frontiers in Genetics

ORIGINAL RESEARCH
published: 13 July 2017
doi: 10.3389/fgene.2017.00096

Quantifying Gene Regulatory Relationships with Association Measures: A Comparative Study

Zhi-Ping Liu *

Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, China

In this work, we provide a comparative study of the main available association measures for characterizing gene regulatory strengths. Detecting the association between genes (as well as RNAs, proteins, and other molecules) is very important to decipher their functional relationship from genomic data in bioinformatics. With the availability of more and more high-throughput datasets, the quantification of meaningful relationships by employing association measures will make great sense of the data. There are various quantitative measures have been proposed for identifying molecular associations. They are depended on different statistical assumptions, for different intentions, as well as with different computational costs in calculating the associations in thousands of genes. Here, we comprehensively summarize these association measures employed and developed for describing gene regulatory relationships. We compare these measures in their consistency and specificity of detecting gene regulations from both simulation and real gene expression profiling data. Obviously, these measures used in genes can be easily extended in other biological molecules or across them.

Keywords: gene regulatory network, gene coexpression, association measure, high-throughput data, bioinformatics

OPEN ACCESS

Edited by:
Shihua Zhang,
Academy of Mathematics and Systems Science (CAS), China

Reviewed by:
Xingming Zhao,
Tongji University, China
Lin Gao,
Xidian University, China

***Correspondence:**
Zhi-Ping Liu
zpliu@sdu.edu.cn

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *18 April 2017*
Accepted: *28 June 2017*
Published: *13 July 2017*

Citation:
Liu Z-P (2017) Quantifying Gene Regulatory Relationships with Association Measures: A Comparative Study. Front. Genet. 8:96. doi: 10.3389/fgene.2017.00096

INTRODUCTION

The high-throughput technologies, such as microarray (Schena et al., 1995) and RNA-Seq (Wang et al., 2009) in transcriptomic level, generate bunch of data of describing various perspectives of cell state. These data provide unprecedented opportunity to quantify molecular expressions and their relationships. From a systematic perspective, the molecules in a cell orchestrate together to form various integrated and condense network systems of performing comprehensive functions (Liu, 2015). For instance, transcriptional interactions between transcription factor (TF) and target genes are often formulated into gene regulatory network of modeling biological processes (Liu et al., 2014, 2015). Deciphering gene relationships from high-throughput data are crucial to reversely engineer their inner interaction scenarios, as well as profoundly reveal the dysfunctions in certain disorders, such as complex diseases (Liu et al., 2012).

Quantifying the relationship between molecular components becomes fundamental in the new research paradigm from data to knowledge. The data analysis techniques of association support the kind of investigation. Traditionally, when we explore the relationship between two variables, Pearson's correlation coefficient (PCC) is employed to qualify their linear relationship (Zou et al., 2003). From entropy aspects, mutual information (MI) is often used for defining the non-linear relationship between gene variables (Butte and Kohane, 2000). Mathematically, the assumptions underlying these measures are considerable in real applications. Association measures have been

developed to meet the requirements of appropriateness and precision in defining relationships from various perspectives.

Detecting gene associations is a fundamental method to reconstruct gene regulatory network from gene expression profiling data (Liu, 2015). Although more integrated methods such as ordinary differential equations are available to model the differential dynamics among genes, the association-based methods are direct, simple, and easy for interpretation as well. With introducing the independence, these measures have been extended to quantify the associations between many genes simultaneously (Stuart et al., 2003). In typical microarray experiments, the gene expression data can often be represented by matrix **G**,

$$G = \begin{pmatrix} G_1 \\ G_2 \\ \vdots \\ G_m \end{pmatrix} = \begin{pmatrix} G_{11} & \dots & G_{1j} & \dots & G_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ G_{i1} & \dots & G_{ij} & \dots & G_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ G_{m1} & \dots & G_{mj} & \dots & G_{mn} \end{pmatrix}.$$

