*Bioinformatics*, 31(15), 2015, 2452–2460
doi: 10.1093/bioinformatics/btv175
Advance Access Publication Date: 26 March 2015
Original Paper



Genome analysis

# DEOD: uncovering dominant effects of cancer-driver genes based on a partial covariance selection method

**Bayarbaatar Amgalan and Hyunju Lee***

School of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Republic of Korea

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on December 16, 2014; revised on March 20, 2015; accepted on March 23, 2015

## Abstract

**Motivation:** The generation of a large volume of cancer genomes has allowed us to identify disease-related alterations more accurately, which is expected to enhance our understanding regarding the mechanism of cancer development. With genomic alterations detected, one challenge is to pinpoint cancer-driver genes that cause functional abnormalities.

**Results:** Here, we propose a method for uncovering the dominant effects of cancer-driver genes (DEOD) based on a partial covariance selection approach. Inspired by a convex optimization technique, it estimates the dominant effects of candidate cancer-driver genes on the expression level changes of their target genes. It constructs a gene network as a directed-weighted graph by integrating DNA copy numbers, single nucleotide mutations and gene expressions from matched tumor samples, and estimates partial covariances between driver genes and their target genes. Then, a scoring function to measure the cancer-driver score for each gene is applied. To test the performance of DEOD, a novel scheme is designed for simulating conditional multivariate normal variables (targets and free genes) given a group of variables (driver genes). When we applied the DEOD method to both the simulated data and breast cancer data, DEOD successfully uncovered driver variables in the simulation data, and identified well-known oncogenes in breast cancer. In addition, two highly ranked genes by DEOD were related to survival time. The copy number amplifications of MYC (8q24.21) and TRPS1 (8q23.3) were closely related to the survival time with *P*-values = 0.00246 and 0.00092, respectively. The results demonstrate that DEOD can efficiently uncover cancer-driver genes.

**Availability and implementation:** DEOD was implemented in Matlab, and source codes and data are available at http://combio.gist.ac.kr/softwares/.

**Contact:** hyunjulee@gist.ac.kr

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

## 1 Introduction

By building systematic knowledge about genetic alterations in cancer, we can enhance our understanding concerning the mechanisms of cancer development and so can identify actionable target genes for cancer treatment. Genomic events in cancer are a mixture of driving events that promote cancer development and passenger events that represent random somatic alterations (Beroukhim *et al.*, 2007). Hence, it is important to distinguish the two events since effective therapies against cancer should target dominant driver genes that promote cell migration and invasion into malignant derivatives

(Danussi *et al.*, 2013). Although many methods based on aberrations in genome sequences, such as copy number changes or mutations, have been proposed to address this challenge, limitations still remain. For example, Beroukhim *et al.* (2007) developed an analytic approach (GISTIC) for identifying significantly aberrant regions of genomes that are frequently found in multiple tumors. However, the identified aberrant regions were often large and contained many neighborhood genes of cancer-driver genes. By incorporating prior knowledge, several frameworks were designed to represent the potential effects of disease-causing alterations. Using eight sequence-based and three structure-based predictive features, a prediction method was developed to estimate the probability of the damaging effect of a missense mutation (Adzhubei *et al.*, 2010). A similar approach (Kumar *et al.*, 2009) was developed to assess the probability that a non-synonymous single nucleotide polymorphism or a single amino acid substitution affects protein functions. A scoring function measuring the effect of a particular mutation was described to cover activation of neighborhoods in a local pathway (Ng *et al.*, 2012). However, following this method, pathway information was used to define the existence of edges in the network so that interactions in only known pathways were considered in the scoring function.

Since disease-causing genetic alterations can be observed from several data types, such as copy numbers, gene expressions or mutations, various statistical approaches to integrate different data types were developed. Bayesian network-based methods such as CONEXIC (Akavia *et al.*, 2010) and Multi-Reg (Danussi *et al.*, 2013) recommend the highest scoring candidate driver within the aberrant regions. For example, in CONEXIC, genes in significant copy number aberrant regions were initially identified as candidate drivers by using the JISTIC method (Sanchez-Garcia *et al.*, 2010), a modified version of GISTIC. Then, for each candidate, a Bayesian scoring function was used to measure the influences of the candidate driver on the genes in its module, which was constructed by the Module Networks algorithm (Segal *et al.*, 2003). In this process, relationships between driver genes and their target genes were calculated based on the expression levels of the genes. Hence, the effects of copy number changes of driver genes on the expression levels of target genes were not fully incorporated.

In this study, we propose a method for uncovering the dominant effects of cancer-driver genes (DEOD) on their target genes by incorporating both chromosomal changes and expression changes. We first describe a statistical graph estimation model to construct a gene network from an integration of matched copy number, gene expression and mutation data. It is formulated as a convex optimization problem, which minimizes a least square error under a sparsity constraint. To measure the dominant effect of each gene throughout the entire network, a scoring function is proposed to compare the downstream consequences of a gene's activity to influences from upstream regulators.

We first tested the performance of our method using the simulated datasets, which are generated based on the principle of a conditional multivariate normal distribution. Then, we applied the proposed method to breast cancer data. From TCGA, we collected genes from three groups, which consist of genes having significant genomic alteration of copy numbers, mutations or expression changes. Then, the method was used to uncover the driver genes that make a dominant contribution to the union of the three types of alterations. Based on the estimation of directional interaction behaviors, the dominant drivers are correctly pinpointed with the highest scores of contributions, which increase our confidence in the prediction of novel drivers. The results demonstrated that our method efficiently uncovers the dominant behaviors of driver genes under the investigated conditions.

## 2 Methods

An overview of the workflow is presented in Figure 1. We first describe a statistical graph estimation model based on a convex optimization technique (Fig. 1A and B). For each gene, it estimates the incoming effects from the other genes, which are edge scores in a network in Figure 1C. We then propose a scoring function to measure the dominant effect of a gene throughout an entire network in Figure 1D.

### 2.1 A partial covariance selection model

The entire network is represented as a directed-weighted graph, $\mathbb{G} = (\mathbb{V}, \mathbb{E})$, where a set of nodes $\mathbb{V}$ represents genes, and a set of edges $\mathbb{E}$ represents the relationships among these genes. We propose a partial covariance selection (PCS) model to construct a gene network as a directed-weighted graph by integrating DNA copy numbers, gene expressions and single nucleotide mutations from the same matched samples, and protein–protein interactions (PPI). For each gene, expression changes of the gene are modeled as a combinational effect of DNA copy number changes and DNA sequence mutations of other genes in the network. We assume that alterations in gene expression are affected by chromosomal aberrations in cancer, because several studies have shown that copy number aberrations often influence the expression of genes via changes in expression of the driver gene (Akavia *et al.*, 2010; Cervigne *et al.*, 2014). Furthermore, expression changes of a particular gene in cancer might not only be inherent from its own copy number or mutation

