathway through which the expression of a variable X modulates the effect of a variable G on outcome Y. Figure 1 corresponds to the linear system

$$\begin{aligned} Y &= \beta_{Y|X.G}X + \beta_{Y|G.X}G + \epsilon_{Y|X,G}, \\ X &= \beta_{X|G}G + \epsilon_{X|G}, \\ G &= \epsilon_G, \end{aligned} \tag{1}$$

where $\beta_{Y|X.G}$ is the regression coefficient for X in the model including the additional regressor G and $\beta_{Y|G.X}$ is the regression coefficient for G in that equation. Likewise, $\beta_{X|G}$ is the coefficient of G in the simple regression of X on G. Variables $\epsilon_{Y|X.G}$, $\epsilon_{X|G}$ and ϵ_G are uncorrelated random error terms. We assume linearity, and tracing the path $G \to X \to Y$ calculates that effect as $\beta_{X|G}\beta_{Y|X.G}$. This is motivated by the familiar formula (Cochran (1938)) for least squares regression coefficients:

$$\beta_{Y|G} = \beta_{Y|G.X} + \beta_{X|G}\beta_{Y|X.G}. \tag{2}$$

The term $\beta_{Y|G.X}$ encompasses remaining effects of G on Y through the expression of other genes or other genetic effects.

The linear structure given by (1) defines the covariance of (Y, X, G). To define the distribution of the data implied by (1) we rewrite (1) as $\mathbf{Bz} = \mathbf{e}$ where $\mathbf{z} = (Y, X, G)^T$, $\mathbf{e} = (\epsilon_{Y|X,G}, \epsilon_{X|G}, \epsilon_G)^T$, and

$$\mathbf{B} = \begin{bmatrix} 1 & -\beta_{Y|X.G} & -\beta_{Y|G.X} \\ 0 & 1 & -\beta_{X|G} \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Then $\mathbf{B\Sigma B}^T = \mathbf{D}$ where $\mathbf{\Sigma}$ is the variance of $\mathbf{z}$ and $\mathbf{D}$ is the diagonal matrix with elements $(\sigma_{Y|X,G}, \sigma_{X|G}, \sigma_G)$ where the first two elements are the residual variances from the first two equations in (1) and $\sigma_G = Var(G)$ (Wermuth (1992)). Thus $\mathbf{\Sigma} = \mathbf{B}^{-1}\mathbf{D}\mathbf{B}^{-T}$ can be computed once $\mathbf{B}$ and $\mathbf{D}$ are defined. This yields

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{Y|X,G} + \sigma_{X|G}\beta_{Y|X.G}^2 + \sigma_G\tau^2 & \sigma_{X|G}\beta_{Y|X.G} + \sigma_G\beta_{X|G}\tau & \sigma_G\tau \\ \sigma_{X|G}\beta_{Y|X.G} + \sigma_G\beta_{X|G}\tau & \sigma_{X|G} + \sigma_G\beta_{X|G}^2 & \sigma_G\beta_{X|G} \\ \sigma_G\tau & \sigma_G\beta_{X|G} & \sigma_G \end{bmatrix} \tag{4}$$

where $\tau = \beta_{Y|G.X} + \beta_{X|G}\beta_{Y|X.G}$. Alternatively, $\mathbf{\Sigma}$ can also be obtained by the usual covariance computation if the inverse is difficult to compute.

For the outcome variable Y, we define α as the importance of the G, X path to be

$$\alpha = \frac{\beta_{X|G}\beta_{Y|X.G}}{\beta_{Y|G}}. \tag{5}$$

2.2 Algorithm

2.2.1 Network Summary Matrix (NSM)

We consider the triple Y, X, G to be a latent causal process. For one subject, sets of observed variables $\{g_i, i = 1, \ldots, a\}$, $\{x_j, j = 1, \ldots, b\}$, and y are indicators of a latent G, X, and Y respectively. The distributions of all $a + b$ indicators are independent conditional on the latent process given by Figure 1. Each realization of the (Y, X, G) process determines the distributions of the observed variables related to the causal process, so the distribution of the $\{g_i\}$ is centered at G, the distributions of the $\{x_j\}$ are centered at X, and y is sampled around Y. For N subjects, the sets $\{g_i\}$ and $\{x_j\}$ are assembled into matrices $\boldsymbol{\mathcal{G}}^{Sig}_{N\times a}$ and $\boldsymbol{\mathcal{X}}^{Sig}_{N\times b}$ respectively, where "Sig" denotes signal indicating these are the matrices containing the biological signal. I.e., the causal pathway. For $p \geq a$ we have the $N \times p$ matrix $\boldsymbol{\mathcal{G}} = (\boldsymbol{\mathcal{G}}^{Sig}, \boldsymbol{\mathcal{G}}^{Noi})$, where "Noi" means noise and $\boldsymbol{\mathcal{G}}^{Noi} = (\boldsymbol{g}_{a+1}, \ldots, \boldsymbol{g}_p)$ of which $\boldsymbol{g}_i$ is a length N vector of standardized sample genotypes for the ith SNP and $\boldsymbol{\mathcal{G}}^{Noi}$ are observations that are not in the causal process depicted in Figure 1. Similarly, for $q \geq b$, $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}^{Sig}, \boldsymbol{\mathcal{X}}^{Noi})$ is a $N \times q$ matrix with each column a sampled expression profile of the genes and $\boldsymbol{\mathcal{X}}^{Noi} = (\mathbf{x}_{b+1}, \ldots, \mathbf{x}_q)$ comes from non-causal processes, e.g. noise.

We construct the $p \times q$ matrix $\mathbf{C}$ where $c_{i,j} = \hat{\beta}_{y|x_j.g_i}\hat{\beta}_{x_j|g_i}$, and $\mathbf{D_G} = diag(\hat{\beta}_{y|g_i})$. The $\hat{\beta}$'s are obtained by ordinary least squares (OLS) method. Define the network summary matrix, $\mathbf{NSM} = [nsm_{i,j}] = \mathbf{D_G C}$ with dimension $p \times q$. Note that

$$\begin{aligned} nsm_{i,j} &= \hat{\beta}_{y|g_i}\hat{\beta}_{x_j|g_i}\hat{\beta}_{y|x_j.g_i} \\ &= \big[\frac{\hat{\beta}_{x_j|g_i}\hat{\beta}_{y|x_j.g_i}}{\hat{\beta}_{y|g_i}}\big]\hat{\beta}^2_{y|g_i} \\ &= \hat{\alpha}_{i,j}\hat{\beta}^2_{y|g_i} \end{aligned} \tag{6}$$

The expectation of the $\hat{\beta}$'s corresponding to given latent variables from a pathway is an attenuation toward zero of the original coefficients from the same latent process as the indicators can be viewed as the latent variables observed with measurement error. That is, the expectations have the same sign as in the underlying process, but are attenuated. Then the sign of $\left[\frac{\hat{\beta}_{y|x_j g_i}\hat{\beta}_{x_j|g_i}}{\hat{\beta}_{y|g_i}}\right]\hat{\beta}^2_{y|g_i}$ is the same as $\alpha\beta^2_{Y|G}$ in the underlying process. Note that $nsm_{i,j}$ is large when g_i and x_j are strongly related to y and to each other, signaling the importance of g_i and the existence of the mechanism depicted in Figure 1.

We make the model assumption that α is a proportion of the overall effect of G on Y when it exists, that is, $0 \leq \alpha \leq 1$ and this assumption stipulates two bounds:

1) $0 \leq \alpha = \frac{\beta_{X|G}\beta_{Y|X.G}}{\beta_{Y|G}}$: If this is violated the effect of G on Y mediated through X (expressed by $\beta_{X|G}\beta_{Y|X.G}$) is of different sign (hence different causal direction) than the overall G effect. Such contradictory effects make the interpretation of G complex and counterintuitive. This constraint is checked by verifying that $nsm_{i,j} \geq 0$ and **NSM** is filtered by setting elements in violation to zero.

2) $\alpha = \frac{\beta_{X|G}\beta_{Y|X.G}}{\beta_{Y|G}} \leq 1$: By 1) $\alpha = |\alpha| = \frac{|\beta_{X|G}\beta_{Y|X.G}|}{|\beta_{Y|G}|} \leq 1$. This is violated when $|\beta_{X|G}\beta_{Y|X.G}| \geq |\beta_{Y|G}|$. Since $\beta_{Y|G}$ is the total effect of G on Y it includes the effect mediated through X so there must be some components of the overall effect of G on Y that oppose the path mediated by X, complicating the causal narrative for G. Filtering by setting elements of **NSM** for which $nsm_{i,j} > \beta^2_{Y|g_i}$ to zero avoids considering these paths. Alternatively, if they are retained a more complex overall mechanism needs to be considered.

Note from here on, we assume **NSM** is the filtered version. The filtered **NSM** matrix can be considered as a consequence of the interaction network between $\boldsymbol{\mathcal{G}}$, $\boldsymbol{\mathcal{X}}$ and outcome $\mathbf{Y}$. The interpretation of $nsm_{i,j}$ is the portion of the squared change of y for every unit change of g_i which is contributed by going through x_j. With this interpretation, our goal is to select large $nsm_{i,j}$, which correspond to g, x pairs for which a substantial portion of the $g \to y$ effect is mediated by x. Formally, we form sets $\mathcal{A} = \{i; nsm_{i,j} \text{ is selected for some } j\}$ and $\mathcal{B} = \{j; nsm_{i,j} \text{ is selected for some } i\}$. We find those sets by decomposing **NSM** via instability regularized Penalized Matrix Decomposition (PMD). The goal of Section 2.2 is to find $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$, the estimators of $\mathcal{A}$ and $\mathcal{B}$, from the observed data.

2.2.2 Decomposition of Network Summary Matrix via Instability (DNSMI)

The **NSM** matrix defined in (6) has p rows and q columns. Because the order of the columns and the smoothness (Tibshirani et al. (2005)) are not among our primary interest of finding the elements of **NSM** that are relatively larger than

the others, we use the so called penalized matrix decomposition "PMD(L_1, L_1)" method where L_1 denotes the L_1-norm for a vector (Witten et al. (2009)). The L_1-norm of a vector $\mathbf{z}$ is denoted as $\|\mathbf{z}\|$ and defined as $(\sum_i |z_i|)$. Briefly, the PMD is a penalized version of singular value decomposition (SVD). For a given matrix $\mathbf{M}_{p\times q}$ of rank $K \leq \min(p, q)$, SVD decomposes $\mathbf{M}$ into

$$\mathbf{M} = \mathbf{U}\mathbf{D}\mathbf{V}^T,\ \mathbf{U}^T\mathbf{U} = \mathbf{I}_p,\ \mathbf{V}^T\mathbf{V} = \mathbf{I}_q,\ d_1 \geq d_2 \geq \cdots \geq d_K > 0. \tag{7}$$

The kth column of $\mathbf{U}$ is denoted as $\mathbf{u}_k$ and called the kth left singular vector, the kth column of $\mathbf{V}$ is denoted as $\mathbf{v}_k$ and called the kth right singular vector, and d_k denotes the kth diagonal element of the diagonal matrix $\mathbf{D}$ and called the kth singular value. Then for any $r \leq K$, the first r components of the SVD will give the best rank-r approximation to $\mathbf{M}$ in the sense of the Frobenius norm (Eckart and Young (1936)) below

$$\sum_{k=1}^{r} d_k \mathbf{u}_k \mathbf{v}_k^T = \arg \min_{\widehat{\mathbf{M}} \in N(r)} \|\mathbf{M} - \widehat{\mathbf{M}}\|_F^2 \tag{8}$$

where $N(r)$ is the set of rank-r $p \times q$ matrices. PMD imposes a penalty on size of $\mathbf{U}$ and $\mathbf{V}$ to particularly make some elements zero. If the PMD is applied on the rank-1 approximation of the original matrix $\mathbf{M}$ according to (9)

$$\begin{aligned} &\text{minimize}_{d,\mathbf{u},\mathbf{v}} \tfrac{1}{2}\|\mathbf{M} - d\mathbf{u}\mathbf{v}^T\|_F^2 \\ &\text{subject to } \|\mathbf{u}\|_2^2 = 1,\ \|\mathbf{v}\|_2^2 = 1,\ P_1(\mathbf{u}) \leq c_1,\ P_2(\mathbf{v}) \leq c_2,\ d \geq 0 \end{aligned} \tag{9}$$

where P_1 and P_2 are penalty functions, it will shrink some small elements of the left and right singular vectors, $\mathbf{u}$ and $\mathbf{v}$, to zero, and hence produce sparse solutions. More details of PMD and its applications are available in Witten et al. (2009) and Witten and Tibshirani (2009).

In our study, the network summary matrix (**NSM**) is the matrix to be decomposed by PMD, the P_1 and P_2 penalty functions are both L_1-norms and $\mathbf{u}$ and $\mathbf{v}$ represent the information from row and column dimensions of **NSM**. In order to apply the PMD(L_1, L_1) method, two tuning parameters c_1 and c_2 are required with the ranges $1 \leq c_1 \leq \sqrt{p}$ and $1 \leq c_2 \leq \sqrt{q}$ where c_1 and c_2 control the sparsity of $\mathbf{u}$ and $\mathbf{v}$, respectively. Smaller c_1 leads to sparser $\mathbf{u}$, the same for c_2 and $\mathbf{v}$. Tuning c_1 and c_2 and thus sparsity for $\mathbf{u}$ and $\mathbf{v}$ is achieved via the instability framework.

