ABSTRACT: The construction of gene regulatory networks (GRNs) is an essential component of biomedical research to determine disease mechanisms and identify treatment targets. Gaussian graphical models (GGMs) have been widely used for constructing GRNs by inferring conditional dependence among a set of gene expressions. In practice, GRNs obtained by the analysis of a single data set may not be reliable due to sample limitations. Therefore, it is important to integrate multiple data sets from comparable studies to improve the construction of a GRN. In this article, we introduce an equivalent measure of partial correlation coefficients in GGMs and then extend the method to construct a GRN by combining the equivalent measures from different sources. Furthermore, we develop a method for multiple data sets with a natural missing mechanism to accommodate the differences among different platforms in multiple sources of data. Simulation results show that this integrative analysis outperforms the standard methods and can detect hub genes in the true network. The proposed integrative method was applied to 12 lung adenocarcinoma data sets collected from different studies. The constructed network is consistent with the current biological knowledge and reveals new insights about lung adenocarcinoma.

KEYWORDS: Gaussian graphical model, gene regulatory network, integrative analysis, multiple hypothesis test, partial correlation coefficient

Introduction

A gene regulatory network (GRN) describes the interactions among genes and how they work together to maintain life. The inference of GRNs leads to a systematic understanding of disease mechanisms at molecular levels and identification of potential therapeutic treatment targets for diseases. The development of high-throughput technologies has made it possible to simultaneously measure the activities of genes at the whole genome level, which greatly facilitates the study of GRNs. Gaussian graphical models (GGMs) have been widely used for inferring GRNs by estimating conditional dependence relationships among the expression levels of genes. The concept underlying GGMs is to use the partial correlation coefficient as a measure of dependency between the expressions of any 2 genes. Hence, inferring the GRN amounts to estimating its partial correlation coefficients or precision matrix (the inverse of covariance matrix) in GGMs, where partial correlation coefficients can be computed by entries in the precision matrix. However, applications of GGMs to high-dimensional data, where the number of genes $p$ is much larger than the number of patients $n$, are not trivial because the sample covariance matrix is singular and thus the precision matrix cannot be directly estimated.

In recent literature, there has been much work on sparse estimation of the precision matrix in GGM using regularization methods with a Lasso penalty. In addition to the regularization methods, other methods based on limited-order partial correlations have been proposed. These methods are intended to use a low-order partial correlation in lieu of the full-order partial correlation. Recently, Liang et al. proposed a new approach to construct a high-dimensional GGM based on an equivalent measure of partial correlation coefficients. However, in practical applications, GRNs constructed from the analysis of a single data set may lack reliability due to the limitation in sample size. An alternative approach is to use multiple data sets from comparable studies and conduct an integrative analysis.

Recent technologies have made it feasible to collect diverse and multiple genome-scale data sets in biomedical studies. For example, The Cancer Genome Atlas has collected genome, transcriptome, and epigenome data on more than 20 types of cancer from thousands of patients. The availability of such plentiful data enables researchers to construct more reliable GRNs to capture the heterogeneity of biological processes and phenotypes by borrowing strength across multiple sources of data. Guo et al. proposed the joint estimation method of multiple graphical models that share the same genes and dependence relationships among genes. They used the likelihood-based method with a hierarchical lasso penalty to encourage similar patterns of sparsity across multiple data sets. Similarly, Danaher et al. proposed the joint
graphical lasso (JGL) which uses the sparse group Lasso\textsuperscript{13} or the fused lasso penalty.\textsuperscript{14} However, when integrating multiple data sets measured using different microarray platforms, these penalty-based methods may suffer from discrepancies among different platforms. For example, when the expressions of several genes are missing in all patients in a specific platform but not in the other platforms, the penalty-based methods are not applicable because they require a complete design matrix without any missing values from multiple data sets. Even though the complete design matrix can be obtained by the deletion or imputation of missing values, it may suffer from a loss of information and severe bias.

Liang et al\textsuperscript{9} proposed a new method, namely, the $\psi$-learning method, to construct a GGM based on an equivalent measure of partial correlation coefficients, and they briefly introduced $\psi$-learning with data integration for multiple data sets. However, they focused only on the standard $\psi$-learning method for a single data set.

In this article, we study the integrative $\psi$-learning method to construct a GRN by integrating multiple sources of data. This approach is based on the equivalent measures of partial correlation coefficients in GGMs, and hence, it could be applied to high-dimensional multiple data sets because equivalent measures are evaluated with a reduced conditional set. Moreover, we provide an extension to the integrative $\psi$-learning method for multiple data sets with a natural missing mechanism caused using different platforms in multiple sources of data. The proposed method outperformed other standard methods in simulation studies. Finally, the proposed method was applied to study the gene regulation of lung adenocarcinoma (ADC). In this study, we integrated gene expression profiles of 1246 patients with ADC collected from 12 different studies and measured across different platforms. It is thus far the largest study on GRNs of ADC. The resulting GRN reveals several important hub genes in ADC and leads to new biological insights on the disease and its potential therapeutic targets. Finally, our sensitivity analysis shows that the identified hub genes are robust against random noise and selection of genes.

**Methods**

**The equivalent measure of partial correlation coefficients**

Following Liang et al\textsuperscript{9} we describe the GGM and some notation for defining the equivalent measure of partial correlation coefficients in GGMs. We then introduce the equivalent measure of partial correlation coefficients and the $\psi$-learning algorithm to construct a network using a single data set.