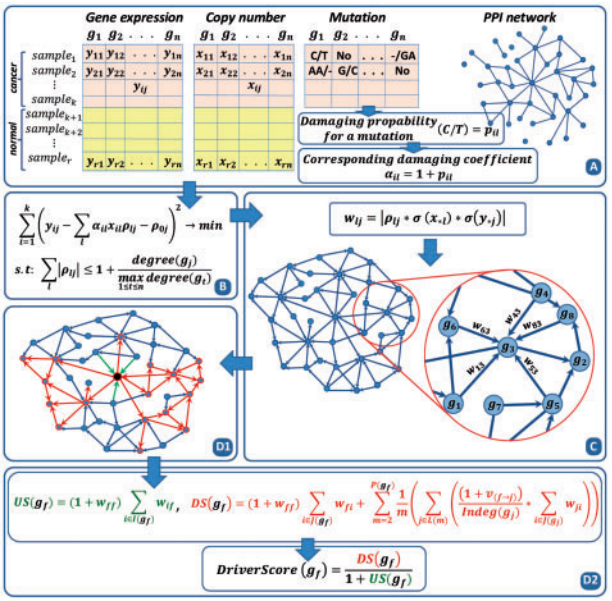


**Fig. 1.** A schematic of our approach. (**A**) An input dataset consists of matched sets of DNA copy numbers, gene expressions and mutations, and PPI. (**B**) A convex optimization problem is formulated to obtain the partial correlation to represent directional influences from regulators to their targets. All genes in the network are considered as candidate targets, and a solution to the optimization problem represents weights of influence from regulators to target genes. (**C**) For each pair of genes *j* and *l* with a non-zero partial correlation coefficient, a partial covariance $w_{lj}$ is expressed as a partial correlation between the copy number change of regulator gene $g_l$ and expression change of the target gene $g_j$. It is embedded as an edge score in the entire network. (**D**) Finally, we describe a driver score for each gene $g_f$ by comparing the downstream consequences (red edges in D1) of a gene's activity to the incoming influences (green edges in D1) from its regulators. The colors of the edges in D1 match with the colors of the equation in D2

status, but also driven by the effects of alteration in copy numbers and mutations of its dependent neighbors. Based on this assumption, we simultaneously estimate the weights of all incoming influences for each gene by solving the following convex optimization problem in Equation (1). In the optimization problem, the expression of each gene can be rewritten as a linear combination of the copy numbers and the mutation status of its incoming neighbors. In other words, although each patient could have a different combination of deletions, amplifications and mutations, the optimization problem assigns more weights to the most relevant combination of copy number changes and mutation effects.

Let $k$ and $n$ denote the numbers of cancer samples and genes, respectively. For the $i$th cancer sample, let $y_{ij}$ denote the gene expression value of gene $g_j$ and $x_{il}$ denote the copy number of gene $g_l$. $\alpha_{il} = 1 + p_{il}$ is a damaging coefficient of mutations (multiple mutations for some samples) occurring in the $i$th sample of gene $l$, where $p_{il}$ is the sum of the probabilities of damaging effects of the mutations and is calculated using a PolyPhen-2 web server (Adzhubei *et al.*, 2010) and a SIFT algorithm (Kumar *et al.*, 2009) that predict the harmful effect of a mutation occurring in DNA sequences. Then, we can write the linear relationships mentioned above as $y_{ij} = \sum_{l=1}^{n} \alpha_{il} x_{il} \rho_{lj} + \rho_{0j}$. Note that $\alpha_{il}$ is multiplied to $x_{il}$. The reason for adding a value of 1 to $p_{il}$ in $\alpha_{il}$ ($\alpha_{il} = 1 + p_{il}$) is that the copy number has still the same value when there is no mutation in the corresponding sample, while a higher aberration value is assigned to the copy number when there is a mutation with a harmful effect. Thus, our goal is to find the minimum of the least square error function subject to an $l_1$ norm constraint on $\rho_{*j}$ in Equation (1).

$$\begin{aligned} \underset{\rho_{*j} \in R^{n+1}}{\text{minimize}} \quad & \sum_{i=1}^{k} \left( y_{ij} - \sum_{l=1}^{n} \alpha_{il} x_{il} \rho_{lj} - \rho_{0j} \right)^2 \\ \text{subject to} \quad & \sum_{l=1}^{n+1} |\rho_{lj}| \leq 1 + \frac{\text{degree}(g_j)}{\max\limits_{1 \leq t \leq n} \text{degree}(g_t)}, \end{aligned} \tag{1}$$

where degree($g_j$) is the degree of gene $g_j$ in the PPI network. The right side in the constraint inequality of Equation (1), ranging from 1 to 2, denotes the expected weight of the total incoming effect from its neighbors to gene $g_j$. For each gene $j$, an $n$-dimensional vector $(\rho_{1j}, \rho_{2j}, \ldots, \rho_{nj})$, obtained by solving the problem in Equation (1), represents the weights of incoming effects from the other genes, and $\rho_{jj}$ represents that expressions of gene $j$ itself can be affected by its own copy number status with mutation effects. $\rho_{0j}$ denotes the intercept adjusting the fitness between random variables. Note that an optimization variable $\rho_{lj}$ is a partial correlation coefficient representing the directional relationship between genes $l$ and $j$. Consider the ordinary correlation $\beta_{12}$ between two random variables $X_1$ and $X_2$. If $X_1$ and $X_2$ are correlated with $n - 2$ other variables $X_3, X_4, \ldots, X_n$, we may regard $\beta_{12}$ as a mixture of a direct correlation between $X_1$ and $X_2$ and an indirect portion due to the presence of other variables correlating with $X_1$ and $X_2$. The partial correlation measuring the direct portion of the total correlation can be defined as a correlation between $X_1$ and $X_2$ after removing effects due to other variables by a linear regression. Therefore, the least square linear regression coefficients are proportional to the partial correlation coefficients (for a detailed description, see Fujikoshi *et al.*, 2010).

An accurate solution to Equation (1) is critical for the robust estimation of relationships among genes in the large-scale network. Although the objective function in Equation (1) is convex, the $l_1$ norm in the sparsity constraint is a non-smooth function and derivative-based optimality conditions such as Lagrangian multipliers and Karush–Kuhn–Tucker (KKT) conditions are not, in general, directly applicable. An extension of function differentiation such as the subdifferentiation might be required due to the non-smoothness. The $l_1$ norm constraint can be decomposed as $2^n$ inequality constraints (Tibshirani, 1996). However, the direct application of the procedure for handling $2^n$ constraints might not be practically useful for the large-scale problem. To overcome this issue, we applied the projected gradient method (Gafni and Bertsekas, 1984) to obtain the optimal solution. A partial covariance between the copy number status of gene $g_l$ and the expression value of gene $g_j$ is defined as

$$w_{lj} = |\rho_{lj} \sigma(x_{*,l}) \sigma(y_{*,j})|, \tag{2}$$

where $\sigma(x_{*,l})$ and $\sigma(y_{*,j})$ are standard deviations of the copy numbers of gene $g_l$ and the expressions of gene $g_j$ across all cancer and normal samples, respectively. Note that the normal samples are used to measure changes of each gene between normal and cancer conditions (see Supplementary Fig. S1). The partial covariance $w_{lj}$ incorporates three statistical measurements, a directional correlation $\rho_{lj}$ from a regulator $g_l$ to its target $g_j$, copy number changes of regulator $g_l$ and expression changes of target $g_j$, and it is used as the edge score in the network.

## 2.2 A scoring function for cancer-driver genes

The alteration of a gene may change cell physiology by directly or indirectly activating transcriptional cascades involving transcription factors, master regulators and signaling proteins. The local invasion might be spread out to normal tissues surrounding the tumor, resulting in cell proliferation, migration and differentiation (Akavia *et al.*, 2010; Danussi *et al.*, 2013).