Where G_{ij} represents the gene expression value of the i-th gene ($1 \leq i \leq m$) in the j-th experiment ($1 \leq j \leq n$). It is noted that j refers to a sample or a time point with specific phenotype meaning. The association between gene X and gene Y ($X, Y \in \{G_1, G_2, \cdots, G_m\}$) is often to indicate their regulatory relationship (Zhang and Horvath, 2005). Let gene expressions be $X = (X_1, X_2, ..., X_n)$ and $Y = (Y_1, Y_2, ..., Y_n)$. Based on the two vectors, we employ or define an association measure to assess their regulatory strength. Recently, some novel measures besides PCC and MI have been proposed to define the association between two variables (Reshef et al., 2011). It is of great interest to investigate their performances in the reconstruction of gene regulatory network from gene expression data. **Figure 1** demonstrates the strategy of inferring gene regulatory network by gene coexpression analysis. Gene regulation, in a particular form of transcriptional regulation, often specifies the regulation from TF to target gene. The quantified gene coexpression evaluates the simultaneous patterns of two gene's redundancy across samples. The expression level of upstream TF's gene is often to approximate its downstream protein product. As shown in **Figure 1C**, if we set up which ones are TFs by prior knowledge in the gene association network, we can infer a directed gene regulatory network via an undirected association measure.

The coexpression pattern between two genes implies their regulatory aspects. As shown in **Figure 1C**, it firstly indicates a direct regulatory interaction. In some biological state, gene coexpression exactly responds to the activation or inhibition regulation from a TF to its target gene. The regulation between them is reflected by their highly-related gene expression redundancy. Secondly, gene coexpression is about gene co-regulation. That is to imply the two genes are regulated by the same TF(s) and then they contain highly-related gene expression patterns. Third means that the two genes are functionally-related by participating in the same regulatory circuit or particular signaling pathway. Generally, the dynamic regulations in a cell are inherently embedded with temporal features. Gene regulation is often reflected by time-delayed gene expression patterns from the activation of TF's gene to the downstream target responds (Bar-Joseph et al., 2012). For the simplicity of association measure, the coexpression-based methods are popular in inferring gene regulatory network from gene expression data (Zhang and Horvath, 2005).

In this paper, we provide a comparative study on these available association measures of quantifying gene relationships in regulatory network. Fourteen most-popular association measures or indices will be summarized and compared. Based on some benchmark datasets of gene regulatory network inference challenges, we evaluate their individual performances in the reconstruction of gene regulatory networks. This provides a concise comparison of accuracy and quality in network inference by the association measures. In a case study, we compare the differences of these inferred regulations during the infection of hepatitis C virus on host cells. In data-driven network inference, the characteristics of the association measures in statistics and computations are also analyzed and discussed.

ASSOCIATION MEASURES

Numerous association measures have been proposed to define the relationship between two random variables. For gene regulations, we collect 14 of them for our assessments of network inference power from data. **Table 1** lists the 14 association measures with brief introduction of their statistical assumptions and fundamental properties individually. Some measures are well-known such as PCC, while some become available recently such as maximal information correlation (MIC). For the completeness of introduction and reference, we describe them in details respectively in this section.

Pearson's Correlation Coefficient

PCC describes the linear relationship between two variables X and Y (Pearson, 1895). In the microarray data of gene expression, it defines the correlation coefficient between gene X and Y as

$$r(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{(n-1) S_X S_Y},$$

where $\bar{X} = \sum_{i=1}^{n} X_i$, $\bar{Y} = \sum_{j=1}^{n} Y_j$ refer to the mean of two variables of gene expression in samples, and $S_X = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$, $S_Y = \sqrt{\frac{\sum_{j=1}^{n} (Y_j - \bar{Y})^2}{n-1}}$ are their standard deviations. Generally, it assesses their linear relationship into a value between -1 and 1, where 1 refers to total positive correlation and -1 refers to total negative correlation, and 0 refers to no correlation.