Instability is generally interpreted as a measure of disagreement of results across subsamples for a given method and the associated settings. It was originally developed for choosing regularization parameters for high dimensional graphical models, see Meinshausen and Bühlmann (2010) and Liu et al. (2010) for more details. However, the specific formation can be transformed for our research setting. In this proposal, the element-wise instability for vectors $\mathbf{u}$ and $\mathbf{v}$ given tuning parameters c_1 and c_2 is defined as

$$\xi_i^u(c_1, c_2) = 2\text{Pr}(u_i \text{ is selected})(1 - \text{Pr}(u_i \text{ is selected})) \tag{10}$$

$$\xi_j^v(c_1, c_2) = 2\text{Pr}(v_j \text{ is selected})(1 - \text{Pr}(v_j \text{ is selected})) \tag{11}$$

where $1 \leq i \leq p$ and $1 \leq j \leq q$. To estimate (10) and (11), we define

$$\widehat{\xi_i^u}(c_1, c_2) = 2\widehat{\Pr}(u_i \text{ is selected})(1 - \widehat{\Pr}(u_i \text{ is selected})) \quad (12)$$

$$\widehat{\xi_j^v}(c_1, c_2) = 2\widehat{\Pr}(v_j \text{ is selected})(1 - \widehat{\Pr}(v_j \text{ is selected})) \quad (13)$$

where

$$\widehat{\Pr}(u_i \text{ is selected}) = \frac{1}{R}\sum_{f=1}^{R} I_f^{u_i} \quad (14)$$

$$\widehat{\Pr}(v_j \text{ is selected}) = \frac{1}{R}\sum_{f=1}^{R} I_f^{v_j} \quad (15)$$

of which R is the total number of subsamples and

$$I_f^{u_i} = \begin{cases} 1 & \text{if } u_i \neq 0 \text{ for subsample } S_f,\ i = 1, \ldots, p,\ f = 1, \ldots, R \\ 0 & \text{if } u_i = 0 \text{ for subsample } S_f,\ i = 1, \ldots, p,\ f = 1, \ldots, R \end{cases} \quad (16)$$

likewise for $I_f^{v_j}$. The instability for $\mathbf{u}$ under the same setting, denoted as $\xi_u(c_1, c_2)$, is the mean instability averaged over all the elements

$$\xi_u(c_1, c_2) = \frac{1}{p}\sum_{i=1}^{p} \xi_i^u(c_1, c_2) \quad (17)$$

and it is estimated by $\widehat{\xi_u}(c_1, c_2)$ via $\widehat{\xi_i^u}(c_1, c_2)$ in (12). Same as for the instability of $\mathbf{v}$, denoted as $\xi_v(c_1, c_2)$ and estimated by $\widehat{\xi_v}(c_1, c_2)$ via $\widehat{\xi_j^v}(c_1, c_2)$. In order to have one scalar that can represent the combined instability derived from the pair (c_1, c_2), the maximum between $\xi_u(c_1, c_2)$ and $\xi_v(c_1, c_2)$ is used and denoted as $\xi_{u,v}(c_1, c_2)$. Because $\xi_{u,v}(c_1, c_2)$ is not necessarily a monotone function of either c_1 or c_2 when the other parameter is fixed, it is then monotonized by substituting the supremum instability up to (c_1, c_2) for $\xi_{u,v}(c_1, c_2)$ given a pair of (c_1, c_2). The supremum instability at (c_1, c_2) is defined as

$$\overline{\xi}(c_1, c_2) = \sup_{1 \leq s \leq c_1, 1 \leq t \leq c_2} \xi_{u,v}(s, t). \quad (18)$$

The optimal pair of (c_1, c_2) is obtained by grid search and such a grid has h elements in c_1 direction and l elements in c_2 direction. At the end, we choose a pair of (c_1, c_2) such that

$$(c_1, c_2) = \arg \max_{(c_1, c_2) \in E} \left\| \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \right\|_2 \quad (19)$$

where $E = \{(c_1, c_2) \mid \overline{\xi}(c_1, c_2) \leq \delta \text{ over the } h \times l \text{ search grid}\}$ for a preset threshold δ which ranges from 0 to 0.5. The detailed algorithm for DNSMI used on given observation matrices $\mathbf{Y}_{N\times 1}$, $\boldsymbol{\mathcal{X}}_{N\times q}$ and $\boldsymbol{\mathcal{G}}_{N\times p}$ to find the pathway elements $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$ for a given search grid is presented as Algorithm 1 in Appendix 1.

3 Simulations

3.1 Extended latent model

To create a realistic covariance matrix additional variables are added to the latent pathway model in Figure 1 to make an extended model (20):

$$\begin{aligned} Y &= \beta_{Y|X.G}X + \beta_{Y|G.X}G + \beta_{Y|S}S + \beta_{Y|H}H + \epsilon_{Y|X,G,S,H}, \\ X &= \beta_{X|G}G + \epsilon_{X|G}, \\ G &= \epsilon_G. \end{aligned} \tag{20}$$

We include a variable S with the same variance as X which is independent of X and G but adds $\beta_{Y|S}S + \epsilon_{Y|S}$ to the system equation for Y in (1). The indicators of S will be a cluster of variables (typically transcripts) related to Y. Similarly, we include a variable H with the same variance as G which is independent of X, G and S but adds $\beta_{Y|H}H + \epsilon_{Y|H}$ to the system equation for Y in (1). The indicators of H will be a cluster of variables (typically genes or microbial species abundance) related to Y. This results in the covariance matrix $\mathbf{\Sigma}'$ for $(Y, X, G, S, H)^T$ as

$$\mathbf{\Sigma}' = \begin{bmatrix} \sigma_{Y|X.G.S.H} + \sigma_{X|G}\beta^2_{Y|X.G} + \sigma_G\tau^2 + \sigma_S\beta^2_{Y|S} + \sigma_H\beta^2_{Y|H} & \sigma_{X|G}\beta_{Y|X.G} + \tau\sigma_G\beta_{X|G} & \sigma_G\tau & \beta_{Y|S}\sigma_S & \beta_{Y|H}\sigma_H \\ \sigma_{X|G}\beta_{Y|X.G} + \tau\sigma_G\beta_{X|G} & \beta^2_{X|G}\sigma_G + \sigma_{X|G} & \sigma_G\beta_{X|G} & 0 & 0 \\ \sigma_G\tau & \sigma_G\beta_{X|G} & \sigma_G & 0 & 0 \\ \beta_{Y|S}\sigma_S & 0 & 0 & \sigma_S & 0 \\ \beta_{Y|H}\sigma_H & 0 & 0 & 0 & \sigma_H \end{bmatrix}, \tag{21}$$

where arbitrarily $\sigma_S = \beta^2_{X|G}\sigma_G + \sigma_{X|G}$ so that S has the same variance as X.

We add another null case. X', G' represent a pathway with the same covariance structure as $(X, G)^T$, but are independent of Y. The covariance of $(X', G')^T$ will be the submatrix formed from the second and third rows and columns of $\mathbf{\Sigma}'$ in (21), call it $\mathbf{\Sigma^{ind}}$. Then the covariance for $(Y, X, G, S, H, X', G')^T$ is

$$\mathbf{\Sigma}'' = \begin{bmatrix} \mathbf{\Sigma}' & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma^{ind}} \end{bmatrix} \tag{22}$$

The extended model is illustrated in Figure 2.

3.2 Other algorithms

In order to evaluate the performance of the proposed algorithm DNSMI, we compare it to three different algorithms: AMSE-PMD, sSCCA-P and sSCCA-W where AMSE-PMD stands for Average Mean Squared Error tuned PMD decomposition of **NSM**, sSCCA refers to supervised Sparse Canonical Correlation Analysis and "P" and "W" means that the SCCA takes the form in Parkhomenko et al. (2009) or Witten and Tibshirani (2009), respectively. AMSE is a regularization method suggested in Witten et al. (2009) used to choose parameters for PMD. The supervision is carried out by implementing univariate simple regression on each of the columns of $\boldsymbol{\mathcal{X}}$ with $\mathbf{Y}$ and then to select features with

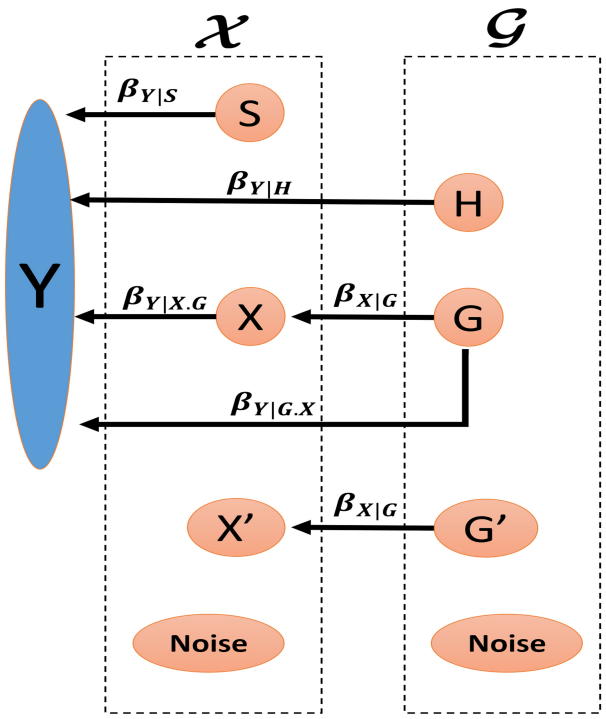


Figure 2: One pathway extended latent model. The left dashed box includes latent variables from whom the observations will be generated and assembled into matrix $\boldsymbol{\mathcal{X}}$ and the right dashed box includes latent variables from whom the observations will be generated and assembled into matrix $\boldsymbol{\mathcal{G}}$.

the p-values controlled for a level of false discovery rate (FDR) using Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg (1995)), for example, at 0.2. This filtering will result in a new matrix $\widetilde{\boldsymbol{\mathcal{X}}}$ consisting of the selected features under the preset FDR rate. Analogously, the selected features from $\boldsymbol{\mathcal{G}}$ will be stored in matrix $\widetilde{\boldsymbol{\mathcal{G}}}$. After that, both of the SCCA algorithms will be applied on the resulted $\widetilde{\boldsymbol{\mathcal{G}}}$ and $\widetilde{\boldsymbol{\mathcal{X}}}$ matrices. The details of AMSE-PMD, sSCCA-P and sSCCA-W are completely specified in Algorithm 2, 3 and 4 in Appendix 1.

3.3 Simulation settings and realization

We simulate a model governed by a system of latent variables conforming to (20). A primary parameter of our simulation is α given by (5), the proportion of the G effect mediated by X. Without loss of generality we set the scale by letting $\beta_{Y|G} = 1$ so $\alpha = \beta_{X|G}\beta_{Y|X.G}$. The relation (2) implies that $\tau = 1$ and $\beta_{Y|G.X} = 1 - \alpha$. We further assume that $\beta_{X|G} = \beta_{Y|X.G} = \sqrt{\alpha}$ so that the links in the $G \to X \to Y$ path are of equal strength. We also set the variance of G to be 1 and calculate the variance parameters $\sigma_{X|G}$ and $\sigma_{Y|X.G.S.H}$ to specify the predictability of Y and X in (20) to be a specified $R^2 = R^2(X) = R^2(Y)$. The $R^2(X)$ is defined as

$$R^2(X) = \frac{Var(X) - \sigma_{X|G}}{Var(X)}, \tag{23}$$

likewise for $R^2(Y)$. Using the relation $R^2(X) = \alpha\sigma_G/(\sigma_{X|G} + \alpha\sigma_G) = \alpha/(\sigma_{X|G} + \alpha)$ yields $\sigma_{X|G} = \alpha\gamma$, with $\gamma = (1-R^2)/R^2$. Similarly, $\sigma_{Y|X.G.S.H} = \gamma + \alpha^2\gamma(1+\gamma)(1+\alpha)$. In addition, we set $\beta_{Y|S} = \beta_{Y|H} = \alpha$ so that each effect is the same as that of the $G \to X \to Y$ pathway.

The latent variable X interpreted as gene expression will act as a driver for a pathway of transcripts which we will generate as a set of variables with a normal distribution $x_j \sim N(X, \sigma_x^2), j = 1, \ldots, b$ which are independent given X. Likewise $g_j \sim N(G, \sigma_g^2), j = 1, \ldots, a$ independently. The correlation of x_j with the latent variable X driving that module can be set to r by setting $\sigma_x^2 = Var(X)(1 - r^2)/r^2$. Likewise the correlation of g_j with the latent variable G driving that module is set to r by setting $\sigma_g^2 = Var(G)(1 - r^2)/r^2$ and y is sampled from $N(Y, Var(Y)(1 - r^2)/r^2)$. In addition, we do the same for observations generated from latent variables S, X', H, G'. The noise variables are generated according to $N(0, \pi^2 Var(Y)(1 - r^2)/r^2)$ where π^2 is a parameter controlling the variance ratio between noise and y.

We thus have the parameters α (the importance of X as a mediator of Y, G association), R^2 (the predictability of the latent variables Y and X), r (specifying the correlations of the observed variables with their underlying latent variable), and π^2 (variance ratio between noise and y), which can be manipulated to model various scenarios through simulation of the defined distributions.

We are particularly interested in the performance of different methods as a function of 3 factors: signal to nonsignal ratio, sample size and sparsity. The signal here consists of elements indexed by the sets $\mathcal{A}$ and $\mathcal{B}$ introduced in Section 2.2.1, and the nonsignals indicate $\{i; nsm_{i,j}\} \setminus \mathcal{A}$ and $\{j; nsm_{i,j}\} \setminus \mathcal{B}$. For each

Table 1: Scenario setting for various signal to nonsignal ratios.

Parameter	Signal to nonsignal ratio (Scenario index)		
	Low (A)	Middle (B)	High (C)
α	0.35	0.35	0.35
R^2	0.35	0.85	0.85
r	0.35	0.35	0.85
π	0.5	1	2
N	500	500	500
a	10	10	10
Number of g'	63	63	63
Number of h	63	63	63
Number of noise	64	64	64
b	15	15	15
Number of x'	95	95	95
Number of s	95	95	95
Number of noise	95	95	95
Number of columns of $\boldsymbol{\mathcal{G}}$	200	200	200
Number of columns of $\boldsymbol{\mathcal{X}}$	300	300	300

factor, three different scenarios are set with each scenario having a different level of that factor while keeping the other factors the same. Scenarios A, B and C are used to test the signal to nonsignal ratio factor, Scenarios B, D and E are used for the sample size factor and Scenarios B, F and G are for the sparsity factor. The outcomes are evaluated via three measures, true positive rate (TPR) or sensitivity, true negative rate (TNR) or specificity and Cohen's Kappa statistic, κ (Cohen (1960)). Cohen's Kappa is a robust measurement of agreement that takes into account the agreement that occurs by chance. κ ranges from -1 to 1 with 1 representing complete agreement, 0 indicating no agreement and -1 being complete disagreement. For information of its interpretation and relationship between TPR and TNR, see McHugh (2012) and Feuerman and Miller (2008). Because the outcomes are specific to the rows ($\mathbf{u}$) and columns ($\mathbf{v}$), all three summaries (TPR, TNR, κ) will be evaluated on each of $\mathbf{u}$, $\mathbf{v}$ and the sum of κ on $\mathbf{u}$ and $\mathbf{v}$ will be also included as the total accuracy, i.e. $\kappa_{Tot} = \kappa_{\mathbf{u}} + \kappa_{\mathbf{v}}$ where $\kappa_{\mathbf{u}}$ and $\kappa_{\mathbf{v}}$ are the Kappa statistic for $\mathbf{u}$ and $\mathbf{v}$, respectively.

We perform 100 Monte Carlo simulations for each of the methods under each of the scenarios. We chose $\delta = 0.05$ and FDR $= 0.2$ for methods DNSMI, sSCCA-P, and sSCCA-W.