In this study, we propose a scoring function to measure the potential effect of driver genes throughout the entire network. By comparing the downstream effects of a gene's activity to the influences from its upstream regulators, the cancer-driver score of gene $g_f$ is defined as

$$\text{DriverScore}(g_f) = \begin{cases} \dfrac{DS(g_f)}{1 + US(g_f)}, & E_{\text{out}}(g_f) > M + E_{\text{in}}(g_f) \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $E_{\text{out}}(g_f) = \sum_{i \in J(g_f)} w_{fi}$ and $E_{\text{in}}(g_f) = \sum_{i \in I(g_f)} w_{if}$ denote the total direct outgoing and incoming edge scores for the gene $g_f$, respectively, and where $I(g_f)$ and $J(g_f)$ denote index sets for the incoming and outgoing neighbors of $g_f$, respectively. $M = \frac{1}{n} \sum_{i=1}^{n} E_{\text{in}}(g_i)$ denotes the mean value of $E_{\text{in}}(g_i)$ of all $n$ genes. It implies that $g_f$ is more likely a dominator if its outgoing effect is greater than the sum of its incoming effect and the mean value of the incoming effects of all genes. US($g_f$) describes the effects on $g_f$ from its local upstream regulators and DS($g_f$) describes the effect on downstream targets driven by the regulator $g_f$. We assume that each regulator can be activated by the direct neighbors on its upstream cascades, and then produces carry-over effects on its direct and indirect targets on the downstream of the regulator. US($g_f$) is defined as

$$\text{US}(g_f) = (1 + w_{ff}) \sum_{i \in I(g_f)} w_{if},$$

where $w_{ff}$ denotes a partial covariance representing the effect of the copy number status on its own expression values, and in the network, it is a self-loop that can be either an incoming or an outgoing edge for $g_f$. Therefore, $w_{ff}$ is used to emphasize the effect of the copy numbers to its own gene expressions in both US($g_f$) and DS($g_f$)

(for more explanation, see Supplementary Text and Fig. S2). $\mathrm{DS}(g_f)$ is defined as

$$\mathrm{DS}(g_f) = (1 + w_{ff}) \sum_{i \in J(g_f)} w_{fi} + \sum_{m=2}^{P(g_f)} \frac{1}{m} \left( \sum_{j \in L(m)} \left( \frac{(1 + v_{(f \to j)})}{\mathrm{Indeg}(g_j)} \sum_{i \in J(g_j)} w_{ji} \right) \right),$$

where $P(g_f)$ is the longest distance from $g_f$ to its target genes in the network, and $L(m)$ is the index set of downstream target genes of $g_f$, which are located $m$ distance from the gene $g_f$. $v_{(f \to j)} = \sum_{l \in I_A(g_j)} w_{lj}$ is the total influence to gene $g_j$ from its direct incoming neighbors and it describes how strong $g_j$ is affected by the regulator $g_f$ via its intermediate targets on the paths from $g_f$ to $g_j$. If the affection score is high, the outgoing influences of $g_j$ on its downstream targets would be more valuable, where $I_A(g_j)$ is an index set of direct-active incoming edges into gene $g_j$, which are on the paths from $g_f$ to $g_j$. $\mathrm{Indeg}(g_j)$ is the number of all incoming edges into gene $g_j$. Note that if an intermediate target gene $g_j$ is located $m$ distance from its focusing regulator $g_f$, then outgoing effect of $g_j$ should be scaled with $\frac{1}{m}$, which means that target genes longer distances from $g_f$ have less contributions to the accumulation of $\mathrm{DS}(g_f)$.

To select candidate-driver genes, we estimate a threshold value of $\mathrm{DriverScore}(g_f)$ based on a $P$-value obtained by comparing the observed network with random networks. A random network is generated from a random copy number matrix and gene expression matrix, where the copy number or expression values are randomly permuted in each matrix. The null hypothesis is that gene expression values are independent of copy numbers and mutations. By taking the ratio of random networks ($N_{\mathrm{max>obs}}$), in which the maximum driver score from the random network is larger than the observed driver score, to the total number of random networks ($N_{\mathrm{all}}$), $P$-value is calculated as ($N_{\mathrm{max>obs}}/N_{\mathrm{all}}$).

## 3 Results

### 3.1 Simulation study

Suppose that we have an $n$-dimensional multivariate normal distributed random variable $Y = (y_1, y_2, \ldots, y_n)^T \sim \mathcal{N}(\mu, \Sigma)$. Consider partitioning $Y = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ into $x_1 = (y_1, y_2, \ldots, y_p)^T$ and $x_2 = (y_{p+1}, y_{p+2}, \ldots, y_n)^T$ with a similar partition of a mean vector and a covariance matrix $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then, $(x_1 | x_2 = s)$, the conditional distribution of the first partition given the second, is also multivariate normal $\mathcal{N}(\overline{\mu}, \overline{\Sigma})$, with mean

$$\overline{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(s - \mu_2)$$

and covariance matrix

$$\overline{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

For a comprehensive review of the conditional distribution, see Johnson and Wichern (2007). Assuming the conditional distribution of the partitioned multivariate normal random variables, we constructed six simulation datasets: five case datasets (cancer samples) with different covariance matrices and one reference dataset (healthy samples). Each dataset represents matched samples of copy numbers with mutation effects and gene expressions. Each dataset consists of 500 genes (variables) with 100 samples. For each gene $i$, we draw the mean $\mu_i$ and the standard deviation $\sigma_i$ from the uniform distributions on the observed range of normal data $[-0.5, 0.5]$, and then calculate the correlation coefficient $\rho_{ij}$ for each pair of genes $(i, j)$. Let $\mu = (\mu_1, \mu_2, \ldots, \mu_{500})^T$ denote the mean vector, and $\Sigma = \{\sigma_{ij}\}_{i,j=1,\ldots,500}$ denote the covariance matrix, where $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ is the covariance between genes $i$ and $j$. The reference data were simulated from joint normal distribution $\mathcal{N}(\mu, \Sigma)$. For each of the case datasets, 10 of 500 genes were selected as driver genes and 200 of the remaining 490 genes were selected as target genes. For each of the 10 drivers, 20–35 targets were randomly selected from the 200 target genes. For the selected 210 (drivers and targets) genes, mean values were shifted. For correlations, significantly higher correlations between each driver and its own targets were assigned to represent interactions between the driver gene and its targets, and slightly higher correlations were assigned among the 10 driver genes and among the 200 target genes to represent alternations on the part of the entire network. In the covariance matrix of the illustration example in Figure 2A and B, correlations between each driver and its own targets are colored in blue, and correlations among drivers and among targets are colored in red and magenta, respectively. Let $\mathbb{I}_{\mathrm{d}}$ and $\mathbb{I}_{\mathrm{t}}$ denote sets of indexes for the driver genes and the target genes, respectively.