When we implement the statistical test of its significance, PCC assumes the two variables are from two normal distributions and the two vectors are the corresponding pairs with independence in the observations (Zou et al., 2003). It has been widely used to quantify the gene coexpression relationships in many studies,

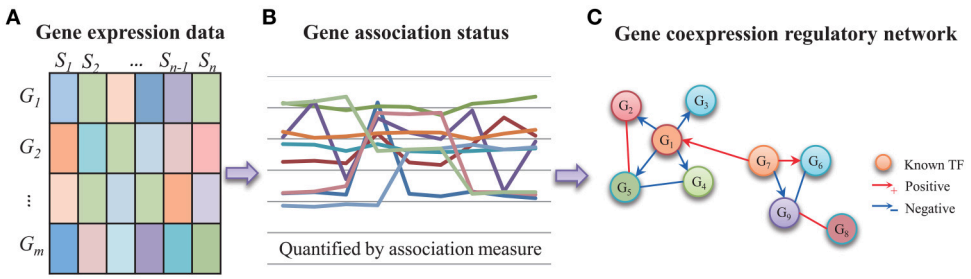

FIGURE 1 | The strategy of building gene coexpression-based regulatory network from gene expression data. **(A)** The gene expression patterns of *m* genes in *n* samples. **(B)** The gene coexpression patterns quantified by association measure. **(C)** With some prior knowledge of TFs, the gene coexpression relationships can be improved to be a gene regulatory network.

TABLE 1 | Summary of some association measures used to quantify gene regulations.

Abbre.	Method	Symbol	Description	References
Pearson	Pearson's	r	Linear, widely-used, no parameter, coeff. ∈ [−1, 1]	Pearson, 1895
Spearman	Spearman's	ρ	Monotonic, rank-based, no parameter, coeff. ∈ [−1, 1]	Spearman, 1904
Kendall	Kendall's	τ	Monotonic, rank-based, no parameter, coeff. ∈ [−1, 1]	Kendall, 1938
Hoeffding	Hoeffding's	D	Non-linear, rank-based, no parameter, coeff. ∈ [0, 1]	Hoeffding, 1948
Blomqvist	Blomqvist's	β	Monotonic, rank-based, no parameter, coeff. ∈ [−1, 1]	Blomqvist, 1950
Goodman	Goodman and Kruskal's	γ	Monotonic, cross classifications, rank-based, no parameter, coeff. ∈ [−1, 1]	Goodman and Kruskal, 1954
WWH	Wang, Waterman, Huang's	*wwh*	Monotonic, rank-based, no parameter, coeff. ∈ [0, +∞]	Wang et al., 2014
MI	Mutual information	I	Non-linear, entropy-based, no parameter, coeff. ∈ [0, +∞]	Shannon, 1948
MIC	Maximum information correlation	*mic*	Non-linear, entropy-based, 1 parameter, coeff. ∈ [0, 1]	Reshef et al., 2011
Wilks	Wilks'	W	Linear, covariance-based, no parameter, coeff. ∈ [0, 1]	Wilks, 1935
KCCA	Kernel canonical correlation analysis	*kcca*	Non-linear, covariance-based, 1 parameter, coeff. ∈ [0, 1]	Bach and Jordan, 2002
dCor	Distance correlation	*dCor*	Non-linear, covariance-based, 1 parameter, coeff. ∈ [0, 1]	Szekely and Rizzo, 2009
CMMD	copula-based maximum mean discrepancy	*cmmd*	Non-linear, copulas-based, 1 parameter, coeff. ∈ [0, 1]	Poczos et al., 2012
RDC	Randomized dependence coefficient	*rdc*	Non-linear, copulas-based, 2 parameters, coeff. ∈ [0, 1]	Lopez-Paz et al., 2013

such as WGCNA (Zhang and Horvath, 2005; Langfelder and Horvath, 2008).

Spearman's Rank Correlation

Spearman's rank correlation ρ is a non-parametric measure of the relationship between two variables (Spearman, 1904). The association between two variables X and Y is formulated as a monotonic function

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

Where $d_i = X_i - Y_i$, $1 \leq i \leq n$. Instead of using the element values directly, it transforms the two vectors to the two rank vectors of these elements respectively. The differential rank vector is generated by the difference between two rank vectors.