Let $X = (X_1, \ldots, X_p)^T$ denote a random vector drawn from the multivariate Gaussian distribution $N_p(\mu, \Sigma)$, where $\mu$ and $\Sigma$ are the mean vector and covariance matrix, respectively. The partial correlation coefficient between $X_i$ and $X_j$, denoted by $\rho_{ij} \mid V \setminus \{i, j\}$, where $V = \{1, \ldots, p\}$ is the index set of all variables. It is well known that the partial correlation coefficient in GGM can be expressed as follows:

$$\rho_{ij} \mid V \setminus \{i, j\} = -\frac{\omega_{ij}}{\sqrt{\omega_{ii} \omega_{jj}}}$$

where $\omega_{ij}$ is the $(i, j)$ entry of the precision matrix denoted by $\Omega = \Sigma^{-1} = (\omega_{ij})$. The Gaussian random vector $X$ can be represented by the undirected graph $G = (V, E)$, where $V$ is the set of vertices corresponding to $p$ variables and $E = (e_{ij})$ is the adjacency matrix which specifies the edges included in the graph $G$. For a GGM, the following relation holds:

$$e_{ij} = e_{ji} = 1 \Leftrightarrow \rho_{ij} \neq 0 \Leftrightarrow \rho_{ij} \mid V \setminus \{i, j\} \neq 0$$

In the context of GRNs, $X$ represents the expression levels of $p$ genes measured on each individual. Hence, constructing GRNs amounts to identifying their nonzero partial correlation coefficients.

Let $r_{ij}$ be the correlation coefficient between $X_i$ and $X_j$, and let $G_{ij}$ denote a reduced graph of $G$ with $e_{ij}$ being set to 0. We define $r_{G_{ij}}(i)$ as a set of vertices for which the corresponding variable is correlated with $X_i$ in $G_{ij}$, i.e.,

$$r_{G_{ij}}(i) = \{v : r_{ij} \neq 0 \} \setminus \{j\}.$$

Similarly, we define $r_{G_{ij}}(j) = \{v : r_{ij} \neq 0 \} \setminus \{i\}$. For any pair of vertices $i$ and $j$, Liang et al\textsuperscript{9} proposed the $\psi$-partial correlation coefficient defined by the following equation:

$$\psi_{ij} = \rho_{ij} \mid S_{ij}$$  \hspace{1cm} (1)

where $S_{ij} = r_{G_{ij}}(i)$ if $|r_{G_{ij}}(i)| > |r_{G_{ij}}(j)|$ and $S_{ij} = r_{G_{ij}}(j)$ otherwise, and $|D|$ is the cardinality of a set $D$. They showed that $\psi_{ij}$ and $\rho_{ij} \mid V \setminus \{i, j\}$ are equivalent in the sense that

$$\psi_{ij} = 0 \Leftrightarrow \rho_{ij} \mid V \setminus \{i, j\} = 0$$

under the faithfulness assumption for GGM. Based on the equivalent measure of partial correlation coefficient $\psi_{ij}$, the Gaussian graphical network can be constructed by the $\psi$-learning algorithm summarized in Algorithm 1. Furthermore, Liang et al\textsuperscript{9} established the asymptotic consistency property of the $\psi$-learning method under mild conditions.

**Algorithm 1. $\psi$-learning algorithm**

1. **Step 1. (Correlation screening)—**Conduct a multiple hypothesis test to identify the pairs of vertices for which the correlation coefficient is significantly different from zero.

2. **Step 2. (\psi-calculating)—**For each pair of vertices $i$ and $j$, identify the conditional set $S_{ij}$ based on the results in Step 1 and calculate $\psi_{ij}$ by inverting the sample covariance matrix of the variables indexed by $S_{ij} \cup \{i, j\}$.

3. **Step 3. (\psi-screening)—**Conduct a multiple hypothesis test to identify the pairs of vertices for which $\psi_{ij}$ is significantly different from zero.
Integrative $\psi$-learning method

The $\psi$-learning method described in Algorithm 1 focuses on constructing a network based on only a single data set. However, with the recent development of high-throughput technologies, it is common to gather multiple genome-scale data sets to study the mechanisms of a disease. In real-world applications, it is much more efficient to integrate such plentiful and compatible data, which could lead to more reliable GRNs. In this section, we describe how to construct a GRN by integrating multiple sources of data under the framework of the $\psi$-learning method.

Suppose that we have $K$ sources of data, all of which are normally distributed. Let $\hat{\psi}_{ij}^{(k)}$ be the estimated $\psi$-partial correlation coefficient in equation (1) from the $k$th source of data. We first apply the Fisher transformations to obtain the following equation:

$$
Z_{ij}^{(k)} = \frac{1}{2} \log \left( \frac{1 + \hat{\psi}_{ij}^{(k)}}{1 - \hat{\psi}_{ij}^{(k)}} \right)
$$

(2)

whose distribution is approximately normally distributed with mean zero and variance $1/(n_k - |S_{ij}^{(k)}| - 3)$ under the null hypothesis $H_0: \psi_{ij}^{(k)} = 0$, where $n_k$ is the sample size of the $k$th source and $(n_k - |S_{ij}^{(k)}| - 3)$ is called the effective sample size of the $\psi$-partial correlation coefficient.\(^9\) For convenience, we call the scaled $z$-score (equation (2)) a $\psi_z$-score defined by $z_{ij}^{(k)} = z_{ij}^{(k)} \sqrt{n_k - |S_{ij}^{(k)}| - 3}$, whose distribution approximately follows a standard normal distribution under $H_0$. We then combine the $\psi_z$-scores from different sources of data using the Stouffer meta-analysis method\(^15\) as follows:

$$
Z_y = \frac{\sum_{k=1}^{K} \omega_k z_{ij}^{(k)}}{\sqrt{\sum_{k=1}^{K} \omega_k^2}}, \quad i, j = 1, \ldots, p
$$

(3)

where $\omega_k$ is a nonnegative weight assigned to the $k$th source of data. The assignment of $\omega_k$ may depend on the sample size or data quality for different sources known in advance. If a prior knowledge for each source of data is not available, we simply use the weight proportional to the sample size: for example, $\omega_k = n_k / \sum_k n_k$. We notice that the weight $\omega_k$ for the $k$th source of data in equation (3) is set to be the same for all edges.