$$\hat{\mu}_i = \begin{cases} \mu_i + a, & i \in \mathbb{I}_{\mathrm{d}} \cup \mathbb{I}_{\mathrm{t}} \\ \mu_i, & i \notin \mathbb{I}_{\mathrm{d}} \cup \mathbb{I}_{\mathrm{t}}, \end{cases}$$

$$\hat{\rho}_{ij} = \begin{cases} \rho_{ij} + b, & i, j \in \mathbb{I}_{\mathrm{d}} \text{ or } i, j \in \mathbb{I}_{\mathrm{t}} \\ \rho_{ij} + c, & (i \in \mathbb{I}_d \text{ and } j \in \mathbb{I}_{\mathrm{t}}) \text{ or } (j \in \mathbb{I}_{\mathrm{d}} \text{ and } i \in \mathbb{I}_{\mathrm{t}}) \\ \rho_{ij}, & \text{otherwise.} \end{cases}$$

When we generate samples for the five case datasets, $a = 0.3$, $b = 0.3$ and $c = 0.4$, 0.5, 0.6, 0.7 and 0.8 were used to obtain the mean vector $\hat{\mu}$ and the correlation matrix $\hat{\rho}$. In this process, we attempted to restrict the effects from targets to drivers. Hence, we first simulate samples for 10 driver genes $y_{491}, \ldots, y_{500}$ from the multivariate normal distribution, and then simulate samples of other genes from the

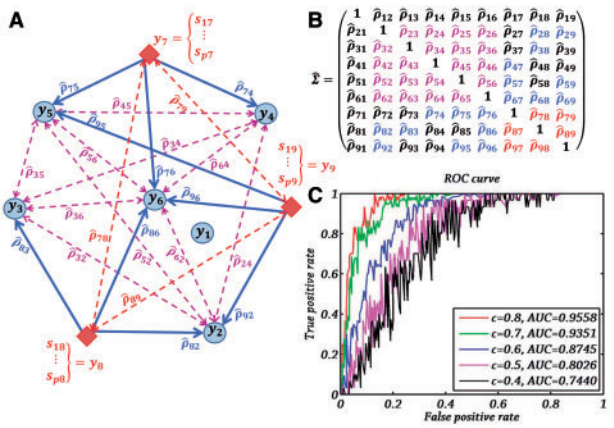


**Fig. 2.** Simulation of the DEOD method. (**A**) Relationships among driver genes and target genes are represented in a network. Squares in red represent driver genes, while circles in gray represent target genes or free genes. Blue, red-dotted and magenta-dotted edges, respectively, represent the directed relationships from drivers to their own targets, correlation between drivers and correlation between targets. (**B**) A correlation matrix for the network in (A) is shown. The highlighted colors in the covariance matrix are matched with the edge colors in (A). (**C**) Performances of the DEOD method from the five simulated datasets with different covariance matrices are shown

conditional multivariate normal distribution depending on the given drivers.

Let $\hat{\Sigma} = \{\hat{\sigma}_{ij}\}_{i,j=1,\dots,500}$ denote the covariance matrix, where $\hat{\sigma}_{ij} = \hat{\rho}_{ij}\sigma_i\sigma_j$ is the covariance between genes $i$ and $j$, and $S$ is a $100 \times 500$ matrix whose entry $s_{lj}$ represents the value of the $l$th sample of gene $j$. Then, the samples of the 10 driver genes $y_{491}, \dots, y_{500}$ were first simulated as

$$s_{(1\dots100),(491\dots500)} \Leftarrow \text{Simulate}(\mathcal{N}(\hat{\mu}_{(10)}, \hat{\Sigma}_{(10\times10)})).$$

Assuming the conditional multivariate normal distribution of the 490 genes given the 10 drivers, we simulated the samples for the 490 genes from multivariate normal distribution $\mathcal{N}(\overline{\mu}, \overline{\Sigma})$ with mean

$$\overline{\mu}_{(490)} = \hat{\mu}_{(490)} + \hat{\Sigma}_{(490\times10)}\hat{\Sigma}^{-1}_{(10\times10)}(s - \hat{\mu}_{(10)})$$

and covariance matrix

$$\overline{\Sigma}_{(490\times490)} = \hat{\Sigma}_{(490\times490)} - \hat{\Sigma}_{(490\times10)}\hat{\Sigma}^{-1}_{(10\times10)}\hat{\Sigma}_{(10\times490)},$$

where $\overline{\mu}$ is a 490-dimensional vector and $\overline{\Sigma}$ is a $490 \times 490$ matrix. For each sample $l$, we finally simulated cancer data for the 490 genes $y_1, y_2, \dots, y_{490}$ as

$$s_{(l),(1\dots490)} \Leftarrow \text{Simulate}(\mathcal{N}(\overline{\mu}(s_{(l),(490\dots500)}), \overline{\Sigma})).$$

Figure 2A shows an example for simulating nine genes. Let $Y = (y_1, y_2, \dots, y_9)^T$ be a multivariate normal random variable with a mean vector $\hat{\mu}$ and a covariance matrix $\hat{\Sigma}$, where the variances are fixed at 1. The covariance matrix is then the same as the correlation matrix given in Figure 2B, and the conditional distribution of $x_1 = (y_1, y_2, \dots, y_6)^T$ given $x_2 = (y_7 = s_{l7}, y_8 = s_{l8}, \dots, y_9 = s_{l9})^T$ is also multivariate normal with mean

$$\overline{\mu} = \hat{\mu}_{(1\dots6)} + \hat{\Sigma}_{(1\dots6),(7\dots9)}\hat{\Sigma}^{-1}_{(7\dots9),(7\dots9)}(s - \hat{\mu}_{(7\dots9)})$$

and covariance matrix

$$\overline{\Sigma} = \hat{\Sigma}_{(1\dots6),(1\dots6)} - \hat{\Sigma}_{(1\dots6),(7\dots9)}\hat{\Sigma}^{-1}_{(7\dots9),(7\dots9)}\hat{\Sigma}_{(7\dots9),(1\dots6)}$$

where $\overline{\mu}$ is a 6-dimensional vector and $\overline{\Sigma}$ is a $6 \times 6$ matrix.

In Figure 2A, if the highlighted part of the entire network represents a complete graph and the correlation coefficients for all pairs in that part are similar enough, then the covariance matrix tends to be symmetric positive semi-definite and the simulation process can be easily constructed without any theoretical complications. Unfortunately, the significant part forms an incomplete graph structure and correlations between the drivers and their own targets are sufficiently higher than the correlations either between targets or between drivers. In addition, drivers may share only some of their targets (in Fig. 2A, driver $y_8$ is connected to its targets $y_2$, $y_3$ and $y_6$ while it is not connected to $y_1$, $y_4$ and $y_5$). Due to this kind of randomness, the covariance matrix itself is neither positive definite nor positive semi-definite. Hence, we need to find the nearest positive definite matrix of the covariance matrix.