When there are no repeated values in X and Y (no duplicated ranks), ρ reaches 1 and −1 when a variable is a perfect monotone function of the other variable. The statistical independence between them refers to $\rho = 0$. In the statistical test, it still requires the dependence between the two ranking of two variables (Zar, 1972). Compared to PCC, it contains a larger application scope because it does not require the normal distribution assumptions. It is equivalent to PCC between two ranked variables (Conover and Iman, 1981). The following non-linear rank-based correlations contain the similar properties.

Kendall's Tau Coefficient

Similar to the former coefficients, Kendall's tau coefficient (Kendall, 1938) is another measure of rank correlation between X and Y. It is defined as

$$\tau = \frac{n_c - n_d}{n(n-1)/2},$$

where $n_c = \#(concordant\ pairs)$ and $n_d = \#(discordant\ pairs)$. Any pair of observations (X_i, Y_i) and (X_j, Y_j) in X and Y, where $i \neq j$, are defined as concordant if the ranks for both elements agree, i.e., if both $X_i > X_j$ and $Y_i > Y_j$ or if both $X_i < X_j$ and $Y_i < Y_j$. They are classified to be discordant if $X_i > X_j$ and $Y_i < Y_j$ or if $X_i < X_j$ and $Y_i > Y_j$. If $X_i = X_j$ or $Y_i = Y_j$, the pair

is neither concordant nor discordant. Based on τ, Somers' D of Y with respect to X is defined as $D_{YX} = \tau(X,Y)/\tau(X,X)$, where $\tau(X,X)$ is the number of pairs with unequal values (Somers, 1962). It is easy to find that the order of ranks in the two variables plays critical roles in the calculation of these non-parametric estimators.

Hoeffding's Dependence Coefficient

The original idea of Hoeffding's dependence measure D is to assess the independence of two datasets by their distance between distributions for continuous variables (Hoeffding, 1948). It has been extended for the samples of X and Y as

$$D = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)},$$

where $D_1 = \sum_i (Q_i - 1)(Q_i - 2)$, $D_2 = \sum_i (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$ and $D_3 = \sum_i (R_i - 2)(S_i - 2)(Q_i - 1)$, R_i is the rank of X_i, S_i is the rank of Y_i, and Q_i is the bivariate rank, which refers to the number of points with both X and Y values less than the ith point, i.e., $Q_i = \#(X_j, Y_j)\ s.t.\ X_j < X_i\ and\ Y_j < Y_i$.

Blomqvist's β

A measure referred as Blomqvist's β has been developed for the medial correlation coefficient (Blomqvist, 1950). For two random variables X and Y, let "$x - y$"-plane be divided into four regions by the median lines of $\tilde{x}$ and $\tilde{y}$. The relationship of X and Y can be obtained from the number of sample points in the four quadrants. In gene regulations, suppose the sample size takes even number (with minor modifications in odd number), it is defined as

$$\beta = \frac{n_1 - n_2}{n_1 + n_2} = \frac{2n_1}{n_1 + n_2} - 1,$$

where n_1 refers to the number of data in the first or third quadrant, and n_2 refers to that in the second or fourth quadrant. It has some advantages such as its explicit form and low computational complexity in estimation (Blomqvist, 1950).

Goodman and Kruskal's Gamma Coefficient

The Goodman and Kruskal's γ coefficient (Goodman and Kruskal, 1954) is another widely-used rand-based coefficient to measure the dependence between variables. It is defined as

$$\gamma = \frac{P_s - P_d}{P_s + P_d},$$

where P_s, P_d are the probabilities that a randomly selected pair of observations will relocate in the same or opposite order respectively, when ranked by both variables. It represents the symmetric distances between the two paired sets representing the binary relation of ranks. It is very close to Kendall's tau. In gene samples, its maximum likelihood estimation can be regarded as

$$G = \frac{n_s - n_d}{n_s + n_d},$$

where n_s is the number of concordant pairs, which refer to those pairs ranked in the same order one both variables. n_d is the number of discordant pairs, which are the number of pairs of cases ranked in reversed order. It computes the normalized difference between the numbers of concordant and discordant pairs such that it will take values between -1 and $+1$. When it is specified into 2×2 matrices, it is exactly Yule's Q coefficient (Yule, 1900).