3.4 Factor 1: signal to nonsignal ratio

Three different scenarios, Scenario A, B and C, ordered by signal to nonsignal ratio from low, middle to high are set to evaluate the methods in Section 2.2.2 and 3.2. Their settings are listed in Table 1. The **NSM** matrix is plotted in Figure 3. Figure 3 shows that in the Low signal to nonsignal ratio scenario the

Table 2: Summary of TPR and TNR on dimensions **u** and **v** for 100 simulations for low, middle and high signal to nonsignal ratio scenarios. For DNSMI, $\delta = 0.05$. For sSCCA-P and sSCCA-W, FDR = 0.2.

Method	**Dimension**								**Signal to nonsignal ratio (Scenario index)**
	u				**v**				
	TPR		TNR		TPR		TNR		
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
DNSMI	0.12	0.01	0.99	0.00	0.15	0.02	0.98	0.00	
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	Low (A)
sSCCA-P	0.00	0.00	0.03	0.02	0.00	0.00	0.03	0.02	
sSCCA-W	0.00	0.00	0.04	0.02	0.00	0.00	0.04	0.02	
DNSMI	0.40	0.02	1.00	0.00	0.42	0.02	1.00	0.00	
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	Middle (B)
sSCCA-P	0.10	0.02	0.28	0.04	0.07	0.02	0.28	0.04	
sSCCA-W	0.11	0.02	0.31	0.05	0.08	0.02	0.31	0.05	
DNSMI	0.98	0.01	1.00	0.00	1.00	0.00	1.00	0.00	
AMSE-PMD	1.00	0.00	0.01	0.01	1.00	0.00	0.01	0.01	High (C)
sSCCA-P	0.97	0.02	0.99	0.00	0.97	0.02	0.99	0.00	
sSCCA-W	0.95	0.02	0.83	0.02	0.84	0.02	0.92	0.01	

signals are almost undistinguishable from the nonsignals. In the Middle scenario the signals become stronger and the signals are completely separated from the nonsignals in High scenario.

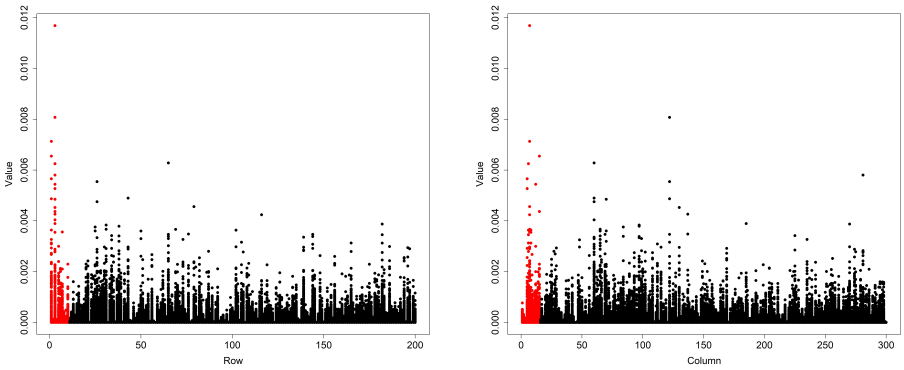
The TPR and TNR are summarized in Table 2. From the outcome, we see that AMSE-PMD selects almost all elements over all three scenarios regardless of the signal to nonsignal ratio. Excluding it, we observe a general trend on both dimensions that TPR and TNR will improve as the signal to nonsignal ratio increases and reach the maximum under the High scenario. Our proposed method DNSMI maintains a very high level (> 0.98) of TNR across all three scenarios and the TPR increases from 0.12 to 0.98 and from 0.15 to 1 on dimensions **u** and **v**, respectively.

The Cohen's Kappa is summarized in Table 3. As previously indicated by TPR and TNR, κ confirms that AMSE-PMD has little detection accuracy while DNSMI at $\delta = 0.05$ level has better performance than the others.

To conclude, signal to nonsignal ratio plays a key role in every method except AMSE-PMD. Larger signal to nonsignal ratio will lead to larger κ.

3.5 Factor 2: sample size N

We use the middle signal to nonsignal ratio scenario (Scenario B) from Section 3.4 to represent the low sample size scenario plus two new scenarios Scenario D and E, ordered by sample size, to evaluate the four methods. Their settings are



(a) Row in Low scenario

(b) Column in Low scenario

(c) Row in Middle scenario

(d) Column in Middle scenario

(e) Row in High scenario

(f) Column in High scenario

Figure 3: **NSM** for one realization for different signal to nonsignal ratio scenarios as a function of row and column. Red points indicate pathway elements, i.e. $a = 10$ elements on row dimension and $b = 15$ elements on column dimension. Low, Middle and High corresponds to scenario A, B and C in Table 1

Table 3: Summary of Cohen's Kappa, κ, on dimensions $\mathbf{u}$ and $\mathbf{v}$ for 100 simulations for low, middle and high signal to nonsignal ratio scenarios. For DNSMI, $\delta = 0.05$. For sSCCA-P and sSCCA-W, FDR = 0.2. Total accuracy, $\kappa_{Tot} = \kappa_{\mathbf{u}} + \kappa_{\mathbf{v}}$.

Method	**Dimension**						**Signal to nonsignal ratio (Scenario index)**
	κ on $\mathbf{u}$		κ on $\mathbf{v}$		Total accuracy		
	Mean	SE	Mean	SE	Mean	SE	
DNSMI	0.15	0.02	0.17	0.02	0.32	0.04	Low (A)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.00	0.00	0.00	0.00	0.01	0.01	
sSCCA-W	0.00	0.00	0.00	0.00	0.01	0.00	
DNSMI	0.53	0.02	0.55	0.02	1.08	0.04	Middle (B)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.11	0.02	0.10	0.02	0.21	0.04	
sSCCA-W	0.13	0.02	0.11	0.02	0.24	0.04	
DNSMI	0.98	0.01	1.00	0.00	1.98	0.01	High (C)
AMSE-PMD	0.00	0.00	0.01	0.01	0.01	0.01	
sSCCA-P	0.91	0.02	0.93	0.02	1.85	0.04	
sSCCA-W	0.51	0.04	0.66	0.03	1.17	0.06	

listed in Table 4 with Low, Middle and High scenarios having 500, 1000, and 1500 sample size. The **NSM** matrices are plotted in Figure 4. It appears that the signals get stronger as the sample size increases.

The TPR and TNR are summarized in Table 5. Similar to Section 3.4, AMSE-PMD selects all elements regardless of the change on sample size. However, the increase in sample size has a large improvement of TPR for method DNSMI while having almost no effect on TNR. For example, on $\mathbf{u}$ dimension, DNSMI has a mean TPR of 0.40, 0.66 and 0.84 for Low, Middle and High scenarios, respectively, and a mean TNR of 1. On the other hand, the increase in sample size has a larger positive effect on both measures on the supervised SCCA. For instance, sSCCA-P has mean TPR's of 0.1, 0.42 and 0.54 and mean TNR's of 0.28, 0.66 and 0.68 on $\mathbf{u}$ dimension under the Low, Middle and High scenarios, respectively.

The results of Cohen's Kappa are summarized in Table 6. Similarly, κ of DNSMI improves greatly as sample size increases. AMSE-PMD is unaffected by sample size. sSCCA-P and sSCCA-W will also have larger κ when sample size increases, however, with larger variance as well.

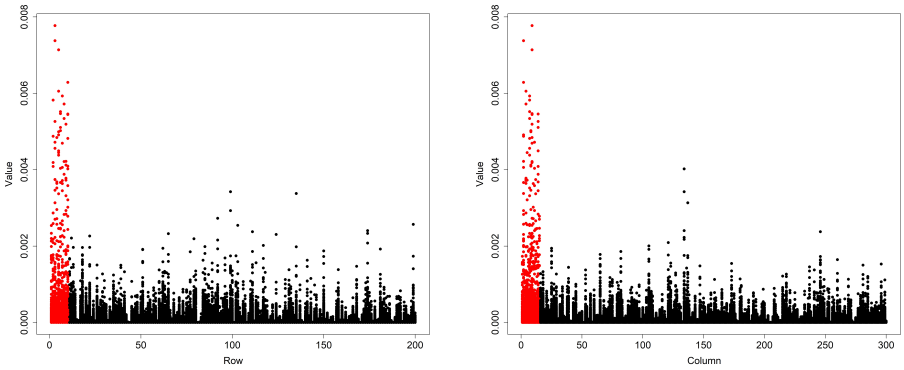
In conclusion, increasing the sample size N will improve the performance of DNSMI, sSCCA-P and sSCCA-W. Particularly, κ_{Tot} for DNSMI greatly increases as a function of sample size.

Table 4: Scenario setting for various sample size N. Low, Middle and High represent 500, 1000 and 1500 sample sizes.

Parameter	Sample Size (Scenario index)		
	Low (B)	Middle (D)	High (E)
α	0.35	0.35	0.35
R^2	0.85	0.85	0.85
r	0.35	0.35	0.35
π	1	1	1
N	500	1000	1500
a	10	10	10
Number of g'	63	63	63
Number of h	63	63	63
Number of noise	64	64	64
b	15	15	15
Number of x'	95	95	95
Number of s	95	95	95
Number of noise	95	95	95
Number of columns of $\boldsymbol{\mathcal{G}}$	200	200	200
Number of columns of $\boldsymbol{\mathcal{X}}$	300	300	300

Table 5: Summary of TPR and TNR on dimensions **u** and **v** for 100 simulations for sample size $N = 500, 1000$ and 1500 scenarios. For DNSMI, $\delta = 0.05$. For sSCCA-P and sSCCA-W, FDR = 0.2.

Method	**Dimension**								**Sample size (Scenario index)**
	u				**v**				
	TPR		TNR		TPR		TNR		
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
DNSMI	0.40	0.02	1.00	0.00	0.42	0.02	1.00	0.00	Low (B)
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	
sSCCA-P	0.10	0.02	0.28	0.04	0.07	0.02	0.28	0.04	
sSCCA-W	0.11	0.02	0.31	0.05	0.08	0.02	0.31	0.05	
DNSMI	0.66	0.02	1.00	0.00	0.64	0.02	1.00	0.00	Middle (D)
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.01	0.01	
sSCCA-P	0.42	0.04	0.66	0.05	0.39	0.03	0.67	0.05	
sSCCA-W	0.42	0.04	0.65	0.05	0.37	0.03	0.67	0.05	
DNSMI	0.84	0.02	1.00	0.00	0.82	0.02	1.00	0.00	High (E)
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.01	0.00	
sSCCA-P	0.54	0.04	0.68	0.05	0.49	0.04	0.69	0.05	
sSCCA-W	0.59	0.04	0.73	0.04	0.51	0.03	0.76	0.04	



(a) Row in Low scenario

(b) Column in Low scenario

(c) Row in Middle scenario

(d) Column in Middle scenario

(e) Row in High scenario

(f) Column in High scenario

Figure 4: **NSM** for one realization for different sample size N scenarios as a function of row and column. Red points indicate pathway elements, i.e. $a = 10$ elements on row dimension and $b = 15$ elements on column dimension. Low, Middle and High correspond to scenario B, D and E in Table 4 and represent N = 500, 1000 and 1500.

Table 6: Summary of Cohen's Kappa, κ, on dimensions $\mathbf{u}$ and $\mathbf{v}$ for 100 simulations for sample size $N = 500, 1000$ and 1500 scenarios. For DNSMI, $\delta = 0.05$. For sSCCA-P and sSCCA-W, FDR = 0.2. Total accuracy, $\kappa_{Tot} = \kappa_{\mathbf{u}} + \kappa_{\mathbf{v}}$.

Method	**Dimension**						**Sample size (Scenario index)**
	κ on $\mathbf{u}$		κ on $\mathbf{v}$		Total accuracy		
	Mean	SE	Mean	SE	Mean	SE	
DNSMI	0.53	0.02	0.55	0.02	1.08	0.04	Low (B)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.11	0.02	0.10	0.02	0.21	0.04	
sSCCA-W	0.13	0.02	0.11	0.02	0.24	0.04	
DNSMI	0.77	0.02	0.76	0.01	1.53	0.03	Middle (D)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.44	0.04	0.44	0.04	0.88	0.07	
sSCCA-W	0.41	0.04	0.43	0.04	0.84	0.07	
DNSMI	0.89	0.02	0.88	0.01	1.78	0.03	High (E)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.53	0.04	0.52	0.04	1.05	0.08	
sSCCA-W	0.49	0.03	0.57	0.04	1.06	0.07	

3.6 Factor 3: sparsity, a/p and b/q

Two additional scenarios are introduced to test the performance for each method under different sparsity settings. They are built by varying the sparsities calculated via a/p and b/q such that they have sparsities 0.05, 0.15 and 0.25 for both $\mathbf{u}$ and $\mathbf{v}$ dimensions. Their settings are listed in Table 7 and the **NSM** matrices are plotted in Figure 5.

Table 8 contains the numerical results of TPR and TNR for each method under each scenario. From the results we see that the most significant change is the decreasing trend of mean TPR as the sparsity decreases for method DNSMI. On $\mathbf{u}$ dimension, TPR decreases from 0.4 to 0.18 and to 0.11 while it drops from 0.42 to 0.20 and to 0.12 on $\mathbf{v}$ dimension. However, the TNR is 1.00 across all three scenarios on both dimensions.

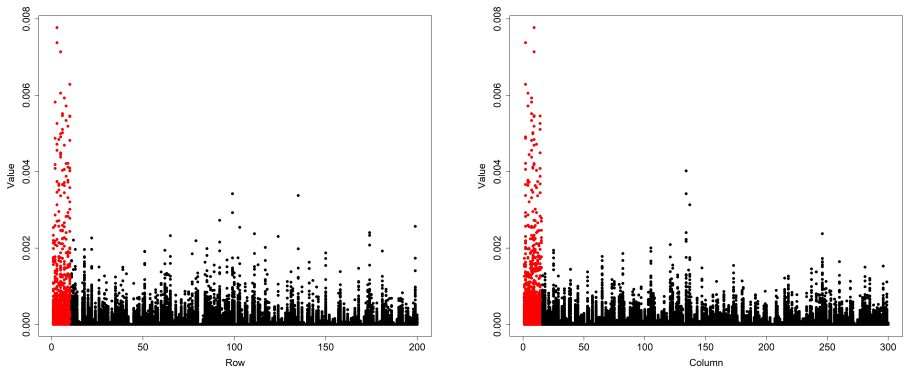
From the results of Cohen's Kappa, Table 9, we also see that the mean total accuracy drops from 1.08 to 0.57 and to 0.32 as the sparsity changes downwards. In conclusion, sparsity level is a crucial factor affecting the performance of DNSMI where the more sparse the better to apply DNSMI. Importantly, we note that DNSMI performs very well in very sparse settings.