In real-world applications, it is common for different data sets to be collected using different microarray platforms. But this will create missing values for some genes in some of the data sets when combining all the data sets together due to the differences among platforms. In this case, a standard approach is to apply a method after deleting the patients or genes with missing values or to impute missing values. As a result, the network may suffer from a loss of information and severe bias. Moreover, both deletion and imputation of missing values are inappropriate in real applications because too many missing values exist for many genes in a source of data.

For the purpose of analyzing the data with missing values, we propose to use different weights for each edge in a source of data. Let $n_i^{(k)}$ be the number of samples except for those with missing values for the $k$th gene in the $k$th source of data, and $n_j^{(k)} = \max(n_i^{(k)}, n_j^{(k)} + 1)$. We modify the method to combine $\psi_z$-scores from different sources in equation (3) in the following way:

$$
Z_y = \frac{\sum_{k=1}^{K} \omega_k^{(i)} z_{ij}^{(k)}}{\sqrt{\sum_{k=1}^{K} \omega_k^{(i)} z_{ij}^{(k)}}^2}, \quad i, j = 1, \ldots, p
$$

(4)

where $\omega_k^{(i)}$ denotes the nonnegative weight for the edge $e_{ij}$ assigned on the $k$th source of data. Similarly, if prior knowledge for each source of data is not available, we simply set the weight proportional to the sample size, $\omega_k^{(i)} = n_k^{(i)} / \sum_k n_k^{(i)}$. For each edge $e_{ij}$, $\omega_k^{(i)} = 0$ if the expressions of gene $i$ or gene $j$ are missing in source $k$, but $\omega_k^{(i)} \neq 0$ otherwise. For a fixed $e_{ij}$, even if many missing values in genes $i$ or $j$ exist in a specific platform (a source of data), the integrative $\psi$-learning method can be applied to $Z_y$ in equation (4) computed by other sources of data unless the expressions of the corresponding genes are missing in all sources of data. This enables us to partially use the information from the other sources of data, which is not achieved by the penalty-based joint estimation method because it requires a complete design matrix from all sources.

Note that $Z_y$ approximately follows a standard normal distribution under the null hypothesis $H_0: e_{ij} = 0$. Then, a multiple hypothesis test can be performed on $Z_y$ to identify the pairs of vertices for which $Z_y$ is differentially distributed from the standard normal $N(0,1)$. The integrative $\psi$-learning algorithm is summarized in Algorithm 2.

Algorithm 2. Integrative $\psi$-learning algorithm.

Step 1. ($\psi_z$-calculation)—Perform Steps 1 and 2 of the $\psi$-learning algorithm independently for each source of data.

Step 2. ($\psi_z$-combination)—Calculate $Z_y$ by combining $\psi_z$-scores in equations (2) and (4).

Step 3. ($\psi_z$-screening)—Conduct a multiple hypothesis test to identify the pairs of vertices for which $Z_y$ is differentially distributed from the standard normal $N(0,1)$.

For the multiple hypothesis tests in Steps 1 and 3, we use the stochastic approximation-based method,\(^16\) which is also used in the $\psi$-learning method in Algorithm 1. Overall, this method works well under general dependence between test statistics. It is obvious that the correlation coefficients and $\psi$-partial correlation coefficients are generally dependent for
GGMs. For the multiple hypothesis test procedure, one important issue is to choose significance levels used as cutoff values of correlation coefficients and $\psi$-partial correlation coefficients. We set the significance level in the Storey $q$ value\(^{17}\) as in the $\psi$-learning of Liang et al.\(^3\) For correlation screening, we generally use a high $q$ value (eg, 0.05 or even larger). When $p$ is extremely large with sample size $n$ and the $q$ value is large, the case $|S_{ij}^{(k)}| > r_{ij}^{(k)} - 3$ might occur where $\psi_{ij}^{(k)}$ is incalculable. A small $q$ value would reduce the computational complexity of $\psi$-partial correlation coefficients by the small size of the conditional set $S_{ij}^{(k)}$, but the calculated $\psi$-partial correlation coefficients may be less reliable. In this article, we set the $q$ value to 0.05 for correlation screening. For $\psi$-screening, a large $q$ value produces a dense network, whereas a small $q$ value leads to a sparse network. Similar to regularization methods such as thelasso,\(^3\) a network path can be obtained with a monotone sequence of $q$ values for $\psi$-screening.

Finally, it is worth noting that the Stouffer meta-analysis method in equation (3) could also be replaced by other probability test methods, such as the Fisher method and the Pearson method.\(^{18}\) This can be done by combining the corresponding $P$ values of each hypothesis instead of the $z$-score. The multiple hypothesis tests in Steps 1 and 3 can be also performed in various methods, such as the positive false discovery rate method,\(^{19}\) empirical Bayesian method,\(^{20}\) 2-stage method,\(^{21}\) and principal factor approximation method.\(^{22}\)

### Results and Discussion

#### Simulation studies

In this section, we investigate the performance of the integrative $\psi$-learning method on simulated data based on 2 real protein-protein interaction networks.\(^{23}\) The underlying network structures were partially selected from the human protein reference database and form an approximate scale-free topology, as shown in Figure 1.\(^{24}\) The strengths of dependencies between 2 proteins with interaction are generated from a normal distribution with mean 0.5 and variance 0.2, and then the sign of dependencies is randomly decided by Bernoulli random variables with probability .5 so that we allow both positive and negative regulations. The expression data are generated from the conditional normal distribution. Specifically, the expression of gene $j$ is simulated from the following equation:

$$X_i | X_{-i} \sim N\left(\sum_{j \in A_i} \beta_{ij} X_j \sigma^2\right)$$

where $X_i$ is the expression level of gene $i$, $A_i$ is the set of genes with connection to gene $i$ in the true underlying network, and $\beta_{ij}$ is the strength of dependency between genes $i$ and $j$.