WWH Order Correlation

The order statistics seems to provide a robust gene coexpression measure by taking local patterns in gene expression profiles into account. Wang, Huang, and Waterman (WWH; Wang et al., 2014) proposed a count statistics method to define a new gene coexpression regulatory measure, i.e.,

$$wwh = \sum_{1 \le i_1 < \cdots < i_k \le n} F(X_{i_1}, ..., X_{i_k}; Y_{i_1}, ..., Y_{i_k}).$$

Where $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$ are genes X and Y with expression levels from n samples. The function F is an indicator function comparing the rank patterns of the two subsequences with a length parameter k. This method aims to identify the consistency of rank orders of the two variables and expect to highlight the local corresponding features in expression profiles. The authors considered a special case in the time-series samples by constraining the consecutive subsequences and another general cases of samples (Wang et al., 2014).

Mutual Information

Mutual information is based on information theory (Shannon, 1948). Suppose $P(X,Y)$ is the joint probability distribution function of gene variables of X and Y, and $P(X)$ and $P(Y)$ are their marginal probability distribution functions respectively. The mutual information between X and Y is defined as

$$I = -\sum_{X_i \in X, Y_j \in Y} P(X_i, Y_j) \log \frac{P(X_i, Y_j)}{P(X_i)P(Y_j)}.$$

The mutual information can also be represented as a Kullback–Leibler divergence (Kullback and Leibler, 1951), which is to measure of the difference between two probability distributions.

Maximal Information Correlation

Based on mutual information, MIC is defined to evaluate the margin probability by calculating the data point frequencies (Reshef et al., 2011), i.e.,

$$MIC = \max_{|X_i||Y_j| < B} \frac{I(X,Y)}{\log_2(\min(|X_i|, |Y_j|))},$$

where (X_i) and (Y_j) are the two gene expressions across the samples individually. I refers to their mutual information. The B is a heuristically setting parameter such as $B = N^{0.6}$, and N is the cells of a grid G induced by X and Y.

Wilks' *W*

Wilks' *W* statistic is the covariance-based measure of two vectors (Wilks, 1935). It is defined as

$$W = 1 - \frac{\det(\Sigma)}{\det(\Sigma_{11})\det(\Sigma_{22})},$$

where $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, and $\Sigma_{ij} = \mathrm{cov}(X_i, Y_j)$. It has close relationship with likelihood-ratio and multivariate analysis of variance (MANOVA) by integrating the covariances of two individual variables and their combinations. Similarly, Pillai's trace criterion performs similar ideas while with low popularity (Pillai, 1955). Here, it is a special case only for two gene expression vectors.

Kernel Canonical Correlation Analysis

Instead of directly calculating the relationship between X and Y, the canonical correlation analysis (CCA) is a statistical technique of maximizing the correlation between sets of projections of the two original vectors.

Let $U = a^T X$, $V = b^T Y$, $Var(U) = a^T \sum_{11} a$, $Var(V) = b^T \sum_{22} b$, $Cov(U, V) = a^T \sum_{12} b$,

where $\Sigma = Var(X, Y) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, $\Sigma_{11} = Var(X)$, $\Sigma_{22} = Var(Y)$, $\Sigma_{12} = Var(X, Y)$, $\Sigma_{21} = Var(Y, X)$.

So

$$Cor(U, V) = \frac{a^T \Sigma_{12} b}{\sqrt{a^T \Sigma_{11} a}\sqrt{b^T \Sigma_{22} b}}.$$

We define the largest canonical correlation as $\rho_1 = \sup_{a,b} Cor(U, V)$, where we set the second floor as a fix number. When we maximize the first floor by solving an optimization problem is to achieve the largest canonical correlation coefficient between the original X and Y.