The problem of finding the nearest positive definite matrix can be formulated as the following optimization problem

$$\varepsilon(\hat{\Sigma}) = \underset{X - X^T \geq 0}{\text{minimize}} \ ||\hat{\Sigma} - X||.$$

In other words, any positive definite $X$ satisfying $||\hat{\Sigma} - X|| = \varepsilon(\hat{\Sigma})$ can be a positive approximation of $\hat{\Sigma}$ in the given norm. We applied the following analytic result from Higham (1988) that gives the solution to the problem of positive approximation in the Frobenius norm. Let $\hat{\Sigma} \in R^{n\times n}$ be an arbitrary matrix, its symmetric and skew-symmetric parts be $A = (\hat{\Sigma} + \hat{\Sigma}^T)/2$ and $B = (\hat{\Sigma} - \hat{\Sigma}^T)/2$, respectively, and $A = UP$ be a polar decomposition. Then, according to the theorem in Higham (1988), $X^*_F = (A + P)/2$ is the unique positive approximation of $\hat{\Sigma}$ in the Frobenius norm, and the approximation error is estimated as

$$\varepsilon_F(\hat{\Sigma})^2 = \sum_{\lambda_i(A)<0} \lambda_i(A)^2 + ||B||^2_F,$$

where $A = Y\Lambda Y^T$ is a spectral decomposition of $A$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n)$ and $Y^TY = I$. The performances of the DEOD method on the simulated datasets with 500 genes are summarized in Figure 2C. When five different covariance matrices were used, the prediction accuracies varied. An area under the curve (AUC) value of the true positives and false positive rates was highest when $c = 0.8$. Although the accuracy decreases when the covariance decreases, DEOD still gives a high AUC value for even $c = 0.4$, demonstrating that DEOD can successfully uncover a large fraction of driver genes.

## 3.2 Analysis on breast cancer data

### 3.2.1 Datasets

The matched copy number, gene expression and mutation datasets of the breast cancer samples were downloaded from the TCGA data portal (http://cancergenome.nih.gov/). We used 506 cancer samples without missing values in both copy numbers and gene expressions, and 58 normal samples. We collected 11 852 genes in the union of the following three datasets.

- A total of 2592 genes located in copy number aberrant regions were obtained by the GISTIC method (Beroukhim *et al.*, 2007). The copy number value of a gene in a sample was determined using the hg19 build of the genome (Kent *et al.*, 2002).
- Differentially expressed 649 genes forming the condition-specific subnetwork were obtained by applying our previously developed method, WMAXC (Amgalan and Lee, 2014).
- A total of 10 895 genes containing somatic mutations in at least one of the 506 cancer samples were selected. In 506 samples, a total of 40 978 mutations occurred.

In terms of PPI data, we downloaded the Human Protein Reference Database released in 2010 (Prasad *et al.*, 2009), which contains 39 240 PPI from 9617 genes. We used 21 350 interactions among the above 11 852 genes.

### 3.2.2 Driver genes of breast cancer

When we applied the DEOD method to the breast cancer data, 186 genes with a $P$-value $< 0.05$ (the driver score $> 18.09$) were selected as candidate-driver genes. The $P$-value was obtained by comparing with 30 random networks. A list of the 186 genes is shown in Supplementary Table S1. The 10 highest scoring drivers are shown in Table 1, and all 10 genes are previously known oncogenes in breast cancer or in other cancer types, or their cellular changes are related to cancer. MYC, MDM4, TRPS1, ZNF217, PAX5 and BANP are known to contribute to breast cancer progression, and ADAMTSL4, ORAOV1 and AVPR1B have functional roles in other cancers or cancer-related biological processes. Also, F-box only protein 31 (FBXO31) is suggested as a candidate tumor suppressor gene (Kumar *et al.*, 2005).

For each of the 186 candidate cancer-driver genes, gene $g_j$ is selected as its target gene if the partial covariance $w_{lj}$ between the driver gene $g_l$ and the gene $g_j$ is $> 0.05$ in Equation (2). The numbers of target genes for 186 candidate-driver genes are shown in

**Table 1.** The 10 highest scoring driver genes identified by DEOD

| Gene symbol | Score | Aberration | Location | References |
|---|---|---|---|---|
| MYC | 1516.9 | Amplified | 8q24.21 | Xu *et al.* (2010) |
| MDM4 | 344.08 | Amplified | 1q32.1 | Duffy *et al.* (2014) |
| TRPS1 | 106.13 | Amplified | 8q23.3 | Wu *et al.* (2014) |
| ADAMTSL4 | 80.459 | Amplified | 1q21.3 | Le Goff *et al.* (2011) |
| FBXO31 | 63.289 | Deleted | 16q24.2 | Kumar *et al.* (2005) |
| ORAOV1 | 47.893 | Deleted | 11q13.3 | Katoh *et al.* (2005) |
| ZNF217 | 40.840 | Amplified | 20q13.2 | Nguyen *et al.* (2014) |
| PAX5 | 39.904 | Deleted | 9q13.2 | Moelans *et al.* (2011) |
| BANP | 39.515 | Deleted | 16q24.2 | Malonia *et al.* (2011) |
| AVPR1B | 39.144 | Amplified | 1q32.1 | Savas *et al.* (2012) |

Top 10 genes are listed with scores by DEOD, copy number status and cytobands. They are previously known to be related to breast cancer or other cancer types, and supporting literatures are shown in the last column.

Supplementary Table S1, which varies from 11 to 4529 with the mean number = 242. To check whether the target genes of a driver gene are collaboratively working for particular biological functions or pathways, we performed a functional enrichment test for the target genes of each drivers. We applied a hypergeometric test followed by a Benjamin–Hochberg test for multiple comparison corrections using KEGG pathways (Kanehisa *et al.*, 2000) and Gene Ontology (GO) terms (Carbon *et al.*, 2009). Target genes of 184 out of 186 genes (98.9%) were enriched with at least one KEGG pathway or one GO term. The average number of enriched terms was 24. The numbers of enriched terms are shown in Supplementary Figure S3, and the list of enriched KEGG pathways for the ten highest scoring drivers is shown in Supplementary Table S2. The enriched terms include cancer-related pathways such as ECM–receptor interaction, P53 signaling pathway and TGF-$\beta$ signaling pathway. This result implies that the target genes of the candidate-driver genes consist of functionally related genes, and the identified driver genes might play significant roles in the dysregulation of cancer-related pathways.

In addition, we checked the pathways in which 186 drivers were enriched. As shown in Supplementary Table S3, 15 KEGG pathways and four GO terms were enriched, including the ERBB signaling pathway, a well-known breast cancer-related pathway. Interestingly, enriched pathways include pathways of other cancer types such as non-small cell lung cancer, glioma, melanoma, bladder cancer, pathways in cancer, small cell lung cancer and prostate cancer.

### 3.3 MYC deregulation in breast cancer

DEOD identified MYC, a potent activator of tumorigenesis, as the highest-scoring driver gene. The driver score of MYC was significantly higher compared with the second ranked gene. MYC is a transcription factor and a key regulator of cell growth, proliferation, metabolism, differentiation and apoptosis (Xu *et al.*, 2010). Its deregulation has been found in many cancer types including breast cancer (Liu *et al.*, 2012; Xu *et al.*, 2010). Both genomic and functional analyses of MYC responsive genes suggest that MYC regulates up to 15% of all humans genes (Chen and Olopade, 2008).