We consider 2 networks with small $\psi$-learning method on simulated data based on 2 real networks (ie, the set of genes that regulate gene $i$), and $\beta_{ij}$ is the strength of dependency between genes $i$ and $j$.

For comparison, we consider the $\psi$-learning to be applied separately to each data set ($\psi_i$-learning, $k = 1, 2, 3$) and $\psi$-learning to be simply applied to the pooled data set with a size of $3n$ ($\psi_p$-learning) along with the integrative $\psi$-learning method ($\psi$, learning). We further consider the JGL method with the group lasso penalty,\(^{12}\) for which the final network is constructed by the nonzero $\ell_2$-norms of coefficients across 3 sources of data for each edge.

Figure 2 displays the receiver operating characteristic (ROC) and partial ROC curves of all methods for a sample data set in the simulation with a small size network. The partial ROC curve represents the ROC curve on the region where the false-positive rate is less than 0.05. Figure 2 shows that the $\psi_i$-learning method performs much better than other methods. For the separate $\psi$-learning methods based on each data set, as expected, the $\psi_i$-learning method performs better than other ones from data sets with larger noise levels. It is interesting to
note that the $\psi_p$-learning method performs worse than the $\psi_1$-learning and $\psi_2$-learning methods, although the $\psi_p$-learning method uses more samples. This result might be caused by the third data set’s poor quality, which means that simply pooling data sets from multiple sources may not be a good choice, whereas the proposed integrative analysis provides an effective way to improve the performances of the $\psi$-learning method using more information.

For the performance measures, we consider the numbers of true-positive edges and false-positive edges (TPE and FPE), sensitivity and specificity (SEN and SPE), the rate of misspecified edges (MIS), and the area under ROC and partial ROC curves under $\text{FPR} < 0.05$. Table 1 displays the average values of performance measures over 20 replications in which the levels of sparsity for the $\psi$-learning methods are selected by the Storey $q$ value in multiple hypothesis tests. For the JGL method, we selected a level of sparsity similar to the $\psi_i$-learning method because the JGL method using cross-validation or information criteria tends to produce too dense networks. The results indicate that the $\psi_i$-learning method performs much better than all the other methods. The $\psi_3$-learning from data with the highest noise shows the worst performance, and the $\psi_p$-learning method also shows worse performance than the $\psi_1$-learning method.

Figure 3 and Table 2 show the results for the simulations based on a large network. We again see that the $\psi_i$-learning method overall outperforms all other methods and works well for the high-dimensional data. Figure 4 displays the network paths constructed by various $q$ values for the $\psi_i$-learning method. The $\psi_i$-learning method with a large $q$ value produces a sparse network, whereas one with a small $q$ value leads to a dense network, as shown in Figure 4. For a small $q$ value, the resulting network not only shows a lower number of FPEs but also selects a lower number of TPEs. In contrast, the network constructed by a large $q$ value not only identifies the TPEs well but also selects more irrelevant edges, as seen in Tables 1 and 2. Figure 4 indicates that the $\psi_i$-learning method with an appropriate $q$ value detects the hub genes well in the true network. These results suggest that the integrative analysis for multiple data sets would improve the performance of the $\psi$-learning method and would enable us to construct more reliable GRNs.

Finally, we investigate the robustness of the integrative $\psi$-learning method for data sets with missing values. To do this, we consider the large size network with $n = 100$ and $p = 612$. For each simulation, we randomly pick 1 data set and select 10 genes in this data set and set their expression values to be missing in all samples. Note that it mimics the multiple data sets to be collected using different microarray platforms, and other methods except the proposed method are inappropriate for data sets with missing values as mentioned previously. Figure 5 displays the proportion of each true hub gene being detected by the integrative $\psi_i$-learning method among 100 simulations, and Table 3 compares the integrative $\psi$-learning methods under the situations with and without missing values. These results indicate that the integrative $\psi$-learning method is insensitive for multiple data sets with missing values.

**Application to multiple lung cancer adenocarcinoma data sets**

Lung cancer is one of the most prominent types of cancer. It is the leading cause of cancer death in the United States in both men and women. Non–small-cell lung cancer is the most common cause of lung cancer death, accounting for up to 85% of deaths from lung cancer. About 40% of all lung cancers are ADCs. In this study, we constructed a lung cancer adenocarcinoma–specific gene network by integrating several messenger RNA (mRNA) expression data sets collected from the public.
domain to better understand the molecular mechanisms associated with patient survival.