In CCA, the vector of U and V are linear combinations of X and Y. When

$$K_X = \sum\nolimits_i \Phi(X_i)^T \Phi(X_i),$$
$$K_Y = \sum\nolimits_i \Phi(Y_i)^T \Phi(Y_i),$$

where $\Phi : \mathbb{R}^n \to \mathbb{R}^N (n \le N)$ is the kernel function of X and Y (can be different for them).

$$Cor(U, V) = \frac{\alpha^T K_X K_Y \beta}{\sqrt{\alpha^T K_X K_Y \alpha}\sqrt{\alpha^T K_X K_Y \beta}},$$

and the kernel CCA is defined as $kcca(X, Y) = \sup_{\alpha,\beta} Cor(U, V)$.

Distance Correlation

Let (X_i, Y_i), $1 \le i \le n$ be statistical samples for two random variables (X, Y). The pairwise distances are

$$a_{j,k} = \left\| X_j - X_k \right\|, j, k = 1, 2, ..., n,$$
$$b_{j,k} = \left\| Y_j - Y_k \right\|, j, k = 1, 2, ..., n,$$

where $\|\cdot\|$ denotes Euclidean norm, Then, two $n \times n$ distance matrices $(a_{j,k})$ and $(b_{j,k})$ are generated. For each element (j, k), two transformed values are defined as

$$A_{j,k} = a_{j,k} - \bar{a}_{j,\cdot} - \bar{a}_{\cdot,k} + \bar{a}_{\cdot,\cdot},$$
$$B_{j,k} = b_{j,k} - \bar{b}_{j,\cdot} - \bar{b}_{\cdot,k} + \bar{b}_{\cdot,\cdot},$$

where $\bar{a}_{j,\cdot}$ is the j-th row mean, $\bar{a}_{\cdot,k}$ is the k-th column mean, and $\bar{a}_{\cdot,\cdot}$ is the grand mean of the distance matrix of the X samples. The notations for b values have the similar meanings. The distance covariance is defined as the square root of

$$V_{XY}^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{i,j} B_{i,j}.$$

Then, distance correlation (dCor; Szekely and Rizzo, 2009) between X and Y is defined as the square root of

$$dCor = R^2 = \frac{V_{XY}^2}{V_X V_Y}.$$

dCor satisfies $0 \le R \le 1$, and $R = 0$ when X and Y are independent.

Copula-Based Maximum Mean Discrepancy

A copula is a multivariate probability distribution function defined on the unit hypercube with known uniform marginals (Nelsen, 2006). It is popular in high-dimensional statistics for describing the relationships between variables. Specifically, the copula of two random gene variables X and Y is defined as a function

$$C(U, V) = C(F_X(x), F_Y(y)) = F_{XY}(x, y),$$

where $F_X(x) = P(X \le x)$, $F_Y(x) = P(Y \le y)$, and $F_{XY}(x, y) = P(X \le x, Y \le y)$ are the two marginal distributions and the joint distributions (Sklar, 1959).

cMMD is a copula-based kernel association measure between random variables (Poczos et al., 2012). It extends the maximum mean discrepancy (MMD) method (Borgwardt et al., 2006) of measuring dependence to the copula of the joint distribution. Suppose two copulas transformations have been implemented on the original variables, i.e., $U = F_1(X)$ and $V = F_2(Y)$, F_1 and F_2 are the empirical cumulative distribution functions for X and Y respectively (Lopez-Paz et al., 2013). cMMD defines the relationship between X and Y as

$$cmmd\,(X, Y) = mmd\left[F_1(X), F_2(Y)\right] = \frac{1}{n(n-1)} \sum_{i \ne j}^{n} K(U_i, V_j),$$

where $K(U_i, V_j) = \Phi(U_i, U_j) + \Phi(V_i, V_j) - \Phi(U_i, V_j) - \Phi(U_j, V_i)$, and Φ is a specified kernel function, e.g., Gaussian kernel.