Table 7: Scenario setting for various sparsities, a/p and b/q. High, Middle and Low represent $a/p = b/q = 0.05$, 0.15 and 0.25.

Parameter	Sparsity (Scenario Index)		
	High (B)	Middle (F)	Low (G)
α	0.35	0.35	0.35
R^2	0.85	0.85	0.85
r	0.35	0.35	0.35
π	1	1	1
N	500	500	500
a	10	30	50
Number of g'	63	56	49
Number of h	63	56	49
Number of noise	64	58	52
b	15	45	75
Number of x'	95	85	75
Number of s	95	85	75
Number of noise	95	85	75
Number of columns of $\boldsymbol{\mathcal{G}}$	200	200	200
Number of columns of $\boldsymbol{\mathcal{X}}$	300	300	300

Table 8: Summary of TPR and TNR on dimensions **u** and **v** for 100 simulations for sparsities $a/p = b/q = 0.05$, 0.15 and 0.25 scenarios. For DNSMI, $\delta = 0.05$. For sSCCA-P and sSCCA-W, FDR = 0.2.

Method	**Dimension**								**Sparsity (Scenario index)**
	u				**v**				
	TPR		TNR		TPR		TNR		
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	
DNSMI	0.40	0.02	1.00	0.00	0.42	0.02	1.00	0.00	High (B)
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	
sSCCA-P	0.10	0.02	0.28	0.04	0.07	0.02	0.28	0.04	
sSCCA-W	0.11	0.02	0.31	0.05	0.08	0.02	0.31	0.05	
DNSMI	0.18	0.01	1.00	0.00	0.20	0.00	1.00	0.00	Middle (F)
AMSE-PMD	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	
sSCCA-P	0.22	0.03	0.50	0.05	0.19	0.02	0.50	0.05	
sSCCA-W	0.19	0.02	0.50	0.05	0.16	0.02	0.50	0.05	
DNSMI	0.11	0.00	1.00	0.00	0.12	0.00	1.00	0.00	Low (G)
AMSE-PMD	1.00	0.00	0.05	0.01	0.99	0.00	0.13	0.01	
sSCCA-P	0.28	0.03	0.58	0.05	0.25	0.03	0.59	0.05	
sSCCA-W	0.24	0.02	0.63	0.05	0.21	0.02	0.63	0.05	



(a) Row in High scenario

(b) Column in High scenario

(c) Row in Middle scenario

(d) Column in Middle scenario

(e) Row in Low scenario

(f) Column in Low scenario

Figure 5: **NSM** for one realization for different sparsities, a/p and b/q, scenarios as a function of row and column. Red points indicate pathway elements, i.e. $a = 10, 30$ and 50 elements on row dimension and $b = 15, 45$ and 75 elements on column dimension. Low, Middle and High indicate scenario G, F and B in Table 7 and represent sparsity 0.25, 0.15 and 0.05.

Table 9: Summary of Cohen's Kappa, κ, on dimensions $\mathbf{u}$ and $\mathbf{v}$ for 100 simulations for sparsities $a/p = b/q = 0.05$, 0.15 and 0.25 scenarios. For DNSMI, $\delta = 0.05$. For sSCCA-P and sSCCA-W, FDR = 0.2. Total accuracy, $\kappa_{Tot} = \kappa_{\mathbf{u}} + \kappa_{\mathbf{v}}$.

Method	**Dimension**						**Sparsity (Scenario index)**
	κ on $\mathbf{u}$		κ on $\mathbf{v}$		Total accuracy		
	Mean	SE	Mean	SE	Mean	SE	
DNSMI	0.53	0.02	0.55	0.02	1.08	0.04	High (B)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.11	0.02	0.10	0.02	0.21	0.04	
sSCCA-W	0.13	0.02	0.11	0.02	0.24	0.04	
DNSMI	0.27	0.01	0.29	0.01	0.57	0.01	Middle (F)
AMSE-PMD	0.00	0.00	0.00	0.00	0.00	0.00	
sSCCA-P	0.25	0.03	0.22	0.03	0.47	0.05	
sSCCA-W	0.25	0.03	0.21	0.03	0.46	0.05	
DNSMI	0.15	0.00	0.17	0.00	0.32	0.01	Low (G)
AMSE-PMD	0.03	0.01	0.07	0.01	0.10	0.01	
sSCCA-P	0.30	0.03	0.27	0.03	0.58	0.06	
sSCCA-W	0.29	0.03	0.26	0.03	0.55	0.05	

4 Data analysis

4.1 Background

In this section, an example application of DNSMI using data from The Cancer Genome Atlas (TCGA) (Weinstein et al. (2013)) is presented. TCGA is a project supervised by the National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute. Using genome sequencing and bioinformatics as well as applying high-throughput genome analysis techniques, TCGA aims to improve the ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease. TCGA has now expanded to cover 60 primary cancer sites and include 40 different research projects. We choose the Uterine Corpus Endometrial Carcinoma (UCEC) project as our example. An outline of the TCGA-UCEC project is listed in Table 10.

Endometrial cancer is a cancer that originates from the endometrium. In 2012, there were 320,000 new occurrences and 76,000 deaths, which makes it the third lethal cancer in cancers for women following ovarian and cervical cancer (McGuire (2016)). The overall five-year survival rate in the United States is greater than 80% if the disease is diagnosed at an early stage (Sheets (2015)). Endometrial cancers may be tumours derived from epithelial cells (carcinomas), mixed epithelial and mesenchymal tumours (carcinosarcomas), or mesenchymal tumours. There is a strong correlation between the histologic tumor grade, the depth of myometrial invasion and the prevalence of lymph node metastasis and

Table 10: Basic information about TCGA-UCEC project.

Data Category	Data Type	Workflow Type	Cases
Raw Sequencing Data	Aligned Reads	BWA with Mark Duplicates and Cocleaning STAR 2-Pass BWA-aln	559
Transcriptome Profiling	Gene Expression Quantification Isoform Expression Quantification miRNA Expression Quantification	BCGSC miRNA Profiling HTSeq - Counts HTSeq - FPKM HTSeq - FPKM-UQ	559
Simple Nucleotide Variation	Annotated Somatic Mutation Raw Simple Somatic Mutation Aggregated Somatic Mutation Masked Somatic Mutation	12 types see TCGA website for details.	542
Copy Number Variation	Copy Number Segment Masked Copy Number Segment	DNAcopy	547
DNA Methylation	Methylation Beta Value	Liftover	559
Clinical	Clinical Supplement	NA	548
Biospecimen	Biospecimen Supplement	NA	560

the patient survival (Boronow et al. (1984)). The myometrial invasion ratio determines the International Federation of Gynecology and Obstetrics stage and has a direct influence on treatment (Lin et al. (2009)).

From past studies, TCGA researchers have characterized the marked differences between the two types of endometrial tumors (endometrioid and serous), and discovered that some endometrioid tumors have developed a very similar pattern to serous tumors, suggesting they may benefit from a common treatment (Network (2013)). Particularly, the serous and some of the endometrioid tumors are characterized by frequent mutations in *TP53*, extensive copy number alterations and few DNA methylation changes. The rest of the endometrioid tumors are characterized by few copy number alterations, scarce mutations in *TP53* and frequent mutations in *PTEN* and *KRAS*. *TP53* and *PTEN* are abbreviations of tumor protein p53 and phosphatase and tensin homolog, both are tumor suppressors (Surget et al. (2014) and Steck et al. (1997)). The normal *KRAS* protein performs essentially tissue signaling, and the mutation of a *KRAS* gene is an essential step in the development of many cancers (Kranenburg (2005)). *PTEN*, *KRAS* and *TP53* genes are located on chromosome 10, 12 and 17, respectively. In this study, we focus on chromosome 10 due to the size of computations required.

4.2 Data preparation

According to the biological functional hierarchy and the nature of method DNSMI we decide to use DNA methylation beta value as $\boldsymbol{\mathcal{G}}$, transcriptome profiling as

Table 11: RDT0 data retrieval criteria from GDC for TCGA-UCEC project. $\mathbf{Y}$ is specifically percent tumor invasion and it is defined as the value for percent calculated as depth of myometrial invasion divided by depth of myometrial thickness.

	$\boldsymbol{\mathcal{G}}$	$\boldsymbol{\mathcal{X}}$	$\mathbf{Y}$
Data Category	DNA Methylation	Transcriptome Profiling	Clinical
Data Type	Methylation Beta Value	Gene Expression Quantification	Clinical Supplement
Workflow Type	Liftover	HTSeq - FPKM-UQ	NA
Platform	Illumina Human Methylation 450	NA	NA
Dimension	485 x 485577	587 x 56963	548 x 1

$\boldsymbol{\mathcal{X}}$ and percent tumor invasion as $\mathbf{Y}$. The criteria used to retrieve data from Genomic Data Commons Data Portal (GDC) and the resulting data dimension are in Table 11, the data set ($\boldsymbol{\mathcal{G}}$, $\boldsymbol{\mathcal{X}}$, $\mathbf{Y}$) is named Reduction Data 0 (RDT0).

The RDT0 data set is then filtered by only choosing primary solid tumor for sample type and endometrioid endometrial adenocarcinoma for histological type as well as excluding any missing values. The resulted data set is named RDT1 ($\boldsymbol{\mathcal{G}}$ (269 x 10135), $\boldsymbol{\mathcal{X}}$ (269 x 2107), $\mathbf{Y}$ (269 x 1)). For methylation data, we transform Beta-values to M-values since M-values are more statistically valid (Du et al. (2010)). RDT1 is used as input to DNSMI as well as other algorithms in Section 3.2. The range of percent tumor invasion is shown in Figure 6.

4.3 Results

4.3.1 Results of DNSMI

Using $\delta = 0.05$, DNSMI selects 278 DNA methylation composite elements out of 10135 and 39 genes out of 2107. We examine the interactions between the selected g_i's, x_j's and the outcome y, percent tumor invasion, by univariate hypothesis tests. Results show that 101 out of 278 (36.3%) DNA methylation elements and 35 out of 39 (89.7%) genes are individually statistically significantly associated with the outcome y at 0.05 level. Table 1 and 2 of Appendix 3 show the annotations, p-values as well as estimates and correlations for the 278 and 39 found elements. Within the 278 DNA methylation elements and the 39 genes, i.e. 10842 pairs of DNA methylation and genes, 7515 pairs (69.3%) show a significant association between the pair elements. These pairs may be pathway variables that cannot be discovered by standard methods. In conclusion, DNSMI suggests several causal pathway candidates in which each pathway component is significantly associated with each other and with the outcome.

Among the genes and DNA methylation sites found by DNSMI for UCEC project many are demonstrated to be closely associated with endometrial cancer or other epithelial cancers. Qiu et al. (2013) found that *EMX2* (the human

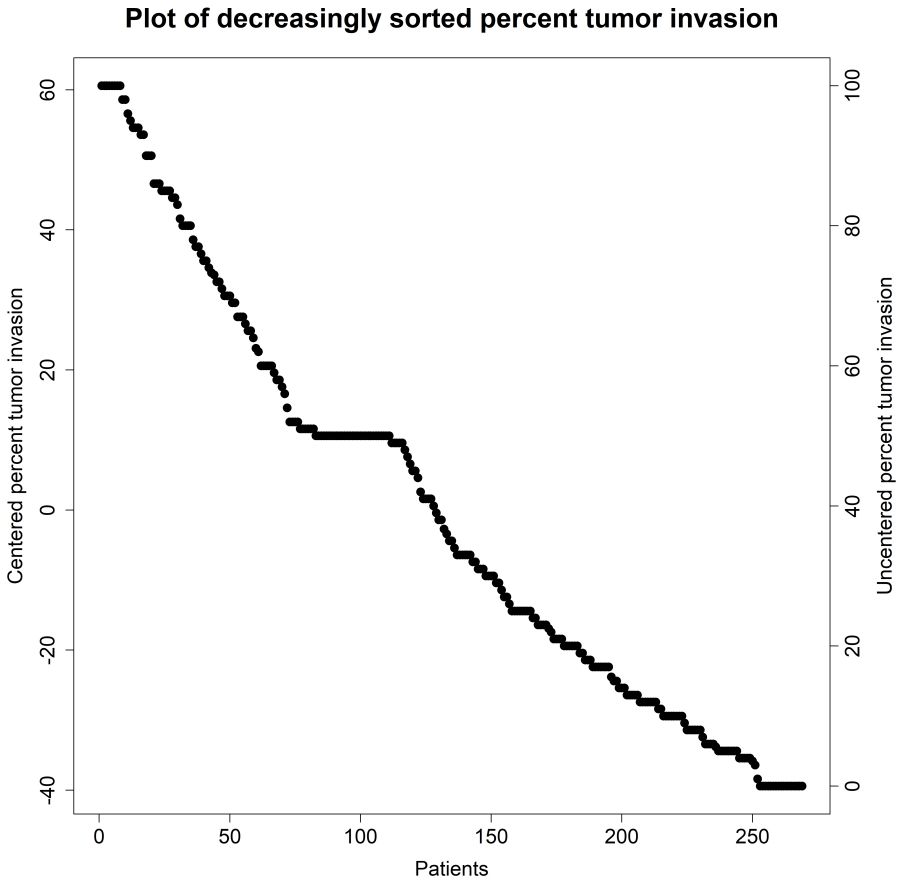


Figure 6: Decreasingly sorted percent tumor invasion variable used as outcome in our analysis.

homologue of *Drosophila* empty spiracles gene 2) was significantly downregulated in endometrial cancer tissues and this was correlated with the tumor stage, grade, and the depth of myometrial invasion. Similar downregulation of *EMX2* was also observed in lung cancer samples in Okamoto et al. (2010) and this downregulation was associated with methylation of the *EMX2* promoter. In Table 1 of Appendix 3 we see that the DNA methylation level of *EMX2* (cg07895186) is positively associated with the degree of tumor invasion and because DNA methylation acts to repress transcription, high level of methylation means low level of expression and this is consistent with the aforementioned findings.

Li et al. (2016) identified that the overexpression of *MCM10* (Minichromosome Maintenance Complex Component 10), a member of *MCM* gene family who are key factors for the initiation of DNA replication, was associated with unfavorable clinicopathological characteristics and independent negative prognostic effects, justifying its potential therapeutic and diagnostic value in urothelial carcinoma, an epithelial cancer. In our example, the DNA methylation level of *MCM10* (cg05505307, cg01237870) is found to be negatively correlated with the degree of tumor invasion, which means that the expression level is positively correlated with the invasion and this also matches the findings in Li et al. (2016).