Data sets with patients annotated as ADC were selected from Lung Cancer Explorer database (http://qbrc.swmed.edu/lce/). Twelve data sets were collected; their inclusion criteria and details are summarized in Table 4. This study integrates mRNA expression data on patients with lung ADC from 12 data sets. The total data set consists of the gene expression levels of 21,353 genes in 1,281 patients with lung ADC. Thus far, this is the largest study to use mRNA expression data to construct a gene network in lung ADC. Genes from different genome-wide platforms are mapped using Probemapper. When multiple probes were mapped to a gene, the arithmetic mean of the probe-level expression was computed as the expression level for the gene. The gene expression profiles of each sample were log_2-transformed and standardized to have zero median and unit variance. Following Tang et al., we first selected 711 genes that are significantly associated with overall patient survival times after adjusting for clinical factors such as age, gender, and tumor stage in the Lung Cancer Consortium study. In this study, we focused on the construction of the global GRN based on the selected 711 genes, and thus, the data set for the analysis had 1,281 patients from 12 different sources and 711 genes. Note that for each data source, the number of patients is still less than the number of genes.

As mentioned previously, the total data set had too many missing values caused by the differences in genome-wide platforms from different studies. Table 4 displays the missing values over 20 replications for each measure.

### Table 1. Comparison of the proposed method with the JGL, separate, and pooled ψ-learning methods in the simulation on small size network.

| Q VALUE | N   | METHOD | TPE   | FPE   | SEN   | SPE   | MIS   | AUC   | PAUC   |
|---------|-----|--------|-------|-------|-------|-------|-------|-------|--------|
| 0.1     | 100 | JGL    | 30.95 | 74.90 | 0.3005 | 0.9773 | 0.0432 | 0.8185 | 0.0087 |
|         |     | ψ1-learning | 61.40 | 10.05 | 0.5961 | 0.9970 | 0.0152 | 0.9464 | 0.0290 |
|         |     | ψ2-learning | 30.95 | 8.95  | 0.3005 | 0.9773 | 0.0238 | 0.8458 | 0.0211 |
|         |     | ψ3-learning | 38.60 | 6.05  | 0.3748 | 0.9982 | 0.0207 | 0.8928 | 0.0233 |
|         |     | ψ4-learning | 26.95 | 3.65  | 0.2617 | 0.9989 | 0.0234 | 0.8545 | 0.0223 |
|         |     | ψ5-learning | 0.75  | 0.70  | 0.0073 | 0.9998 | 0.0303 | 0.6645 | 0.0062 |
| 300     |     | JGL    | 63.30 | 68.10 | 0.6146 | 0.9794 | 0.0317 | 0.9281 | 0.0214 |
|         |     | ψ1-learning | 90.05 | 16.30 | 0.8743 | 0.9951 | 0.0086 | 0.9868 | 0.0360 |
|         |     | ψ2-learning | 64.50 | 11.00 | 0.6262 | 0.9967 | 0.0145 | 0.9395 | 0.0284 |
|         |     | ψ3-learning | 74.05 | 16.85 | 0.7189 | 0.9949 | 0.0135 | 0.9608 | 0.0326 |
|         |     | ψ4-learning | 62.75 | 5.75  | 0.6092 | 0.9983 | 0.0135 | 0.9342 | 0.0326 |
|         |     | ψ5-learning | 16.75 | 2.40  | 0.1626 | 0.9993 | 0.0261 | 0.8239 | 0.0194 |
| 0.3     | 100 | JGL    | 36.40 | 112.25| 0.3534 | 0.9660 | 0.0526 | 0.8185 | 0.0087 |
|         |     | ψ1-learning | 70.90 | 32.95 | 0.6883 | 0.9900 | 0.0191 | 0.9464 | 0.0290 |
|         |     | ψ2-learning | 43.50 | 36.65 | 0.4223 | 0.9889 | 0.0283 | 0.8458 | 0.0211 |
|         |     | ψ3-learning | 50.50 | 22.10 | 0.4903 | 0.9933 | 0.0219 | 0.8928 | 0.0233 |
|         |     | ψ4-learning | 40.25 | 18.75 | 0.3908 | 0.9943 | 0.0239 | 0.8545 | 0.0223 |
|         |     | ψ5-learning | 1.15  | 1.35  | 0.0112 | 0.9996 | 0.0303 | 0.6645 | 0.0062 |
| 300     |     | JGL    | 69.95 | 122.10| 0.6791 | 0.9630 | 0.0456 | 0.9281 | 0.0214 |
|         |     | ψ1-learning | 94.60 | 48.55 | 0.9184 | 0.9853 | 0.0167 | 0.9868 | 0.0360 |
|         |     | ψ2-learning | 72.45 | 39.75 | 0.7034 | 0.9880 | 0.0207 | 0.9395 | 0.0284 |
|         |     | ψ3-learning | 82.65 | 47.20 | 0.8024 | 0.9857 | 0.0199 | 0.9608 | 0.0326 |
|         |     | ψ4-learning | 72.15 | 28.65 | 0.7005 | 0.9913 | 0.0175 | 0.9342 | 0.0326 |
|         |     | ψ5-learning | 31.80 | 14.20 | 0.3087 | 0.9957 | 0.0251 | 0.8239 | 0.0194 |

Abbreviations: AUC, area under ROC; FPE, false-positive edges; JGL, joint graphical lasso; MIS, misspecified edges; PAUC, area under the partial ROC; ROC, receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; TPE, true-positive edges.

Table reports the average values over 20 replications for each measure.
Figure 3. ROC curve and partial ROC curve under FPR < 0.05 for all methods where the sample and network sizes are $n = 100$ and $p = 612$, respectively. FPR indicates false-positive rate; JGL, joint graphical lasso; ROC, receiver operating characteristic curve.