Randomized Dependence Coefficient

Based on the former kernel CCA and copulas, the randomized dependence coefficient (RDC) provides a computationally efficient association measures between multivariate random variables. In details, it is defined as

$$rdc(X, Y; k, s) = \sup_{\alpha,\beta} Cor\left\{\alpha^T \Phi\left[F_1(X); k, s\right],\ \beta^T \Phi\left[F_2(Y); k, s\right]\right\},$$

where the functions are the same as the former ones, $k \in \mathbb{N}^+$ and $s \in \mathbb{R}^+$ are the parameters which are often set as 20 and 0.6 respectively. RDC is proved to be capable of discovering a wide range of functional association patterns in multiple datasets.

RESULTS OF COMPARISON STUDY

For a comparative study of these association measures in inferring gene regulatory relationships, we test these association measures in DREAM3 *in silico* network challenge datasets (Marbach et al., 2010). In the challenges, gene expression datasets have been generated by some specified network structures. Then, the datasets are open without any information about the network structures. The task is to reconstruct the network structures from the open datasets by developing new inference methods. There are three sizes of networks with 10, 50, and 100 nodes respectively, and multiple datasets for each size (4 for 10-node network, 23 for 50-node network, and 46 for 100-node network). The assessment is to evaluate the consistency between the inferred network and the true network structure (gold standards). **Figure 2** illustrate the receiver operating characteristic (ROC) curves of inference performance by these association measures in the 10-node benchmark network. Due to the undirected regulations identified by all these association measures, we omit the regulatory directions when calculating the evaluation metrics of sensitivity (SN), specificity (SP), accuracy (ACC), Matthews correlation coefficient (MCC), F-measure, and area under ROC curve (AUC). **Table 2** demonstrates these detailed values of evaluation metrics of these association measures. We find KCCA performs the best in the 14 association measures for inferring 10-node networks and it reaches the AUC of 0.623 ± 0.083 (mean $\pm$ standard deviation). Overall, the performances of these methods are comparable with each other in the 10-node network.

For the association measures, it becomes more difficult to achieve high inference performances when the network size becomes bigger from 10, 50 to 100. Although each association measure cannot achieve good inferences for big networks, the performances of them decrease with the same tendency. For 50-node networks, mutual information (MI) achieves the best AUC of 0.569 ± 0.046. Blomqvist's β performs the best for 100-node networks in the inference, while it is not stable for the small-size networks. **Figure 3** shows the ranks of their performances according to the mean AUCs in different size of networks individually. From the comparative study, mutual information (MI) performs relatively better with stable ranks for big networks with 50 and 100 nodes. PCC is also stable in the 14 association measures for various sizes of network, as well as KCCA and dCor. This indicates their relative reliability in detecting gene regulatory relationships from expression data. For the other association measures, they accomplish unreliable and unstable regulatory network inferences in the benchmarks.

From the inference performances, we find that most of association-based methods can only achieve limited accuracies in the reconstruction of gene regulatory network from the benchmark datasets, especially for large-size networks. The application scopes of these association measures are mainly determined by the assumptions and characteristics of their definitions listed in **Table 1**. For instances, PCC is for linear regulatory relationship, MI is for non-linear relationship, KCCA and dCor measure the genuine relationship based on covariance, and the rank-based associations are robust