In addition, PAR3 (partition defective 3) protein, encoded by *PARD3* gene, has an important role in mammals in the formation in the epithelia of the tight junctions which is a specialized type of intercellular adhesion complex that defines the apical–lateral border of the cell membrane compartments (Goldstein and Macara (2007), Laprise and Tepass (2011), and Martin-Belmonte and Perez-Moreno (2012)). The deletion and reduced expression of *PARD3* was observed to be a novel mechanism that is behind the progression and metastasis of lung squamous cell carcinomas (LSCC) in Bonastre et al. (2015) and human esophageal squamous cell carcinoma (ESCC) in Zen et al. (2009). In DNSMI result, the expression of *PARD3* is negatively related to the tumor invasion adjusted for other covariates, however, it is not significant after the adjustment.

On the other hand, Xu et al. (2013) reported that *DHTKD1* (dehydrogenase E1 and transketolase domain-containing 1) plays a critical role in energy production in mitochondria which are vital energy factories involved in cell cycle, cell differentiation, metabolic rates and energy requirement (Liesa et al. (2009)). It's also worth to note that there are 4 genes, i.e. *DDX21* (DEAD (Asp-Glu-Ala-Asp)-box RNA helicase), *C10orf91* (chromosome 10 open reading frame 91), *DHTKD1* and *FUT11* (fucosyltransferase 11) are still significant after the adjustment (Table 2 of Appendix 3).

However, the *PTEN* gene mentioned in Section 4.1 is not selected by DNSMI. It may be because the values of the elements in **NSM** that involve *PTEN*, either in methylation or transcription, are very small. The *PTEN* associated values in **NSM**, 115435 of them, have a maximum of 17.86 with a mean of 0.18 while the maximum and the mean of the entire **NSM** is 111.98 and 0.23 (Table 12a) and the maximum and the mean for DNSMI selections are 109.26 and 18.7 (Table 12b).

Table 12: Distribution of elements values for:

(a) **NSM** (10135 × 2107) generated from RDT1

N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
21354445	0.00	0.00	0.00	0.23	0.07	111.98

(b) Submatrix (278 × 39) formed by DNSMI selected elements from **NSM** (10135 × 2107) generated from RDT1

N	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
10842	0.00	11.69	16.80	18.70	23.74	109.26

4.3.2 Results of AMSE-PMD, sSCCA-P and sSCCA-W

Among the 10135 DNA methylation sites and 2107 genes, AMSE-PMD using the same search grid as DNSMI in 4.3.1 selects 9813 DNA methylation sites and all 2107 genes while sSCCA-P and sSCCA-W have a $\widetilde{\boldsymbol{\mathcal{G}}}$ with 145 columns and a $\widetilde{\boldsymbol{\mathcal{X}}}$ of 0 columns after the first filteration step using FDR = 0.2, which means there is no discovery for these two methods.

5 Implementation

5.1 Simulation

The simulation codes have been made public at **`https://github.com/fzhang8/DNSMI-simulation`** including a help file.

5.2 UCEC example

RDT0 data set was generated by using the R packages "TCGAbiolinks" (Colaprico et al. (2016)) and "SummarizedExperiment" (Morgan et al. (2017)), where clinical data contains the variable percent tumor invasion. After RDT1 was generated and according to Step 1 and 2 from Algorithm 1 (Appendix 1), $\mathbf{NSM}_r$, $r = 1, \ldots, 100$ were generated. A search grid of dimension 100 × 46, i.e. $h = 100, l = 46$ from Algorithm 1 (Appendix 1), was used. Then a centralized hub file (CHF) was created with each record being a unique index combination between $\mathbf{NSM}_r$ and 4600 search grid points and thus this hub file has 460000 records. The CHF is then sent to each processor of a high-performance server which has 1024 processors across over 64 nodes and each node has 64 GB memory. For each of the processors, it will independently and randomly draw a record from the CHF and does the PMD using the $\mathbf{NSM}_r$ and (c_1, c_2) that are associated with that record. Due to the duplication from the randomness of the drawing, manual shrinkage on the CHF was carried out periodically to improve the efficiency until there was 0 record left in the

CHF. The results from all the 460000 decompositions were assembled and processed following the rest steps from Algorithm 1 (Appendix 1). The codes to make RDT1 data set from extracting data from TCGA have been made public at `https://github.com/fzhang8/DNSMI_example_TCGA` with a help file of illustrations.

6 Discussion

From the simulations sparsity plays a crucial role in the performance of DNSMI, where, by design, the sparser the better, i.e. small a/p and b/q. The PMD method in DNSMI uses a binary search to find the threshold of the soft thresholding operator which is used to constrain the $\mathbf{u}$ and $\mathbf{v}$ as in (9). Thus large values of $\mathbf{u}$ and $\mathbf{v}$ will be selected before small values. In other words, signals will be selected before low signal noise and large signals before small signals. The selection of large signals across all the subsamples contribute little instability, but they account for a large portion of the constraints to be met, i.e. c_1 and c_2 in (9). This results in a situation where there is only a modest nonsignal added before the instability threshold is achieved. Using a low instability threshold of 0.05 is signal conservative and will force the algorithm to favor sparse settings. As a consequence, in less sparse cases, only a small portion of the signals are selected, leading to a low TPR and a high TNR as shown in Table 8.

Two components for each element of the **NSM** are derived from simple regression using OLS: $\hat{\beta}_{y|g_i}$ and $\hat{\beta}_{x_j|g_i}$. Therefore, the selected elements on both dimensions may contain the effects from other elements that are not included in the regression. And this confounding issue is reflected as shown in Table 2 of Appendix 3. On the other hand, the predictors are in their first order and thus only the linear relations are captured by DNSMI. An implicit assumption of DNSMI is that the hierarchical relation between the two information layers as well as the outcome is known. Therefore, DNSMI may not be well suited under those circumstances where this relation is unknown or where there exists a feedback loop within the hierarchy.

The indices of the selected elements by DNSMI on both dimensions can be used to extract a submatrix from the original **NSM** matrix. For example, such submatrix is of dimension 278×39 for the RDT1 dataset in Section 4.3.1 and the distribution of its elements is listed in Table 12b. Outside of such submatrix, **NSM** may still have some elements that are larger than the maximum value in such a submatrix. Such elements and the associated marginal elements should also be monitored given the goal of the DNSMI and the belief that large elements represent large effects. For instance, there are 3 such elements in **NSM** generated from RDT1 of Section 4.3.1. These elements span 2 DNA methylation sites and 3 genes (Table 13) and the aforementioned *EMX2* also appears in this list.

Note DNSMI is very computationally intensive and memory consuming for certain applications. For example, in Section 5.2, we used a parameter search grid of 4600 points (100 and 46 on $\mathbf{u}$ and $\mathbf{v}$ dimensions) and each point will be used to implement 100 PMD decompositions of which each PMD is on a different **NSM**

Table 13: Annotations and significance of DNA methylation sites and transcriptome that are associated with **NSM** elements which are larger than that of DNSMI selections from RDT1.

(a) DNA methylation

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg04683551	chr10	CDNF|HSPA14	0.02	24.485	0.142
cg07895186	chr10	EMX2|EMX2OS	0.021	23.769	0.141

[a]Simple linear regression using percent tumor invasion as response variable.
[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.
[c]Correlation with percent tumor invasion.

(b) Transcriptome

Ensemble Gene ID	Chromosome	Gene name	p-value[a]	Estimate[b]	Correlation[c]
ENSG00000148773	chr10	MKI67	0.002	0	0.184
ENSG00000186766	chr10	FOXI2	0.055	0	0.117
ENSG00000213551	chr10	DNAJC9	0.020	0	0.142

[a]Simple linear regression using percent tumor invasion as response variable.
[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.
[c]Correlation with percent tumor invasion.

matrix whose dimension is 10135×2107. As a result, 460000 decompositions need to be carried out. The first issue here is the memory capacity. Since numeric vectors occupy 8 bytes for every element in R, loading those 100 **NSM** matrices into memory alone will take about 16 GB. Furthermore, every such decomposition will averagely cost about 55.72 seconds (timed by "microbenchmark" package) on a laptop which is equipped with Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz with 16.0 GB RAM. In other words, it will take roughly 300 days to do the analysis as in Section 4 if it is to be carried out without parallelism on a laptop with the similar settings. Thus, in practice, DNSMI is more suited for use on high performance computing clusters.

7 Conclusions

In this study, we proposed an algorithm called Decomposition of Network Summary Matrix via Instability (DNSMI) which provides a supervised and sparse solution for network detection. Simulations were carried out to test its performance regarding three different factors: signal to nonsignal ratio, sample size and sparsity. DNSMI performed very well for each of the factors compared to other methods, especially in sparse setting. DNSMI is then applied on the TCGA-UCEC project and a sparse solution is obtained which contains several known biologically meaningful pathway candidates. The implementation of DNSMI and its limitations are also discussed.

Main tools: **Software:** RStudio (1.1.383), R (3.4.1), "PMA" package (1.0.9), "microbenchmark" package (1.4.3), "TCGAbiolinks" package (2.5.9), "SummarizedExperiment" package (1.6.5), "biomaRt" package (2.32.1), "stringr" package (1.2.0), "GenomicRanges" package (1.28.6), "parallel" package (3.4.1). **Hardware:** DELL laptop with 8 processors each being Intel(R) Core(TM) i7-6700HQ CPU @ 2.60GHz plus 16.GB RAM and 1 TB hard drive, Roswell Park Comprehensive Cancer Center high-performance computing (HPC) resources for which there are 64 nodes having 2 Intel(R) Xeon(R) CPU E5-2670 HP SL230 G8 Servers with @ 2.60GHz with 8 cores each on the main partition.

Acknowledgement: The authors would like to present special thanks to Martin Morgan, PhD, director of R/Bioconductor project, for his assistance with portions of the R coding and for providing access to Roswell Park Comprehensive Cancer Center high-performance computing resources.

References

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Bonastre, Ester et al. (2015). "PARD3 inactivation in lung squamous cell carcinomas impairs STAT3 and promotes malignant invasion". *Cancer Research* 75.7, pp. 1287–1297.

Boronow, RC et al. (1984). "Surgical staging in endometrial cancer: clinical-pathologic findings of a prospective study." *Obstetrics & Gynecology* 63.6, pp. 825–832.

Cochran, William G (1938). "The omission or addition of an independent variate in multiple linear regression". *Supplement to the Journal of the Royal Statistical Society* 5.2, pp. 171–176.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". *Educational and Psychological Measurement* 20.1, pp. 37–46.

Colaprico, Antonio et al. (2016). "TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data". *Nucleic Acids Research* 44.8, e71. DOI: 10.1093/nar/gkv1507. eprint: /oup/backfile/content_public/journal/nar/44/8/10.1093_nar_gkv1507/1/gkv1507.pdf. URL: http://dx.doi.org/10.1093/nar/gkv1507.

Du, Pan et al. (2010). "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis". *BMC Bioinformatics* 11.1, p. 587.

Eckart, Carl and Gale Young (1936). "The approximation of one matrix by another of lower rank". *Psychometrika* 1.3, pp. 211–218.

Feuerman, Martin and Allen R Miller (2008). "Relationships between statistical measures of agreement: sensitivity, specificity and kappa". *Journal of Evaluation in Clinical Practice* 14.5, pp. 930–933.

Gamazon, Eric R et al. (2015). "A gene-based association method for mapping traits using reference transcriptome data". *Nature Genetics* 47.9, pp. 1091–1098.

Goldstein, Bob and Ian G Macara (2007). "The PAR proteins: fundamental players in animal cell polarization". *Developmental Cell* 13.5, pp. 609–622.

Kranenburg, Onno (2005). *The KRAS oncogene: past, present, and future.*

Laprise, Patrick and Ulrich Tepass (2011). "Novel insights into epithelial polarity proteins in Drosophila". *Trends in Cell Biology* 21.7, pp. 401–408.

Lê Cao, Kim-Anh et al. (2008). "A sparse PLS for variable selection when integrating omics data". *Statistical Applications in Genetics and Molecular Biology* 7.1.

Li, Wei-Ming et al. (2016). "MCM10 overexpression implicates adverse prognosis in urothelial carcinoma". *Oncotarget* 7.47, p. 77777.

Liesa, Marc, Manuel Palacín, and Antonio Zorzano (2009). "Mitochondrial dynamics in mammalian health and disease". *Physiological Reviews* 89.3, pp. 799–845.

Lin, Gigin et al. (2009). "Myometrial invasion in endometrial cancer: diagnostic accuracy of diffusion-weighted 3.0-T MR imaging-initial experience". *Radiology* 250.3, pp. 784–792.

Liu, Han, Kathryn Roeder, and Larry Wasserman (2010). "Stability approach to regularization selection (stars) for high dimensional graphical models". *Advances in Neural Information Processing Systems*, pp. 1432–1440.

Martin-Belmonte, Fernando and Mirna Perez-Moreno (2012). "Epithelial cell polarity, stem cells and cancer". *Nature Reviews Cancer* 12.1, pp. 23–38.

McGuire, Shelley (2016). "World cancer report 2014. Geneva, Switzerland: World Health Organization, international agency for research on cancer, WHO Press, 2015". *Advances in Nutrition: An International Review Journal* 7.2, pp. 418–419.

McHugh, Mary L (2012). "Interrater reliability: the kappa statistic". *Biochemia Medica* 22.3, pp. 276–282.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473.

Miecznikowski, Jeffrey C et al. (2016). "Identification of consistent functional genetic modules". *Statistical Applications in Genetics and Molecular Biology* 15.1, pp. 1–18.

Morgan, M et al. (2017). "SummarizedExperiment: SummarizedExperiment container." *R package version 1.8.1*.

Morgan, Xochitl C et al. (2015). "Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease". *Genome Biology* 16.1, p. 67.

Network, Cancer Genome Atlas Research et al. (2013). "Integrated genomic characterization of endometrial carcinoma". *Nature* 497.7447, p. 67.

Okamoto, J et al. (2010). "EMX2 is epigenetically silenced and suppresses growth in human lung cancer". *Oncogene* 29.44, pp. 5969–5975.

Parkhomenko, Elena, David Tritchler, and Joseph Beyene (2009). "Sparse canonical correlation analysis with application to genomic data integration". *Statistical Applications in Genetics and Molecular Biology* 8.1, pp. 1–34.

Qiu, Haifeng et al. (2013). "EMX2 is downregulated in endometrial cancer and correlated with tumor progression". *International Journal of Gynecological Pathology* 32.2, pp. 193–198.

Sheets, SEER Stat Fact (2015). "Endometrial cancer". *Surveillance, Epidemiology, and End Results Program. seer. cancer. gov*.

Steck, Peter A et al. (1997). "Identification of a candidate tumour suppressor gene, MMAC1, at chromosome 10q23. 3 that is mutated in multiple advanced cancers". *Nature Genetics* 15.4, pp. 356–362.