Table 2. Comparison of the proposed method with the JGL, separate, and pooled $\psi$-learning methods in the simulation on large size network.

| Q VALUE | N   | METHOD  | TPE  | FPE  | SEN  | SPE  | MIS  | AUC  | PAUC |
|---------|-----|---------|------|------|------|------|------|------|------|
| 0.1     | 100 | JGL     | 156.05 | 755.00 | 0.1864 | 0.9959 | 0.0077 | 0.6055 | 0.0095 |
|         |     | $\psi$-learning | 384.55 | 67.70 | 0.4594 | 0.9966 | 0.0028 | 0.9264 | 0.0273 |
|         |     | $\psi$-learning | 191.70 | 144.35 | 0.2290 | 0.9992 | 0.0042 | 0.8381 | 0.0185 |
|         |     | $\psi$-learning | 267.45 | 52.60 | 0.3195 | 0.9997 | 0.0033 | 0.8862 | 0.0207 |
|         |     | $\psi$-learning | 139.00 | 26.90 | 0.1661 | 0.9999 | 0.0039 | 0.8473 | 0.0198 |
|         |     | $\psi$-learning | 0.15   | 1.00  | 0.0002 | 1.0000 | 0.0045 | 0.6002 | 0.0036 |
| 300     |     | JGL     | 458.35 | 1183.30 | 0.5476 | 0.9936 | 0.0084 | 0.9173 | 0.0269 |
|         |     | $\psi$-learning | 675.65 | 162.65 | 0.8072 | 0.9991 | 0.0017 | 0.9811 | 0.0384 |
|         |     | $\psi$-learning | 477.80 | 204.60 | 0.5708 | 0.9989 | 0.0030 | 0.9326 | 0.0295 |
|         |     | $\psi$-learning | 532.50 | 156.05 | 0.6362 | 0.9992 | 0.0025 | 0.9488 | 0.0281 |
|         |     | $\psi$-learning | 449.90 | 61.70  | 0.5375 | 0.9997 | 0.0024 | 0.9267 | 0.0295 |
|         |     | $\psi$-learning | 85.05  | 13.55  | 0.1016 | 0.9999 | 0.0041 | 0.8310 | 0.0183 |
| 0.3     | 100 | JGL     | 156.05 | 755.00 | 0.1864 | 0.9959 | 0.0077 | 0.6055 | 0.0095 |
|         |     | $\psi$-learning | 454.25 | 276.40 | 0.5427 | 0.9985 | 0.0035 | 0.9264 | 0.0273 |
|         |     | $\psi$-learning | 280.80 | 766.60 | 0.3355 | 0.9959 | 0.0071 | 0.8381 | 0.0185 |
|         |     | $\psi$-learning | 335.80 | 220.35 | 0.4012 | 0.9988 | 0.0039 | 0.8862 | 0.0207 |
|         |     | $\psi$-learning | 220.85 | 150.40 | 0.2639 | 0.9992 | 0.0041 | 0.8473 | 0.0198 |
|         |     | $\psi$-learning | 0.45   | 3.40   | 0.0005 | 1.0000 | 0.0045 | 0.6002 | 0.0036 |
| 300     |     | JGL     | 458.35 | 1183.30 | 0.5476 | 0.9936 | 0.0084 | 0.9173 | 0.0269 |
|         |     | $\psi$-learning | 711.75 | 490.85 | 0.8504 | 0.9974 | 0.0033 | 0.9811 | 0.0384 |
|         |     | $\psi$-learning | 539.75 | 825.30 | 0.6449 | 0.9956 | 0.0060 | 0.9326 | 0.0295 |
|         |     | $\psi$-learning | 587.75 | 470.20 | 0.7022 | 0.9975 | 0.0038 | 0.9488 | 0.0281 |
|         |     | $\psi$-learning | 502.35 | 255.45 | 0.6002 | 0.9986 | 0.0032 | 0.9267 | 0.0295 |
|         |     | $\psi$-learning | 151.60 | 83.45  | 0.1811 | 0.9996 | 0.0041 | 0.8310 | 0.0183 |

Abbreviations: AUC, area under ROC; FPE, false-positive edges; JGL, joint graphical lasso; MIS, misspecified edges; PAUC, area under the partial ROC; ROC, receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; TPE, true-positive edges.

Table reports the average values over 20 replications for each measure.
information for each source of data. For some data sets, the expression levels of many genes were missing in all patients for each source. In this case, it is not reasonable to delete missing samples or genes because there were too many. Moreover, the imputation of such expression levels based on information from other sources without missing values may cause severe biases due to discrepancies in the selection of different platforms. Thus, penalty-based joint estimation methods such as the JGL are not applicable for this type of data because it requires a complete design matrix from all sources. However, the integrative $\psi$-learning method is appropriate for multiple data sets with a natural missing mechanism because it is based on meta-analysis using a combination of separate results from each data set. This approach enables us to partially use the information from each source for each edge, which could not be achieved by the penalty-based joint estimation methods because it requires a complete design matrix from all sources. For example, for GAPDH and ANKRD46 genes, there are many missing values in some data sets, but there are no missing values in other data sets. Although the $\psi_z$-scores in equation (2) cannot be computed in the data sets with missing values, the combined $\psi_z$-score for the relationship between GAPDH and ANKRD46 can be obtained by combining the $\psi_z$-scores from other sources, where the corresponding weights $\omega_{ij}^{(4)}$ for data sets with missing values are set to zero in equation (4).