Our analysis showed that the chromosomal region of MYC was significantly amplified and that the number of MYC downstream targets is relatively large. 1098 targets are directly affected by MYC. Figure 3 shows the relationships between the copy number status of MYC and the expressions of its direct targets. The MYC target network is enriched with several GO terms such as cell differentiation (*P*-value = 4.427E−21), regulation of cell proliferation

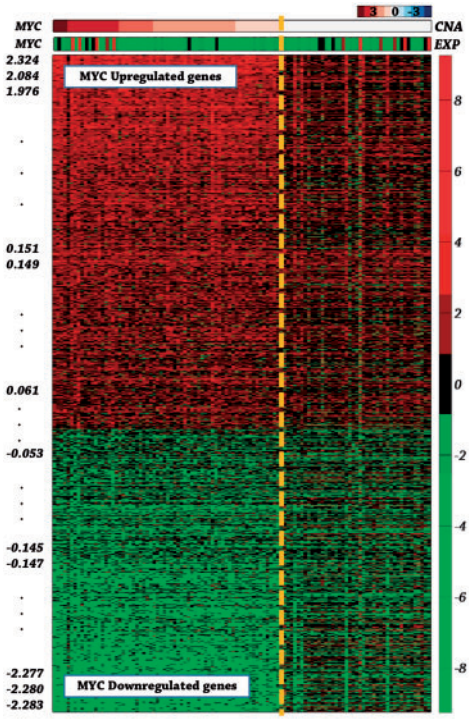


**Fig. 3.** MYC and its target genes. 1098 direct target genes are ordered according to partial covariances ($-2.283 \leq \text{sign}(\rho_{fi}) w_{fi} \leq 2.324$, where *f* and *i* denote the indexes of the driver and its targets) between the copy numbers and mutation effects of MYC and expression levels of its targets, whereas the samples are ordered according to the copy numbers of MYC

(*P*-value = 4.976E − 22), cell migration (*P*-value = 4.986E − 13) and cell adhesion (*P*-value = 3.910E − 19). In addition, 44 KEGG pathways were enriched, including well-known breast cancer-related pathways such as complement and coagulation cascades (*q*-value < 1.0E − 31), focal adhesion (*q*-value = 3.76E − 05) (Zhang and Chen, 2010), cytokine–receptor interaction (*q*-value = 1.14E − 05) (Huan *et al.*, 2013), ECM–receptor interaction (*q*-value = 4.64E − 05) (Krupp *et al.*, 2011) and TGF-$\beta$ signaling pathway (*q*-value = 1.51E − 04) (Scollen *et al.*, 2011). A complete list of enriched pathways from the KEGG pathways is shown in Supplementary Table S2.

We further analyzed the relationship between the copy number changes of MYC and the survival time. Clinical information on 1040 breast cancer patients was downloaded from the TCGA data portal on October 20, 2014. We first estimated the most aberrant 1% of all copy number values in cancer samples (0.840 for amplification and − 0.707 for deletion), and used them as threshold values to decide samples with copy number aberrations. Then, for each candidate-driver gene, the survival time of a group of patients whose copy number status passes the threshold was compared with that of a group of the same number of patients whose copy number status was on the opposite of the aberrant group. Survival curves were estimated by a Kaplan–Meier analysis at a series of time points of days, and the difference between two survival plots was evaluated by a log rank test. The mean survival time of the high amplified group (158 patients) is shorter (3350 days) than that of the low amplified group (4354 days) with *P*-value = 0.00246, which implies that amplification of MYC copy number is significantly related to the survival time. When we performed the survival time analysis for the other nine candidate-driver genes in Table 1, a similar relationship was

found in the amplifications of TRPS1 ($P$-value = 0.00092), which is the third highest scoring dominator in our result and a well-known oncogene playing an important role in the control of cell cycle and proliferation during breast cancer development (Wu *et al.*, 2014). The Kaplan–Meier curves for these two genes are shown in Figure 4. The survival analysis of all the 186 candidate-driver genes was shown in Supplementary Table S4. Out of 186 genes, the copy numbers of 26 genes (16%) were related to the survival time with significant $P$-values. When the threshold for copy number aberration was estimated based on 5% and 3%, 37 genes (19.8%) and 20 genes (10.7%) were related to the survival time, respectively (Supplementary Table S5).

### 3.4 FBXO31 is a candidate tumor suppressor gene

In addition to the well-known oncogenes, FBXO31 was identified as a high-scoring dominator. A previous study hypothesized that FBXO31 is a candidate tumor suppressor in breast cancer, which induces cellular senescence and has consistent properties of tumor suppressors by generating Skp Cullin F-box containing complex (SCF complex) (Kumar *et al.*, 2005). To support the above hypothesis, we further investigated FBXO31.

Our analysis shows that in many cancer samples, the chromosomal regions of FBXO31 were significantly deleted and FBXO31 was down-regulated, and copy numbers were related with expression levels ($w_{ff} = 0.1231$). A target network of FBXO31 includes 740 positively related genes and 788 negatively related genes. Figure 5 displays the relationships between the copy number status of FBXO31 and the expressions of its direct targets. Target genes of FBXO31 were significantly enriched with cancer-related GO terms such as cell differentiation ($P = 1.04\text{E}-16$), cell development ($P = 6.60\text{E}-09$), cell adhesion ($P = 2.77\text{E}-05$), cell migration ($P = 8.77\text{E}-05$) and regulation of cell proliferation ($P = 1.084\text{E}-12$). Also, 13 KEGG pathways were enriched, including breast cancer-related pathways such as p52 signaling and primary immunodeficiency. A complete list of the pathways is given in Supplementary Table S2. In conclusion, our findings support the notion that FBXO31 is a candidate tumor suppressor gene in breast cancer.

### 3.5 Analysis on brain cancer

We compared DEOD with a Multi-Reg (Danussi *et al.*, 2013) method, which is a modified version of CONEXIC (Akavia *et al.*, 2010) and was developed by the same research group. We downloaded the 242 matched copy number (HG-CGH-244A CN Array), gene expression (HT HG-U133A or Agilent G4502A-07) and mutation datasets of glioblastoma multiforme (GBM) samples from the TCGA data portal, which were previously used in Multi-Reg. In Danussi *et al.* (2013), using Multi-Reg, 83 genes were recommended as GBM-driver genes. We used 191 cancer samples without missing values in both copy numbers and gene expressions. Chromosomal locations of genes were determined using the hg18 build of the genome. For the analysis, we used 8362 genes included in the PPI network with 31 485 interactions. When we applied DEOD to GBM data, 48 genes were selected as candidate GBM cancer-driver genes (driver scores > 49.01 and $P$-value < 0.05). A complete list of the 48 genes is given in Supplementary Table S6. The top 10 ranked genes include well-known oncogenes and tumor suppressors of GBM such as EGFR, CDKN2A, PTEN, PDGFRA and CDKN2B. For comparison with Multi-Reg, we applied the functional enrichment test of GO terms and KEGG pathways to 48 genes identified by DEOD and 83 genes identified by Multi-Reg. As a result, 107 and 87 pathways were enriched for DEOD and Multi-Reg, respectively. The enriched terms include many cancer-related terms such as glioma and p53 signaling pathways (Supplementary Tables S7 and S8), showing that both methods recommend GBM-related genes.