Surget, Sylvanie, Marie P Khoury, and Jean-Christophe Bourdon (2014). "Uncovering the role of p53 splice variants in human malignancy: a clinical perspective". *OncoTargets and Therapy* 7, p. 57.

Tibshirani, Robert et al. (2005). "Sparsity and smoothness via the fused lasso". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, pp. 91–108.

Waaijenborg, Sandra, Philip C Verselewel de Witt Hamer, and Aeilko H Zwinderman (2008). "Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis". *Statistical Applications in Genetics and Molecular Biology* 7.1.

Weinstein, John N et al. (2013). "The cancer genome atlas pan-cancer analysis project". *Nature Genetics* 45.10, pp. 1113–1120.

Wermuth, Nanny (1992). "On block-recursive linear regression equations". *Brazilian Journal of Probability and Statistics*, pp. 1–32.

Witten, Daniela M and Robert J Tibshirani (2009). "Extensions of sparse canonical correlation analysis with applications to genomic data". *Statistical Applications in Genetics and Molecular Biology* 8.1, pp. 1–27.

Witten, Daniela M, Robert Tibshirani, and Trevor Hastie (2009). "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis". *Biostatistics* 10.3, pp. 515–534.

Xu, Wangyang et al. (2013). "DHTKD1 is essential for mitochondrial biogenesis and function maintenance". *FEBS Letters* 587.21, pp. 3587–3592.

Zen, K et al. (2009). "Defective expression of polarity protein PAR-3 gene (PARD3) in esophageal squamous cell carcinoma". *Oncogene* 28.32, pp. 2910–2918.

Appendix 1: List of algorithms

Algorithm 1 DNSMI at δ level for observation matrices $\mathbf{Y}_{N\times 1}$, $\boldsymbol{\mathcal{X}}_{N\times q}$ and $\boldsymbol{\mathcal{G}}_{N\times p}$ for a search grid containing h elements on c_1 direction and l elements on c_2 direction.

1: Generate subsampled $\mathbf{Y}_r^S$ $(0.5N\times 1)$, $r=1,\ldots,R$, by drawing $0.5N$ observations randomly without replacement from $\mathbf{Y}$ where "S" indicates subsample. Likewise for $\boldsymbol{\mathcal{X}}_r^S$ $(0.5N\times q)$ and $\boldsymbol{\mathcal{G}}_r^S$ $(0.5N\times p)$.

2: Calculate $\mathbf{NSM}_r$ matrix for each subsampled $(\mathbf{Y}_r^S,\ \boldsymbol{\mathcal{X}}_r^S,\ \boldsymbol{\mathcal{G}}_r^S)$.

3: Given a (c_{1m}, c_{2n}) pair, $m=1,2,\ldots,h, n=1,2,\ldots,l$.

4: Calculate $\boldsymbol{\mathcal{U}}_{p\times R}=(\mathbf{u}_1,\ldots,\mathbf{u}_r,\ldots,\mathbf{u}_R), \boldsymbol{\mathcal{V}}_{q\times R}=(\mathbf{v}_1,\ldots,\mathbf{v}_r,\ldots,\mathbf{v}_R)$ where $\mathbf{u}_r$ and $\mathbf{v}_r$ are sparse solutions from applying PMD using c_{1m} and c_{2n} from 3) on $\mathbf{NSM}_r$ from 2). Set all nonzero elements in $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$ to 1.

5: Calculate $\boldsymbol{\theta}_u=\widehat{\Pr}\begin{pmatrix} u_1 \text{ is selected} \\ \vdots \\ u_p \text{ is selected}\end{pmatrix}_{p\times 1}, \boldsymbol{\theta}_v=\widehat{\Pr}\begin{pmatrix} v_1 \text{ is selected} \\ \vdots \\ v_q \text{ is selected}\end{pmatrix}_{q\times 1}$ by computing row means for $\boldsymbol{\mathcal{U}}$ and $\boldsymbol{\mathcal{V}}$.

6: Calculate

$$\widehat{\boldsymbol{\xi}}^u(c_{1m},c_{2n})=2\boldsymbol{\theta}_u(1-\boldsymbol{\theta}_u)=\begin{pmatrix}\widehat{\xi_1^u}(c_{1m},c_{2n})\\ \vdots \\ \widehat{\xi_p^u}(c_{1m},c_{2n})\end{pmatrix}_{p\times 1}$$

$$\widehat{\boldsymbol{\xi}}^v(c_{1m},c_{2n})=2\boldsymbol{\theta}_v(1-\boldsymbol{\theta}_v)=\begin{pmatrix}\widehat{\xi_1^v}(c_{1m},c_{2n})\\ \vdots \\ \widehat{\xi_q^v}(c_{1m},c_{2n})\end{pmatrix}_{q\times 1}$$

7: Calculate
$\widehat{\xi_u}(c_{1m},c_{2n})=mean(\widehat{\boldsymbol{\xi}}^u(c_{1m},c_{2n}))$, $\widehat{\xi_v}(c_{1m},c_{2n})=mean(\widehat{\boldsymbol{\xi}}^v(c_{1m},c_{2n}))$

8: Compute $\widehat{\xi}_{u,v}(c_{1m},c_{2n})=max(\widehat{\xi_u}(c_{1m},c_{2n}),\widehat{\xi_v}(c_{1m},c_{2n}))$

9: Compute $\widehat{\overline{\xi}}(c_{1m},c_{2n})=\sup_{1\le s\le c_{1m},1\le t\le c_{2n}}\widehat{\xi}_{u,v}(s,t)$

10: Select

$$(c_1,c_2)=\arg\max_{(c_{1m},c_{2n})\in E}\left\|\begin{pmatrix}c_{1m}\\ c_{2n}\end{pmatrix}\right\|_2$$

where $E=\{(c_{1m},c_{2n})\mid \widehat{\overline{\xi}}(c_{1m},c_{2n})\le\delta$ over the $h\times l$ search grid for a preset $\delta\}$.

11: Apply PMD using the selected (c_1,c_2) from 10) on **NSM** that is generated from $\mathbf{Y}_{N\times 1}$, $\boldsymbol{\mathcal{X}}_{N\times q}$ and $\boldsymbol{\mathcal{G}}_{N\times p}$. The indices corresponding to nonzero elements in the sparse output $\mathbf{u}$ and $\mathbf{v}$ represent the $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$, respectively.

Algorithm 2 Average Mean Squared Error tuned PMD decomposition of **NSM** (AMSE-PMD)

1: Given matrix **NSM**, randomly delete 10% of the data elements over the entire matrix, resulting in $\mathbf{NSM}_i$, $i = 1, 2, \ldots, 10$. Note that for each i, the 10% data are nonoverlapping.
2: Apply PMD on the 10 $\mathbf{NSM}_i$'s using a given pair (c_1, c_2).
3: Calculate for each i the mean squared error only of the missing locations in $\mathbf{NSM}_i$ to that of **NSM**.
4: The AMSE is the average of the 10 means from above step, and each pair of (c_1, c_2) will be associated with one such error.
5: The optimal (c_1, c_2) will be the one that corresponds to the smallest AMSE over the entire search grid if there is one.
6: Apply PMD using the selected (c_1, c_2) from 5) on **NSM**. The indices corresponding to nonzero elements in the sparse output **u** and **v** represent the $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$, respectively.

Algorithm 3 Supervised Sparse Canonical Correlation Analysis with SCCA from Parkhomenko et al. (2009) (sSCCA-P)

1: Prefilter features in $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{X}}$ by Benjamini-Hochberg (BH) procedure for FDR = 0.2, which produce $\widetilde{\boldsymbol{\mathcal{G}}}$ and $\widetilde{\boldsymbol{\mathcal{X}}}$.

2: Center and standardize the $\widetilde{\boldsymbol{\mathcal{X}}}$ and $\widetilde{\boldsymbol{\mathcal{G}}}$ matrices so that they have zero column means and unit variances.

3: Calculate sample correlation matrix between $\widetilde{\boldsymbol{\mathcal{X}}}$ and $\widetilde{\boldsymbol{\mathcal{G}}}$ as K.

4: Given a pair of parameters (ζ_u, ζ_v) each of which ranges from 0 to 2.

5: Select initial values $\mathbf{u}^0$ and $\mathbf{v}^0$ and set $i = 0$.

6: Update $\mathbf{u}$:

(a) $\mathbf{u}^{i+1} \leftarrow K\mathbf{v}^i$

(b) Normalize: $\mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$

(c) Apply soft thresholding to obtain sparse solution: $\mathbf{u}_j^{i+1} \leftarrow (|\mathbf{u}_j^{i+1}| - \frac{1}{2}\zeta_u)_+ Sign(\mathbf{u}_j^{i+1})$ for $j = 1, \ldots, p$ ▷

- $(.)_+$ equals to x if $x \geq 0$ and 0 if $x < 0$
- $Sign(x) = \begin{cases} -1 & \text{if } x < 0, \\ 1 & \text{if } x > 0, \\ 0 & \text{if } x = 0. \end{cases}$

(d) Normalize: $\mathbf{u}^{i+1} \leftarrow \frac{\mathbf{u}^{i+1}}{\|\mathbf{u}^{i+1}\|}$

7: Update $\mathbf{v}$:

(a) $\mathbf{v}^{i+1} \leftarrow K^{`}\mathbf{u}^{i+1}$

(b) Normalize: $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

(c) Apply soft thresholding to obtain sparse solution: $\mathbf{v}_j^{i+1} \leftarrow (|\mathbf{v}_j^{i+1}| - \frac{1}{2}\zeta_v)_+ Sign(\mathbf{v}_j^{i+1})$ for $j = 1, \ldots, q$

(d) Normalize: $\mathbf{v}^{i+1} \leftarrow \frac{\mathbf{v}^{i+1}}{\|\mathbf{v}^{i+1}\|}$

8: $i \leftarrow i + 1$

9: Repeat steps 6 and 7 until convergence.

10: The optimal pair of (ζ_u, ζ_v) will be determined by using k-fold cross-validation and will be the one who corresponds to the highest Δ_{cor} in the search grid where

$$\Delta_{cor} = \frac{1}{k}\sum_{j=1}^{k} |cor(\boldsymbol{\mathcal{X}}_j \hat{\mathbf{v}}^{-j}, \boldsymbol{\mathcal{G}}_j \hat{\mathbf{u}}^{-j})|,$$

11: Repeat steps 5–9 using the selected parameters from 10) and the indices corresponding to nonzero elements in the sparse output $\mathbf{u}$ and $\mathbf{v}$ represent the $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$, respectively.

Algorithm 4 Supervised Sparse Canonical Correlation Analysis with SCCA from Witten and Tibshirani (2009) (sSCCA-W)

1: Prefilter features in $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{X}}$ by Benjamini-Hochberg (BH) procedure for FDR = 0.2, which produce $\widetilde{\boldsymbol{\mathcal{G}}}$ and $\widetilde{\boldsymbol{\mathcal{X}}}$.
2: Center and standardize the $\widetilde{\boldsymbol{\mathcal{G}}}$ and $\widetilde{\boldsymbol{\mathcal{X}}}$ matrices so that they have zero column means and unit variances.
3: Given a pair of parameters (c_1, c_2).
4: Set $\mathbf{w}_2$ to have L_2 norm 1.
5: Iterate (a) and (b) until convergence:

(a) $\mathbf{w}_1 \leftarrow \frac{S(\widetilde{\boldsymbol{\mathcal{G}}}^T \widetilde{\boldsymbol{\mathcal{X}}} \mathbf{w}_2,\ \Delta_1)}{\|S(\widetilde{\boldsymbol{\mathcal{G}}}^T \widetilde{\boldsymbol{\mathcal{X}}} \mathbf{w}_2, \Delta_1)\|_2}$, where $\Delta_1 = 0$ if this results in $\|\mathbf{w}_1\|_1 \leq c_1$; otherwise, $\Delta_1 > 0$ is chosen so that $\|\mathbf{w}_1\|_1 = c_1$.

(b) $\mathbf{w}_2 \leftarrow \frac{S(\widetilde{\boldsymbol{\mathcal{X}}}^T \widetilde{\boldsymbol{\mathcal{G}}} \mathbf{w}_1,\ \Delta_2)}{\|S(\widetilde{\boldsymbol{\mathcal{X}}}^T \widetilde{\boldsymbol{\mathcal{G}}} \mathbf{w}_1, \Delta_2)\|_2}$, where $\Delta_2 = 0$ if this results in $\|\mathbf{w}_2\|_1 \leq c_2$; otherwise, $\Delta_2 > 0$ is chosen so that $\|\mathbf{w}_2\|_1 = c_2$. ▷ $S(.)$ denotes the soft-thresholding operator; that is, $S(a, c) = sgn(a)(|a| - c)_+$.

6: Compute $z = Cor(\widetilde{\boldsymbol{\mathcal{X}}} \mathbf{w}_1, \widetilde{\boldsymbol{\mathcal{G}}} \mathbf{w}_2)$.
7: For $i \in 1, \ldots, N$, N is a large number for permutation purpose.

I Permute the rows of $\widetilde{\boldsymbol{\mathcal{G}}}$ to obtain the matrix $\widetilde{\boldsymbol{\mathcal{G}}}^i$, and compute canonical vectors $\mathbf{w}_1^i$ and $\mathbf{w}_2^i$ using data $\widetilde{\boldsymbol{\mathcal{G}}}^i$ and $\widetilde{\boldsymbol{\mathcal{X}}}$ and tuning parameter (c_1, c_2).

II Compute $z_i = Cor(\widetilde{\boldsymbol{\mathcal{G}}}^i \mathbf{w}_1^i, \widetilde{\boldsymbol{\mathcal{X}}} \mathbf{w}_2^i)$.

8: Calculate the p-value $p = \frac{1}{N} \sum_{i=1}^{N} I(z_i \geq z)$.
9: Select the pair of (c_1, c_2) having the smallest p-value over the search grid.
10: Apply PMD using the selected (c_1, c_2) from 9) on $\widetilde{\boldsymbol{\mathcal{G}}}^T \widetilde{\boldsymbol{\mathcal{X}}}$. The indices corresponding to nonzero elements in the sparse output $\mathbf{u}$ and $\mathbf{v}$ represent the $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$, respectively.