Figure 4. Path of networks constructed by various $q$ values in the simulation on large size network: (A) true, (B) $q$ value = 0.000001, (C) $q$ value = 0.001, (D) $q$ value = 0.01, (E) $q$ value = 0.1, and (F) $q$ value = 0.3, where the large nodes with red color represent the hub gene node whose node degrees are greater than 9.
Figure 5. Proportion of each true hub gene being detected by the proposed method in the simulation on large size network: (upper panel) $q$ value = 0.1; (bottom panel) $q$ value = 0.3.

| Q VALUE | MISSING | TPE     | FPE     | SEN     | SPE     | MIS     | AUC     | PAUC    |
|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 0.1     | Without | 384.55  | 67.70   | 0.4594  | 0.9996  | 0.0028  | 0.9264  | 0.0273  |
|         | With    | 368.00  | 64.65   | 0.4397  | 0.9996  | 0.0028  | 0.9236  | 0.0275  |
| 0.3     | Without | 454.25  | 276.40  | 0.5427  | 0.9985  | 0.0035  | 0.9264  | 0.0273  |
|         | With    | 444.65  | 259.65  | 0.5312  | 0.9986  | 0.0035  | 0.9236  | 0.0275  |

Abbreviations: AUC, area under ROC; FPE, false-positive edges; JGL, joint graphical lasso; MIS, misspecified edges; PAUC, area under the partial ROC; ROC, receiver operating characteristic curve; SEN, sensitivity; SPE, specificity; TPE, true-positive edges.

Table 4. Summary of 12 lung ADC data sets from different sources.

| SOURCE    | SAMPLES | $P_{mis} / N_{mis}$ | SOURCE    | SAMPLES | $P_{mis} / N_{mis}$ |
|-----------|---------|---------------------|-----------|---------|---------------------|
| Shedden et al$^{28}$ | 442     | 2/256               | Matsuyama et al$^{29}$ | 94      | 160/94              |
| Kuner et al$^{30}$   | 40      | —                   | Bhattacharjee et al$^{31}$ | 138     | 124/138             |
| Zhu et al$^{32}$     | 28      | —                   | Tang et al$^{27}$       | 133     | 87/133              |
| Hou et al$^{33}$     | 50      | —                   | Wright et al$^{34}$     | 36      | 221/+5 (176/36)     |
| Lee et al$^{35}$     | 62      | —                   | Larsen et al$^{36}$     | 48      | 226/+5 (176/48)     |
| Bild et al$^{37}$    | 58      | —                   | Tomida et al$^{38}$     | 117     | 112/+5              |

$P_{mis} / N_{mis}$ means that $P_{mis}$ genes are missing in $N_{mis}$ patients, and +5 stands for 5 more patients.
Figure 6 displays the structure of networks constructed by the integrative $\psi$-learning method at the $q$ value of 0.0001, in which 602 edges were detected for 456 genes. The constructed networks consisted of many genes with few connections and a few hub genes with many connections. The activity of hub genes may affect many genes in the biological network, and hence, it is expected to play an important role in biology. In the constructed network, we identified 24 potential hub genes whose degrees are greater than 95% quantile $\tau$ of node degrees for each network.

Table 5 summarizes the hub genes identified by the method at the $q$ value of 0.0001. Among the identified genes, most are cancer-related genes previously identified in the literature, and several are known lung cancer genes. For example, CDH1 is a classical tumor suppressor gene and it encodes for epithelial cadherin, which is important for cell-cell adhesion. Disruption of the cadherin-catenin complex at the cell-cell junction from CDH1 loss or mutation often leads to decreased cell-cell adhesion, increased mobility, and induction of epithelial-mesenchymal transition, which favors the formation of metastasis.\(^{39}\) SMAD5 encodes for one of the Smad proteins that are transcription factors downstream of the transforming growth factor $\beta$ (TGF-$\beta$) pathway. The TGF-$\beta$ pathway is often dysregulated in cancer and modulates multiple processes in cancer development.\(^{40}\) MST1R encodes for a receptor tyrosine kinase (RON) that uses macrophage-stimulating proteins (MSPs) as ligands. The MSP-RON signaling has been implicated in the invasive growth of several types of cancer.\(^{41}\) DUSP6 encodes for dual specific phosphatase 6 that dephosphorylates and inactivates ERK2 to inhibit mitogen-activated protein kinase signaling. The expression of DUSP6 is regulated by epidermal growth factor receptor (EGFR) signaling, and depletion of DUSP6 from EGFR inhibition has been implicated in resistance to EGFR inhibitors.\(^{42}\) In summary, the proposed method...
seems appropriate for multiple data sets and should perform well in real applications.

Finally, we performed sensitivity analysis to test the stability of the identified hub genes. We randomly selected 100 genes and generated their expression levels by randomly shuffling their original expression across different patients. Then, we added these 100 random genes to the 711 survival-related genes and constructed the network of the total 811 (711 + 100) genes. At the similar sparsity level (605 edges detected for 444 genes among the 811 genes, compared with 602 edges detected for 456 genes among the 711 genes in the original analysis), we found that 20 of the 24 hub genes (except FOXM1, LDHA, PI4KA, ZNF3) in the original analysis remain to be hub genes when adding the 100 random genes, which indicates that the identified hub genes are reasonably robust against random noise and selection of genes.

Conclusions

In this article, we developed an approach for constructing a GRN by integrating multiple sources of data. We have introduced an integrative ψ-learning method for construction of GRNs using multiple data sets. Numerical results show that the integrative ψ-learning method improves the performances of the standard ψ-learning method and identifies hub genes well in the true network. Furthermore, we have proposed an extension of the integrative ψ-learning method for the data with a natural missing mechanism caused using different platforms for each source of data. The proposed integrative ψ-learning method was applied to multiple lung cancer adenocarcinoma data sets with many missing values and detected hub genes well that are more likely to be meaningful in biological networks.

Author Contributions

GX and FL proposed the research project. SL performed the statistical analysis and drafted the manuscript. FL provided the implementation of our method and insightful comments on the manuscript. GX and LC provided a biological interpretation of the results in the real application. All authors wrote and approved the final manuscript.