In Danussi *et al.* (2013), the target genes of the identified drivers were compared against the gene expression signatures of GBM subtypes in order to show functional properties of candidate-driver genes. We performed the same analysis with 48 candidate-driver genes identified by DEOD. For this task, we obtained three groups of genes whose expression signatures were associated with mesenchymal, proneural and proliferative GBM subtypes (Carro *et al.*, 2010). Among the 140 mesenchymal, 242 proneural and 181 proliferative genes obtained, 92, 137 and 103 genes were included in the set of 8362 genes. The candidate drivers selected by Multi-Reg and DEOD were, respectively, classified into the three molecular signature groups based on the enrichment of their target genes in the subtype-related genes using a hypergeometric test ($q$-value < 0.05). When the driver gene was related with multiple subtypes, the

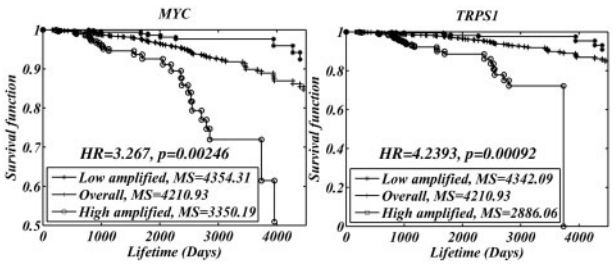


**Fig. 4.** Survival time analysis depending on the copy number changes of MYC and TRPS1. 'HR' represents a hazard ratio measuring the survival time difference between low- and high-amplified groups and 'p' represents a $P$-value of the statistical significance of the survival time difference. 'MS' denotes the mean survival in days

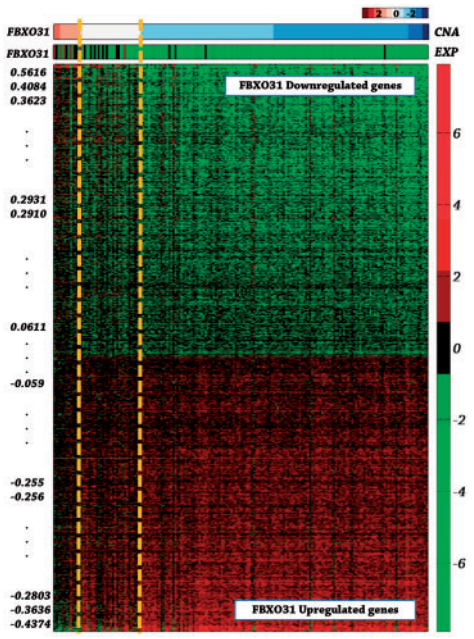


**Fig. 5.** FBXO31 and its target genes. 1528 direct target genes are ordered according to partial covariances ($-0.437 \leq \text{sign}(\rho_{fj}) w_{fj} \leq 0.561$, where $f$ and $j$ denote the indexes of the driver and its targets) between the copy numbers and mutation effects of FBXO31 and expression levels of its targets, whereas the samples are ordered according to the copy numbers of FBXO31

**Table 2.** Comparison on the performances of the two methods on GBM

| Methods | Candidate genes | Mesenchymal | Proneural | Proliferation | Overall |
|---|---|---|---|---|---|
| Multi-Reg | 83 | 23 | 11 | 14 | 48 (58%) |
| DEOD | 48 | 24 | 5 | 8 | 37 (77%) |

The candidate drivers identified by Multi-Reg and DEOD were classified into the three molecular signature GBM subtypes: mesenchymal, proneural and proliferation.

subtype with the minimal $q$-value was selected. Table 2 shows the numbers of candidate drivers associated with the subtypes; 77 and 58% of candidate-driver genes of DEOD and Multi-Reg, respectively, were related to subtypes.

In addition, we analyzed the relationships between the survival time and the copy number changes of candidate-driver genes identified by DEOD and MultiReg. Copy number data (Genome-Wide Human SNP Array 6.0) and clinical information of 523 GBM patients were downloaded from the TCGA data portal on March 12, 2015. Similar to the survival time analysis on breast cancer data, the most aberrant 1% copy numbers of all copy number values in the GBM data were selected as the threshold values (0.5093 for amplification and − 0.8576 for deletion). We found that the copy number changes of 13 out of 48 driver genes (27.1%) selected by DEOD and 14 out of 83 driver genes (16.8%) selected by Multi-Reg were related to the survival time with significant $P$-values. (See Supplementary Table S9 for a list of genes, $P$-values, and the number of samples with copy number aberrations.) Also, fractions of candidate-driver genes related to the survival time were shown in Supplementary Table S10 for the different thresholds of 5%, 3% and 1%. Across the three thresholds, higher fractions of candidate-driver genes identified by DEOD were related to the survival time than those by MultiReg.

## 4 Discussion

The main advantages of DEOD are as follows: Based on an optimal combination of copy number deletions and amplifications with mutation effects, DEOD estimates a large-scale graph by integrating different data types. Different from other methods, DEOD measures genetic alterations and directional relationships between genes across different data types (copy number aberration with mutation effects of regulators to expression change of their targets). A convex minimization method with a strong theoretical validation of optimality, projected gradient, was implemented to select incoming edges of each gene as a partial correlation effect from the other genes in the entire network. A partial covariance is then taken into account as a combination of three statistical measurements, the partial correlation from a regulator to its target, copy number changes of the regulator, and differential expressions of the target to represent edge scores in the network. For each gene, the scoring function measures all effects of the candidate-driver gene on the entire network by comparing the accumulation of contributions in downstream cascades to the total direct incoming effect from its upstream.

We further investigated the effect of mutations on gene expressions. For this task, we ran the complete procedure on the breast cancer data after removing the damaging coefficients of mutations, $\alpha_{ij}$, in Equation (1). In this modified experiment, 216 genes were selected as candidate-driver genes with the threshold score of 10.8609, and the driver scores of several genes with mutations in the multiple samples were significantly decreased. For example, the driver score of USH2A decreased from 24.2752 to 1.107 because 27 samples were mutated in this gene with a high probability of damaging effects. The numbers of samples with mutations in the driver genes are shown in Supplementary Table S1. Also, 47 genes were removed from the original list of drivers and are listed in Supplementary Table S11. When we manually checked the 10 genes in the list from literature in PubMed, they were previously known to be related to breast cancer, showing that mutation information is another important source in finding driver genes.

Similarly, we investigated the effect of the PPI network. The binary interactions in the PPI network were used to describe the expected total incoming effect of each gene in the network. In Equation (1), genes that have high degrees in the PPI network would be expected to have more total incoming effects due to the inequality constraint. We ran the DEOD method on breast cancer without PPI. After excluding the PPI information, 32 candidate drivers whose driver scores were less than the threshold score of 10.2955 were removed from the original list of drivers and are listed in Supplementary Table S12. A comparison statistic for the incoming effects of genes is given in Supplementary Table S13. The incoming effects and degrees of the genes were higher when the PPI information was included than when the PPI information was excluded. This result shows that integration of the PPI information in DEOD successfully incorporated the observation that cancer genes have a higher degree in the PPI network (Jonsson and Bates, 2006).

Our results showed that only some of candidate-driver genes were related to the survival time in both breast cancer and GBM. For example, although PTEN is a well-known GBM-driver gene and the deletion of PTEN is one of the hallmarks in GBM development, the relationship between the copy numbers and the survival time was not significant in our study ($P$-value = 0.0596) and its prognostic significance still remains controversial (Carico *et al.*, 2012; Xu *et al.*, 2014). Also, the statistical significance of the survival analysis depends on the thresholds for copy number aberrations, as shown in Supplementary Tables S5 and S10. For example, in breast cancer, the copy numbers of FBXO31 were related to the survival time when the threshold was chosen based on 5% of all the copy numbers ($P$-value = 0.04375, the number of samples with aberrations = 262), while its $P$-value was not significant when 1% was used ($P$-value = 0.22730, the number of samples with aberrations = 61). There might be several complex relationships between genes and the survival time. One possible explanation might be the combined effects of a gene and other molecules (Gross *et al.*, 2014) and another reason might be the chromosomal positions of aberrations or the role of aberrant domains within the gene (Minaguchi *et al.*, 2001). More analyses are required to reveal the complex relationships between genes and the survival time.