Appendix 2: Notation dictionary

G : Latent variable for genes
G' : Latent variable for genes but is independent of Y
H : Latent variable for genes that is associated with Y but is independent of others
X : Latent variable for transcripts
X' : Latent variable for transcripts but is independent of Y
S : Latent varialbe for transcripts that is associated with Y but is independent of others
Y : Latent variable for outcome
$\boldsymbol{\mathcal{G}}, \boldsymbol{\mathcal{X}}, \mathbf{Y}$: Observation matrices for genes, transcripts and outcome
$\widetilde{\boldsymbol{\mathcal{X}}}, \widetilde{\boldsymbol{\mathcal{G}}}$: Filtered observation matrices by supervision criterion
a : Number of pathway genes
$\mathcal{A}$: set of indices of g_i's elements involved in pathway
$\widehat{\mathcal{A}}$: estimator of $\mathcal{A}$
b : Number of pathway transcripts
$\mathcal{B}$: set of indices of x_i's elements involved in pathway
$\widehat{\mathcal{B}}$: estimator of $\mathcal{B}$
c_1 : Tuning parameter of PMD decomposition method on row direction
c_2 : Tuning parameter of PMD decomposition method on column direction
d : Singular value or sparse singular value for Sd-PMD, depending on the context
h : Number of elements on c_1 direction of search grid
l : Number of elements on c_2 direction of search grid
N : Total subjects number or sample size
p : Number of rows of matrix $\mathbf{W}$, same as number of genes in analysis
q : Number of columns of matrix $\mathbf{W}$, same as number of transcripts in analysis
r : Correlation between observation and the corresponding latent variable
R : Number of subsamples to use IPMDW
R^2 : Predictability
$\mathbf{u}$: First left singular vector, sparse or not depends on the context
$\mathbf{v}$: First right singular vector, sparse or not depends on the context
α : Importance of the G $\rightarrow$ X $\rightarrow$ Y path to the total effect of G on Y
γ : $(1 - R^2)/R^2$
δ : Preset instability level
κ : Cohen's Kappa statistic
$\xi_i^u(c_1, c_2)$: Instability of ith element of vector $\mathbf{u}$
$\xi_j^v(c_1, c_2)$: Instability of jth element of vector $\mathbf{v}$
$\xi_u(c_1, c_2)$: Mean instability of $\mathbf{u}$ vector
$\xi_v(c_1, c_2)$: Mean instability of $\mathbf{v}$ vector
$\xi_{u,v}(c_1, c_2)$: Combined instability from $\mathbf{u}$ and $\mathbf{v}$ vectors
$\overline{\xi}(c_1, c_2)$: Supremum instability at (c_1, c_2)
π^2 : Variance ratio coefficient between noise and y
τ : Total effect of G on Y, $\tau = \beta_{Y|G.X} + \beta_{X|G}\beta_{Y|X.G}$

Appendix 3: Annotations of DNA methylation sites and transcriptome from DNSMI on UCEC project.

Table 1: Annotations and significance of 278 DNSMI selected DNA methylation sites from **NSM** generated from RDT1.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg00282704	chr10	CASC10|MIR1915	0.002	-10.236	-0.18
cg00363811	chr10	BTRC	0.021	-11.686	-0.14
cg00451513	chr10	ASCC1	0.022	12.044	0.14
cg00520540	chr10	CDNF|HSPA14	0.028	-11.545	-0.13
cg00766678	chr10	NPM3	0.036	-11.231	-0.13
cg00997424	chr10	PI4K2A|RP11-548K23.11	0.029	-9.053	-0.13
cg01068136	chr10	DCLRE1A|NHLRC2	0.001	-11.327	-0.2
cg01087392	chr10	GBF1	0.036	-9.99	-0.13
cg01237870	chr10	MCM10	0.004	-18.268	-0.18
cg02016328	chr10	RAB18	0.035	-15.245	-0.13
cg02024446	chr10	C10orf111|RPP38	0.035	-10.933	-0.13
cg02156071	chr10	FAM204A	0.015	-12.841	-0.15
cg02180545	chr10	C10orf2|MRPL43	0.017	-9.124	-0.14
cg02452627	chr10	ZNF438	0.01	-11.107	-0.16
cg02550110	chr10	DDX50	0.019	-12.005	-0.14
cg02733266	chr10	GSTO1	0.005	-11.123	-0.17
cg02878913	chr10	SH3PXD2A	0	-23.138	-0.27
cg02956254	chr10	RP11-298J20.4	0.04	-11.503	-0.13
cg03539850	chr10	PANK1|RP11-80H5.2	0.018	-13.527	-0.14
cg03576467	chr10	DNAJC9-AS1|MRPS16 RP11-152N13.5	0.033	-15.373	-0.13
cg03727700	chr10	DCLRE1A|NHLRC2	0.006	-9.508	-0.17
cg03801898	chr10	ADD3|ADD3-AS1	0.048	-8.146	-0.12
cg04036272	chr10	CCDC6	0.032	-10.619	-0.13
cg04126427	chr10	EIF3A	0.002	-12.379	-0.18
cg04290666	chr10	WNT8B	0.016	12.09	0.15
cg04446777	chr10	BTRC	0.024	-11.41	-0.14

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg04683551	chr10	CDNF|HSPA14	0.02	24.485	0.14
cg04959674	chr10	MMS19|UBTD1	0.018	-13.782	-0.14
cg05505307	chr10	MCM10	0.049	-12.503	-0.12
cg06206603	chr10	RP11-574K11.24|SEC24C	0.014	-10.294	-0.15
cg07203258	chr10	DDX50	0.04	-8.986	-0.13
cg07217563	chr10	WDR37	0.012	-12.144	-0.15
cg07895186	chr10	EMX2|EMX2OS	0.021	23.769	0.14
cg08069263	chr10	MXI1	0.032	-12.995	-0.13
cg08096168	chr10	CCDC6	0.045	-8.923	-0.12
cg08299755	chr10	ZFYVE27	0.025	-13.161	-0.14
cg09152955	chr10	NPM3	0.026	-10.445	-0.14
cg09269103	chr10	NFKB2	0.004	-10.915	-0.18
cg09333812	chr10	ARHGAP19|ARHGAP19-SLIT1	0.044	-12.06	-0.12
cg09478103	chr10	CAP1P2|ZNF485	0.007	-15.592	-0.17
cg09655100	chr10	TCF7L2	0.003	-11.73	-0.18
cg09747456	chr10	PANK1|RP11-80H5.2 RP11-80H5.5	0.027	-11.051	-0.14
cg09886360	chr10	CSGALNACT2|RP11-351D16.3	0.014	-10.535	-0.15
cg10325336	chr10	RP11-574K11.24|SEC24C	0.021	-10.008	-0.14
cg10708548	chr10	ARID5B	0.014	-9.667	-0.15
cg10739686	chr10	KAT6B	0.05	-9.895	-0.12
cg10800082	chr10	PDCD4|PDCD4-AS1	0.023	-11.631	-0.14
cg10878076	chr10	NDUFB8|RP11-411B6.6	0.015	9.511	0.15
cg10905918	chr10	RPS24	0.027	-14.052	-0.13
cg11223711	chr10	EIF3A	0.012	-11.642	-0.15
cg11423178	chr10	HNRNPH3|PBLD	0.04	-12.218	-0.13
cg11499984	chr10	BLOC1S2	0.006	-13.085	-0.17
cg11996395	chr10	NOLC1	0.033	-10.888	-0.13
cg12198729	chr10	BTRC	0.011	-12.527	-0.15
cg12226046	chr10	PANK1|RP11-80H5.2	0.012	-11.05	-0.15
cg12563239	chr10	ANKRD26	0.02	-10.82	-0.14
cg13800022	chr10	ITGB1|RP11-462L8.1	0.014	-16.688	-0.15
cg13830636	chr10	RP11-574K11.24|SEC24C	0.014	-10.169	-0.15
cg13962355	chr10	BTRC	0.03	-11.111	-0.13
cg14039939	chr10	LRRC27|STK32C	0.001	12.222	0.21
cg14052593	chr10	BMS1P4|DUSP8P5 |GLUD1P3|RP11-464F9.1	0.018	-13.196	-0.14
cg14461522	chr10	NPM3	0.006	-14.121	-0.17

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg14562081	chr10	TCF7L2	0.008	-10.501	-0.16
cg14700647	chr10	ASAH2B	0.031	-11.586	-0.13
cg15322766	chr10	POLR3A	0.035	-10.921	-0.13
cg15384821	chr10	EGR2	0.049	-11.094	-0.12
cg15523443	chr10	POLR3A	0.012	-12.58	-0.15
cg15831653	chr10	DNAJC1	0.001	-17.937	-0.2
cg15939466	chr10	VTI1A|ZDHHC6	0.004	-11.788	-0.17
cg15952994	chr10	VTI1A|ZDHHC6	0.021	-11.298	-0.14
cg16754967	chr10	RSU1	0.002	-11.832	-0.18
cg17555499	chr10	CHCHD1	0.015	-10.032	-0.15
cg18493566	chr10	C10orf76	0.032	-12.277	-0.13
cg18510056	chr10	ZNF503-AS2	0.013	-8.904	-0.15
cg18762013	chr10	ZNF33A	0.022	-8.76	-0.14
cg18913254	chr10	ANXA7	0.01	-11.69	-0.16
cg18928584	chr10	CUEDC2	0.04	-11.02	-0.13
cg19014323	chr10	HNRNPF	0.033	-9.228	-0.13
cg19032306	chr10	CPEB3|MARCH5	0.008	-10.692	-0.16
cg19040518	chr10	NDST2|RP11-574K11.31	0.018	-12.549	-0.14
cg19138900	chr10	KLF6	0.038	-7.942	-0.13
cg20444320	chr10	STAM|STAM-AS1	0.014	-11.784	-0.15
cg21374208	chr10	DNAJC1	0.017	-14.929	-0.15
cg21949958	chr10	BUB3	0.027	-9.949	-0.14
cg23087635	chr10	INPP5A	0.004	-17.151	-0.18
cg23319797	chr10	RAB18	0.049	-12.962	-0.12
cg23751407	chr10	RP11-95I16.2	0.028	-9.458	-0.13
cg23936098	chr10	NOLC1	0.007	-15.037	-0.16
cg23991622	chr10	VIM|VIM-AS1	0.016	-12.909	-0.15
cg24166097	chr10	NPM3	0.032	-10.913	-0.13
cg24573310	chr10	CISD1	0.042	-10.687	-0.12
cg25089494	chr10	C10orf131|ENTPD1-AS1 RP11-248J23.7	0.045	-10.352	-0.12
cg25355065	chr10	ARL3|SFXN2	0.048	-10.477	-0.12
cg25648639	chr10	ARHGAP12	0.015	-12.949	-0.15
cg26002628	chr10	ARID5B	0.015	-9.416	-0.15
cg26306372	chr10	VIM|VIM-AS1	0.043	-11.096	-0.12
cg26625369	chr10	CDC123|NUDT5	0.037	-11.197	-0.13
cg26881277	chr10	PDCD4|PDCD4-AS1	0.003	-12.846	-0.18
cg26964061	chr10	MXI1	0.004	-13.658	-0.18

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg27255678	chr10	LZTS2	0.05	-12.631	-0.12
cg27510901	chr10	DNAJC1	0.046	-11.448	-0.12
cg00440043	chr10	ZEB1|ZEB1-AS1	0.062	-9.479	-0.11
cg00588577	chr10	CCAR1	0.161	-9.448	-0.09
cg00879184	chr10	MLLT10	0.16	-6.801	-0.09
cg01021053	chr10	ZWINT	0.292	-6.844	-0.06
cg01042749	chr10	FAM178A|RP11-179B2.2	0.105	-9.749	-0.1
cg01154046	chr10	VIM|VIM-AS1	0.099	-10.199	-0.1
cg01367750	chr10	ACBD5|RP11-85G18.6	0.084	-7.675	-0.11
cg01431972	chr10	ZNF503-AS2	0.053	-10.168	-0.12
cg02150674	chr10	PHYH	0.091	-7.866	-0.1
cg02351056	chr10	METTL10|RP11-12J10.3	0.159	-8.332	-0.09
cg02622557	chr10	EIF3A	0.204	-6.689	-0.08
cg02952711	chr10	LRRC27|STK32C	0.232	-8.554	-0.07
cg03020000	chr10	ARID5B	0.179	-8.616	-0.08
cg03141879	chr10	PITRM1|RP11-298E9.7	0.203	6.307	0.08
cg03211233	chr10	SIRT1	0.273	-7.445	-0.07
cg03361817	chr10	ARID5B	0.178	-8.511	-0.08
cg03524461	chr10	MLLT10	0.092	-9.21	-0.1
cg03588299	chr10	DIP2C	0.276	6.334	0.07
cg03714691	chr10	WDR37	0.149	6.397	0.09
cg03922645	chr10	MEIG1	0.298	-5.658	-0.06
cg03941040	chr10	TFAM	0.468	-5.432	-0.04
cg04167018	chr10	ECD|FAM149B1	0.109	-10.809	-0.1
cg04179819	chr10	TAF3	0.14	13.568	0.09
cg04534276	chr10	PPP3CB|PPP3CB-AS1	0.07	-10.027	-0.11
cg04622176	chr10	MCMBP|SEC23IP	0.25	-8.036	-0.07
cg04646451	chr10	DDX50	0.29	-6.457	-0.06
cg04733624	chr10	ADK	0.074	17.371	0.11
cg04749667	chr10	ECD|FAM149B1	0.064	-9.771	-0.11
cg05088677	chr10	CASC10|MIR1915	0.243	-6.805	-0.07
cg05313070	chr10	ARHGAP12	0.213	-6.889	-0.08
cg05420251	chr10	OLMALINC	0.053	-10.755	-0.12
cg06583105	chr10	PPRC1	0.414	-5.95	-0.05
cg06649808	chr10	RP11-574K11.24|SEC24C	0.133	-8.762	-0.09
cg06782748	chr10	CREM|RP11-297A16.2	0.313	-5.926	-0.06
cg07030336	chr10	VTI1A|ZDHHC6	0.311	-4.968	-0.06
cg07301505	chr10	PI4K2A|RP11-548K23.11	0.102	-7.745	-0.1