REFERENCES

1. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B Met. 1996;58:267–288.
2. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection. J Roy Stat Soc B Met. 2010;72:169–2015.
3. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–441.
4. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse re-

gression models. J Am Stat Assoc. 2009;104:735–746.
5. Magwene PM, Kim J. Estimating genomic coexpression networks using first-

order conditional independence. Genome Biol. 2004;5:1–2373.
6. Castelo R, Roverato A. A robust procedure for Gaussian graphical model search from microarray data with p larger than n. J Mach Learn Res. 2006;7:2621–2650.
7. Wille A, Bühlmann P. Low-order conditional independence graphs for inferring genetic networks. Stat Appl Genet Mol Biol. 2006;5:Article1.
8. Liang F, Song Q, Qiu P. An equivalent measure of partial correlation coefficients for high dimensional Gaussian graphical models. J Am Stat Assoc. 2015;110:1248–1265.
9. Guo J, Levina E, Michailidis G, Zhu J. Joint estimation of multiple graphical models. Biometrika. 2011;98:1–15.
10. Zhou N, Zhu J. Group variable selection via a hierarchical lasso and its oracle property. Stat Interface. 2010;3:557–574.
11. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. J Roy Stat Soc B Met. 2014;76:373–397.
12. Simon N, Friedman J, Hastie T, Tibshirani R. A sparse-group lasso. J Comput Graph Stat. 2013;22:231–245.
13. Hoff P. A path algorithm for the fused lasso signal approximator. J Comput Graph Stat. 2010;19:984–1006.
14. Stouffer S, Suchman E, DeVinney L. The American Soldier: Adjustment during Army Life, vol. 1. Princeton, NJ: Princeton University Press, 1949.
15. Liang F, Zhang J. Estimating the false discovery rate using the stochastic ap-

proximation algorithm. J Comput Graph Stat. 2006;15:187–205.
16. Efron B. Large-scale simultaneous hypothesis testing. J Am Stat Assoc. 2004;99:60–104.
17. Benjamini Y, Krieger AM, Yekutieli D. Adaptive linear step-up procedures that control the false discovery rate. Biometrika. 2006;93:491–507.
18. Fan J, Han X, Gu W. Estimating false discovery proportion under arbitrary co-

variance dependence. J Am Stat Assoc. 2012;107:1019–1035.
19. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. PloS ONE. 2012;7:e29348.
20. Prasad TK, Goel R, Kandasamy K, et al. Human protein reference database? 2009 update. Nucleic Acids Res. 2009;37:D675–D677.
21. Jemal A, Siegel R, Ward E, et al. Cancer statistics, 2008. CA Cancer J Clin. 2008;58:7–45.
22. Allen JD, Wang S, Chen M, et al. Probe mapping across multiple microarray platforms. Brief Bioinform. 2012;13:547–554.
23. Tang H, Xiao G, Behrens C, et al. C12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. Clin Cancer Res. 2013;19:1577–1586.
24. Shkeden K, Taylor JM, Enkemann SA, et al. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. Nat Med. 2008;14:822–827.
25. Matsuyama Y, Suzuki M, Arima C, et al. Prostate-specific nuclear factor PSMD2 as a potential therapeutic target in association with various clinicopath-

ologic features in lung adenocarcinomas. Mol Carcinog. 2011;50:301–309.
26. Kuerer H, Muley T, Meister M, et al. Global gene expression analysis reveals spec-

ific patterns of cell junctions in non-small cell lung cancer subtypes. Lung Cancer. 2009;63:32–38.
27. Bhattacharjee A, Richards WG, Staunton J, et al. Classification of human lung cancers by mRNA expression profiling distinguishes adenocarcinoma sub-

classes. Proc Natl Acad Sci U S A. 2001;98:13790–13795.
28. Zhu C-Q, Ding K, Strumpf D, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non–small-cell lung cancer. J Clin Oncol. 2010;28:4417–4424.
29. Hess J, Aerts J, Den Hemer B, et al. Gene expression-based classification of non-

small cell lung carcinomas and survival prediction. PloS ONE. 2010;5:e10312.
30. Wright CM, Larsen JE, Hayward NK, et al. ADAM28: a potential oncogene involved in adenosquamous lung adenocarcinomas. Genes Chromosomes Cancer. 2010;49:688–698.
31. Lee E-S, Son D-S, Kim S-H, et al. Prediction of recurrence-free survival in postoperative non-small cell lung cancer patients by using an integrated model of clinical information and gene expression. Clin Cancer Res. 2008;14:7397–7404.
32. Larsen JE, Pavey SJ, Passmore LH, Bowman RV, Hayward NK, Fong KM. Gene expression signature predicts recurrence in lung adenocarcinoma. Clin Cancer Res. 2007;13:2946–2954.
33. Bild AH, Yao G, Chang JT, et al. Oncogenic pathway signatures in human can-

cers as a guide to targeted therapies. Nature. 2006;439:353–357.
34. Tomoda S, Takeuchi T, Shimada Y, et al. Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. J Clin Oncol. 2009;27:2793–2799.
35. Van Roy F, Bers G. The cell-cell adhesion molecule E-cadherin. Cell Mol Life Sci. 2006;63:3756–3788.
36. Massague J. TGFbeta in cancer. Cell. 2008;134:215–230.
37. Yu H-P, Zhou Y-Q, Zhang R, Wang M-H. MSP-BDNF signalling in cancer: pathogenesis and therapeutic potential. Nat Rev Cancer. 2013;13:466–481.
38. Tetsu O, Phuchareon J, Eisele DW, Hanger MJ, McCormick F. AKT inacti-

vation causes persistent drug tolerance to EGFR inhibitors. Pharmacol Res. 2015;102:132–137.