In this study, we explained the activities of genes based on copy number changes and mutations, although other molecules or genetic changes such as microRNAs, transcription factors and methylations also play significant roles in cancer development. Hence, the identified driver genes in this analysis may represent a partial list of those genes that drive changes of cellular activities in cancer. Indeed, the PCS method can be used to measure relationships between any explanatory factors and response variables. Hence, it is possible to extend the DEOD method so that driver genes can be identified from several different types of molecular changes. The extension of the DEOD method will be our future work to enhance the understanding of cancer development.

In addition to the DEOD method, we designed a novel data simulation scheme based on conditional multivariate normal random variables given a group of specific variables. Compared with a simulation scheme used in our previous work (Amgalan and Lee, 2014), the novel scheme has the advantages of simulating incomplete-directed subgraphs in an entire background graph.

## Funding

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2013R1A1A2058053) and by the 'Systems biology infrastructure establishment grant' provided by Gwangju Institute of Science and Technology in 2015.

*Conflict of Interest:* none declared.

## References

Adzhubei,I.E. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Akavia,U.D. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Amgalan,B. and Lee,H. (2014) WMAXC: a weighted maximum clique method for identifying condition-specific sub-network. *PLoS One*, **9**, e104993.

Beroukhim,R. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and applications in cancer. *Proc. Natl. Acad. Sci. USA*, **104**, 20007–200012.

Carbon,S. *et al.* (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.

Carico,C. *et al.* (2012) Loss of PTEN is not associated with poor survival in newly diagnosed glioblastoma patients of the temozolomide era. *PLoS One*, **7**, e33684.

Carro,M.S. *et al.* (2010) The transcriptional network for messenchymal transformation of brain tumours *Nature*, **463**, 318–325.

Cervigne,N.K. *et al.* (2014) Recurrent genomic alterations in sequential progressive leukoplakia and oral cancer: drivers of oral tumorigenesis? *Hum. Mol. Genet.*, **23**, 2618–2628.

Chen,Y. and Olopade,O.I. (2008) MYC in breast tumor progression. *Expert Rev. Anticancer Ther.*, **8**, 1689–1698.

Danussi,C. *et al.* (2013) RHPN2 drives mesenchymal transformation in malignant glioma by triggering RhoA activation. *Cancer Res.*, **73**, 5140–5150.

Duffy,M.J. *et al.* (2014) p53 as a target for the treatment of cancer. *Cancer Treat. Rev.*, **40**, 1153–1160.

Fujikoshi,Y. *et al.* (2010) *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. John Wiley, Hoboken, NJ, Chapter 4

Gafni,E.M. and Bertsekas,D.P. (1984) Two-metric projection methods for constrained optimization. *SIAM J. Control Optim.*, **22**, 936–964.

Gross,A.M. *et al.* (2014) Multi-tiered genomic analysis of head and neck cancer ties TP53 mutation to 3p loss. *Nat Genet.*, **46**, 939–943.

Higham,N.J. (1988) Computing a nearest symmetric positive semi-definite matrix. *Linear Algebra Appl.*, **103**, 103–118.

Huan,J. *et al.* (2013) Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17$\beta$-estradion (E2). *Gene*, **533**, 346–355.

Johnson,R.A. and Wichern,D.W. (2007) *Applied Multivariate Statistical Analysis*. Vol. 1, 6th edn. Springer, New York, Chapter 4.

Jonsson,P.F. and Bates,P.A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics*, **22**, 2291–2297.

Kanehisa,M. *et al.* (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Katoh,M. *et al.* (2005) Comparative genomics on mammalian Fgf3-Fgf4 locus. *Int. J. Oncol.*, **27**, 281–285.

Kent,W.J. *et al.* (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.

Krupp,M. *et al.* (2011) The functional cancer map: a systems-level synopsis of genetic deregulation in cancer. *BMC Med. Genom.*, **4**, 53.

Kumar,R. *et al.* (2005) FBXO31 is the chromosome 16q24.3 senescence gene, a candidate breast tumor suppressor, and a component of an SCF complex. *Cancer Res.*, **65**, 11304–11313.

Kumar,P. *et al.* (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1436–1462.

Le Goff,C. *et al.* (2011) The ADAMTS(L) family and human genetic disorders. *Hum. Mol. Genet.*, **20**, R163–R167.

Liu,H. *et al.* (2012) MYC suppresses cancer metastasis by direct transcriptional silencing of $\alpha_V$ and $\beta_3$ integrin subunits. *Nat. Cell Biol.*, **14**, 567–574.

Malonia,N. *et al.* (2011) Gene regulation by SMAR1: role in cellular homeostasis and cancer. *Biochim. Biophys. Acta*, **1815**, 1–12.

Minaguchi,T. *et al.* (2001) PTEN mutation located only outside exons 5, 6, and 7 is an independent predictor of favorable survival in endometrial carcinomas. *Clin. Cancer Res.*, **7**, 2636–2642.

Moelans,N. *et al.* (2011) Frequent promoter hypermethylation of BRCA2, CDH13, MSH6, PAX5, PAX6 and WT1 in ductal carcinoma in situ and invasive breast cancer. *J. Pathol.*, **226**, 143.

Ng,S. *et al.* (2012) PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics*, **28**, i640–i646.

Nguyen,N.T. *et al.* (2014) A functional interplay between ZNF217 and estrogen receptor alpha exists in luminal breast cancers. *Mol. Oncol.*, **8**, 1441–1457.

Prasad,T. *et al.* (2009) Human protein reference database. *Nucleic Acids Res.*, **37**, D767–D772.

Sanchez-Garcia,F. *et al.* (2010) JISTIC: identification of significant targets in cancer. *BMC Bioinformatics*, **11**, 189.

Savas,F. *et al.* (2012) Serotonin transporter gene (SLC6A4) variations are associated with poor survival in colorectal cancer patients. *PLoS One*, **7**, e38953.

Scollen,S. *et al.* (2011) TGF-$\beta$ signaling pathway and breast cancer susceptibility. *Cancer Epidemiol. Biomarkers Prev.*, **20**, 1112–1119.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–177.

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

Wu,L. *et al.* (2014) A central role for TRPS1 in the control of cell cycle and cancer development. *Oncotarget*, **5**, 7677–7690.

Xu,J. *et al.* (2010) MYC and breast cancer. *Monographs Genes Cancer*, **1**, 629–640.

Xu,J. *et al.* (2014) Combined PTEN mutation and protein expression associate with overall and disease-free survival of glioblastoma patients. *Transl. Oncol.*, **7**, 196–205.

Zhang,F. and Chen,J.Y. (2010) Discovery of pathway biomarkers from coupled proteomics and systems biology methods. *BMC Genomics*, **11**, S12.