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg07636870	chr10	ACTR1A|SUFU	0.273	-5.608	-0.07
cg07679896	chr10	RP11-298J20.4	0.084	-10.818	-0.11
cg07855525	chr10	TAF3	0.162	-8.624	-0.09
cg07900823	chr10	NUTM2B-AS1|RP11-182L21.6	0.141	10.984	0.09
cg08395899	chr10	UPF2	0.233	-7.141	-0.07
cg08616269	chr10	CCAR1	0.116	-10.837	-0.1
cg08668510	chr10	IDI1|WDR37	0.112	-8.677	-0.1
cg08797625	chr10	CAMK2G	0.286	-5.494	-0.07
cg08799865	chr10	NT5C2	0.136	-8.556	-0.09
cg08905519	chr10	FAM208B	0.09	-12.262	-0.1
cg09219177	chr10	ACBD5|RP11-85G18.6	0.149	-8.413	-0.09
cg09391093	chr10	RP11-393J16.4|ZNF25	0.077	-9.92	-0.11
cg09526975	chr10	SEPHS1	0.13	-8.258	-0.09
cg09563120	chr10	RP11-108L7.15	0.07	-8.55	-0.11
cg09688285	chr10	XPNPEP1	0.199	-8.006	-0.08
cg09933375	chr10	CCAR1	0.136	-8.499	-0.09
cg10295800	chr10	TFAM	0.146	-8.161	-0.09
cg10436918	chr10	VTI1A|ZDHHC6	0.271	-7.126	-0.07
cg10526556	chr10	SEPHS1	0.338	-6.629	-0.06
cg10609984	chr10	RP11-393J16.4|ZNF25	0.202	-6.433	-0.08
cg10831391	chr10	PFKP	0.064	17.827	0.11
cg10894697	chr10	CASP7	0.192	-7.271	-0.08
cg10928925	chr10	ZCCHC24	0.062	-10.162	-0.11
cg11271505	chr10	ZSWIM8	0.132	-8.726	-0.09
cg11420031	chr10	VPS26A	0.126	-9.04	-0.09
cg11460820	chr10	RPS24	0.065	-11.18	-0.11
cg11504511	chr10	ZMIZ1|ZMIZ1-AS1	0.094	-9.122	-0.1
cg11655691	chr10	MICU1	0.146	-8.186	-0.09
cg11660725	chr10	ANKRD26	0.179	-7.058	-0.08
cg11814667	chr10	PDCD11|USMG5	0.077	-11.843	-0.11
cg11977348	chr10	SMC3	0.076	-11.558	-0.11
cg12143181	chr10	LIPA	0.057	-8.017	-0.12
cg12144272	chr10	ZMIZ1|ZMIZ1-AS1	0.066	-9.189	-0.11
cg12276298	chr10	ECD|FAM149B1	0.234	-7.105	-0.07
cg12294817	chr10	METTL10|RP11-12J10.3	0.124	-7.881	-0.09
cg12536451	chr10	ZFYVE27	0.057	-12.562	-0.12
cg12580870	chr10	CCAR1	0.169	-7.792	-0.08
cg12823012	chr10	CCAR1	0.156	-8.948	-0.09

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg12832988	chr10	CDNF|HSPA14	0.085	-8.824	-0.11
cg13188511	chr10	CUL2	0.292	-6.597	-0.06
cg13320518	chr10	VTI1A|ZDHHC6	0.143	-7.643	-0.09
cg13509954	chr10	MTPAP	0.477	-4.996	-0.04
cg13708259	chr10	C10orf2|MRPL43	0.325	-5.19	-0.06
cg13763308	chr10	ADK|AP3M1	0.111	-11.26	-0.1
cg13799287	chr10	WAC|WAC-AS1	0.253	-9.547	-0.07
cg13817732	chr10	BLOC1S2	0.319	-5.838	-0.06
cg14193565	chr10	INPP5F	0.05	-11.468	-0.12
cg14409890	chr10	PCGF6	0.156	-9.557	-0.09
cg14434255	chr10	ENTPD7	0.244	-6.679	-0.07
cg14440794	chr10	FAM171A1	0.051	12.832	0.12
cg14631462	chr10	ZEB1|ZEB1-AS1	0.191	-6.402	-0.08
cg14694828	chr10	ZMYND11	0.074	-9.632	-0.11
cg14825384	chr10	CASP7	0.41	-6.934	-0.05
cg14924826	chr10	BBIP1|SHOC2	0.144	-7.789	-0.09
cg15055039	chr10	BAG3	0.423	-5.22	-0.05
cg15169471	chr10	LINC00863|NUTM2A-AS1	0.395	-6.242	-0.05
cg15172601	chr10	CDK1	0.064	-12.138	-0.11
cg15317221	chr10	ABI1	0.094	-7.683	-0.1
cg15317837	chr10	GTPBP4|RP11-363N22.3	0.454	-5.371	-0.05
cg15363487	chr10	VIM|VIM-AS1	0.256	9.5	0.07
cg15433901	chr10	NMT2	0.121	-8.11	-0.09
cg15462502	chr10	BMS1	0.29	-7.432	-0.06
cg15563952	chr10	DHTKD1	0.188	-8.475	-0.08
cg15825287	chr10	NRBF2	0.21	-7.643	-0.08
cg15834928	chr10	LRRC27|STK32C	0.346	-5.805	-0.06
cg15872329	chr10	BLOC1S2	0.084	-9.828	-0.11
cg16124546	chr10	ECD|FAM149B1	0.181	-9.093	-0.08
cg16139770	chr10	ARID5B	0.261	-8.344	-0.07
cg16423650	chr10	DNMBP	0.188	-8.155	-0.08
cg16584947	chr10	LARP4B	0.057	-9.119	-0.12
cg16742925	chr10	PDCD11|USMG5	0.337	-6.308	-0.06
cg16905311	chr10	ARL5B|NSUN6	0.235	-7.037	-0.07
cg17012863	chr10	LIPA	0.069	-13.118	-0.11
cg17122475	chr10	DIP2C	0.057	-15.102	-0.12
cg17465063	chr10	MXI1	0.402	-5.149	-0.05
cg17586365	chr10	CDC123|NUDT5	0.103	-8.543	-0.1

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg17629447	chr10	CHUK|RP11-316M21.6	0.34	-7.511	-0.06
cg17710288	chr10	NPM3	0.068	-10.053	-0.11
cg17982459	chr10	SFR1	0.141	-9.884	-0.09
cg18035301	chr10	MAP3K8	0.257	-6.526	-0.07
cg18221224	chr10	ACBD5	0.12	-8.052	-0.1
cg18375586	chr10	FAM171A1	0.25	-7.871	-0.07
cg18409845	chr10	CPEB3|MARCH5	0.054	-7.817	-0.12
cg18691055	chr10	MKI67	0.275	-5.129	-0.07
cg18803045	chr10	PDCD4|PDCD4-AS1	0.173	-6.725	-0.08
cg18827378	chr10	CDK1	0.095	-8.558	-0.1
cg19038917	chr10	GSTO1	0.124	-7.03	-0.09
cg19210816	chr10	EIF3A	0.075	-8.966	-0.11
cg19391892	chr10	DDX50	0.156	-6.993	-0.09
cg19402405	chr10	EGR2	0.304	-5.531	-0.06
cg19535032	chr10	KIF5B|Y_RNA	0.106	-8.799	-0.1
cg19559179	chr10	C10orf76	0.055	-12.012	-0.12
cg19577016	chr10	ANAPC16|ASCC1	0.164	-8.853	-0.09
cg19603966	chr10	DNAJC1	0.336	-6.415	-0.06
cg19716967	chr10	FAM204A	0.203	-7.977	-0.08
cg19839763	chr10	ITPRIP	0.161	-6.85	-0.09
cg19874323	chr10	ARL5B|NSUN6	0.051	-9.793	-0.12
cg20203089	chr10	NFKB2	0.058	-8.164	-0.12
cg20264529	chr10	MTPAP	0.065	-11.269	-0.11
cg20318353	chr10	ABI1	0.06	-10.792	-0.11
cg20355062	chr10	KLF6	0.158	-9.552	-0.09
cg20475000	chr10	PDCD11|USMG5	0.087	-9.876	-0.1
cg20489345	chr10	CASP7	0.072	-11.237	-0.11
cg20778294	chr10	PPRC1	0.112	-11.411	-0.1
cg21201659	chr10	MCMBP|SEC23IP	0.137	-10.717	-0.09
cg22553140	chr10	PRDX3	0.353	-6.281	-0.06
cg22593633	chr10	ZMYND11	0.165	-7.13	-0.08
cg22635723	chr10	ADD3|ADD3-AS1	0.069	-7.749	-0.11
cg22664157	chr10	PWWP2B	0.377	-6.383	-0.05
cg22860891	chr10	RBM17	0.288	-6.565	-0.06
cg23026419	chr10	ITPRIP	0.128	-6.523	-0.09
cg23087130	chr10	ABI1	0.152	-9.211	-0.09
cg23635883	chr10	MASTL|YME1L1	0.143	-8.063	-0.09
cg23638686	chr10	INPP5A	0.096	-13.741	-0.1

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Composite Element Reference	Chromosome	Gene symbol	p-value[a]	Estimate[b]	Correlation[c]
cg23654971	chr10	GBF1	0.184	-8.348	-0.08
cg24182333	chr10	PSAP	0.165	-9.404	-0.08
cg24201716	chr10	CCDC186|MIR2110	0.344	-6.744	-0.06
cg24293903	chr10	ENTPD7	0.29	-7.636	-0.06
cg24315770	chr10	PDCD11|USMG5	0.169	-8.374	-0.08
cg24807448	chr10	SMC3	0.432	-5.864	-0.05
cg24826355	chr10	KIF5B|Y_RNA	0.219	-8.398	-0.08
cg24980609	chr10	DPCD|POLL	0.228	-7.428	-0.07
cg25243854	chr10	BCCIP|UROS	0.075	-13.147	-0.11
cg25713684	chr10	TAF5	0.08	-9.639	-0.11
cg25822326	chr10	NET1	0.16	-8.127	-0.09
cg26022877	chr10	ACADSB|IKZF5	0.08	-9.874	-0.11
cg26075202	chr10	SIRT1	0.407	-6.149	-0.05
cg26097210	chr10	HNRNPH3|PBLD	0.111	-8.36	-0.1
cg26213561	chr10	CASC10|MIR1915	0.336	-6.234	-0.06
cg26273962	chr10	SORBS1	0.112	-10.027	-0.1
cg26358059	chr10	GSTO1	0.348	-5.792	-0.06
cg26485946	chr10	IDI1|WDR37	0.116	-8.325	-0.1
cg26538046	chr10	WDR11|WDR11-AS1	0.207	-7.981	-0.08
cg26538214	chr10	KLF6	0.16	-9.496	-0.09
cg27350398	chr10	PITRM1	0.366	-5.73	-0.06
cg27352063	chr10	PPIF	0.287	10.936	0.07
cg27445265	chr10	BCCIP|UROS	0.103	-9.767	-0.1
cg27503573	chr10	PAOX	0.391	-5.837	-0.05
cg27521563	chr10	ADRB1	0.35	-5.867	-0.06
cg27523141	chr10	ZNF37BP	0.055	-8.999	-0.12
cg27636376	chr10	C10orf111|RPP38	0.381	-7.108	-0.05

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Simple linear regression coefficient estimate using percent tumor invasion as response variable.

[c]Correlation with percent tumor invasion.

Table 2: Annotations and significance of 39 DNSMI selected Transcriptome elements from **NSM** generated from RDT1.

Ensemble Gene ID	Chromosome	Gene name	p-value[a]	p-value[b]	Estimate[c]	Correlation[d]
ENSG00000057608	chr10	GDI2	0.002	0.445	5e-06	0.19
ENSG00000095787	chr10	WAC	0.037	0.175	-7.4e-05	0.13
ENSG00000099194	chr10	SCD	0	0.092	3e-06	0.22
ENSG00000107771	chr10	CCSER2	0.031	0.587	2.4e-05	0.13
ENSG00000108055	chr10	SMC3	0.004	0.31	3.3e-05	0.18
ENSG00000108094	chr10	CUL2	0.039	0.959	4e-06	0.13
ENSG00000119969	chr10	HELLS	0.033	0.566	-5.8e-05	0.13
ENSG00000136758	chr10	YME1L1	0.025	0.71	-1.1e-05	0.14
ENSG00000138107	chr10	ACTR1A	0.02	0.803	-2e-06	0.14
ENSG00000138160	chr10	KIF11	0.005	0.297	-4.1e-05	0.17
ENSG00000138182	chr10	KIF20B	0.004	0.164	0.000208	0.18
ENSG00000148498	chr10	PARD3	0.007	0.726	-1.3e-05	0.16
ENSG00000148660	chr10	CAMK2G	0.011	0.917	-7e-06	0.15
ENSG00000151461	chr10	UPF2	0.006	0.167	6e-05	0.17
ENSG00000151465	chr10	CDC123	0.044	0.839	-3e-06	0.12
ENSG00000155252	chr10	PI4K2A	0.033	0.345	4.1e-05	0.13
ENSG00000165632	chr10	TAF3	0.011	0.843	1.9e-05	0.15
ENSG00000165637	chr10	VDAC2	0.031	0.914	-1e-06	0.13
ENSG00000165732	chr10	DDX21	0.035	0.039	-3.5e-05	0.13
ENSG00000166135	chr10	HIF1AN	0.041	0.816	-2.1e-05	0.12
ENSG00000170759	chr10	KIF5B	0.015	0.586	-6e-06	0.15
ENSG00000171314	chr10	PGAM1	0.002	0.228	1.9e-05	0.19
ENSG00000172731	chr10	LRRC20	0.001	0.168	3.6e-05	0.2
ENSG00000173848	chr10	NET1	0.011	0.492	4e-06	0.16
ENSG00000176171	chr10	BNIP3	0.001	0.416	8e-06	0.21
ENSG00000180066	chr10	C10orf91	0.007	0.036	0.000144	0.16
ENSG00000181192	chr10	DHTKD1	0.004	0.029	4.8e-05	0.17
ENSG00000181915	chr10	ADO	0.005	0.078	7.6e-05	0.17
ENSG00000187522	chr10	HSPA14	0.002	0.214	7.8e-05	0.19
ENSG00000196072	chr10	BLOC1S2	0.05	0.142	-4.7e-05	0.12
ENSG00000196968	chr10	FUT11	0.001	0.04	8.7e-05	0.21
ENSG00000197771	chr10	MCMBP	0.014	0.232	-5.9e-05	0.15

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Multiple linear regression using percent tumor invasion as response variable and all 39 genes as independent variables.

[c]Multiple linear regression coefficient estimate using percent tumor invasion as response variable and all 39 genes as independent variables.

[d]Correlation with percent tumor invasion.

Ensemble Gene ID	Chromosome	Gene name	p-value[a]	p-value[b]	Estimate[c]	Correlation[d]
ENSG00000198825	chr10	INPP5F	0.003	0.164	0.000155	0.18
ENSG00000213390	chr10	ARHGAP19	0.003	0.603	-6.2e-05	0.18
ENSG00000260917	chr10	AL158212.3	0.01	0.119	0.000361	0.16
ENSG00000108239	chr10	TBC1D12	0.079	0.384	-0.000121	0.11
ENSG00000136738	chr10	STAM	0.083	0.628	-3.2e-05	0.11
ENSG00000165660	chr10	ABRAXAS2	0.076	0.684	-3.6e-05	0.11
ENSG00000173145	chr10	NOC3L	0.089	0.457	7.3e-05	0.1

[a]Simple linear regression using percent tumor invasion as response variable.

[b]Multiple linear regression using percent tumor invasion as response variable and all 39 genes as independent variables.

[c]Multiple linear regression coefficient estimate using percent tumor invasion as response variable and all 39 genes as independent variables.

[d]Correlation with percent tumor invasion.