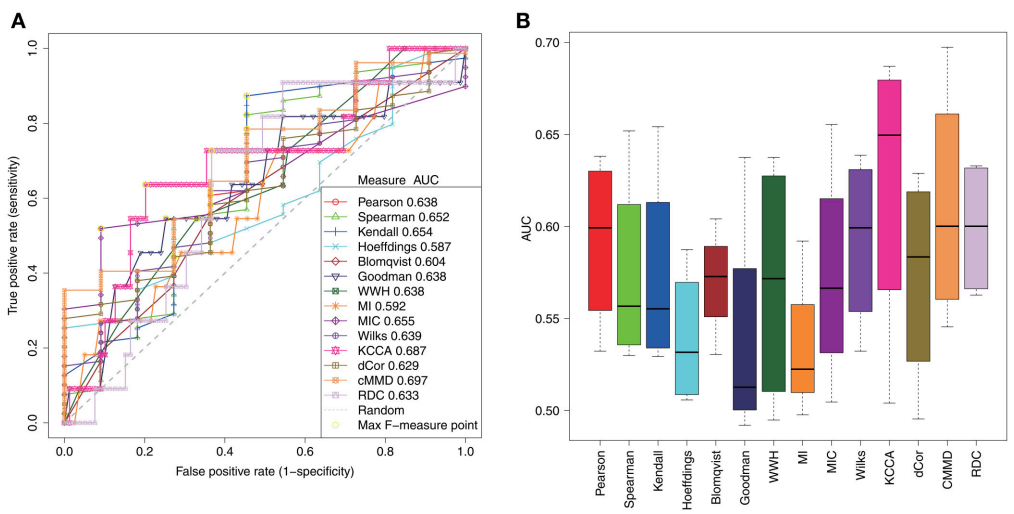

FIGURE 2 | The performances of different association measures in the inference of the 10-node regulatory network of DREAM challenges. **(A)** ROC curve of 14 association measures with maximum AUC in the four datasets. **(B)** Blox plots of AUC of 14 association measures.

TABLE 2 | The performance details of inferring benchmark gene regulatory networks by 14 association measures.

Methods	Node size	SN	SP	ACC	F-measure	MCC	AUC
Pearson	10	0.500 ± 0.093	0.545 ± 0.166	0.506 ± 0.098	0.518 ± 0.121	0.030 ± 0.162	0.592 ± 0.048
	50	0.536 ± 0.102	0.510 ± 0.121	0.535 ± 0.099	0.507 ± 0.074	0.014 ± 0.044	0.554 ± 0.027
	100	0.531 ± 0.047	0.487 ± 0.078	0.530 ± 0.046	0.504 ± 0.048	0.004 ± 0.021	0.536 ± 0.021
Spearman	10	0.617 ± 0.162	0.477 ± 0.155	0.600 ± 0.150	0.526 ± 0.141	0.074 ± 0.191	0.574 ± 0.055
	50	0.511 ± 0.083	0.504 ± 0.071	0.510 ± 0.081	0.502 ± 0.059	0.005 ± 0.036	0.538 ± 0.031
	100	0.501 ± 0.055	0.506 ± 0.086	0.501 ± 0.053	0.497 ± 0.043	0.002 ± 0.019	0.533 ± 0.025
Kendall	10	0.601 ± 0.192	0.500 ± 0.117	0.589 ± 0.175	0.536 ± 0.125	0.082 ± 0.198	0.574 ± 0.057
	50	0.499 ± 0.098	0.518 ± 0.083	0.500 ± 0.095	0.498 ± 0.053	0.005 ± 0.034	0.536 ± 0.031
	100	0.509 ± 0.054	0.503 ± 0.085	0.509 ± 0.053	0.499 ± 0.040	0.003 ± 0.017	0.532 ± 0.025
Hoeffdings	10	0.519 ± 0.591	0.591 ± 0.091	0.528 ± 0.080	0.544 ± 0.042	0.073 ± 0.062	0.539 ± 0.039
	50	0.507 ± 0.072	0.494 ± 0.102	0.507 ± 0.070	0.492 ± 0.064	0.00006 ± 0.038	0.544 ± 0.032
	100	0.504 ± 0.071	0.523 ± 0.061	0.504 ± 0.069	0.508 ± 0.042	0.006 ± 0.018	0.535 ± 0.025
Blomqvist	10	0. 563 ± 0.069	0.409 ± 0.189	0.544 ± 0.060	0.451 ± 0.136	−0.019 ± 0.125	0.570 ± 0.030
	50	0.457 ± 0.126	0.496 ± 0.134	0.458 ± 0.120	0.444 ± 0.069	−0.016 ± 0.028	0.535 ± 0.030
	100	0.550 ± 0.066	0.583 ± 0.056	0.551 ± 0.065	0.560 ± 0.020	0.030 ± 0.008	0.574 ± 0.022
Goodman	10	0.411 ± 0.130	0.500 ± 0.053	0.422 ± 0.073	0.437 ± 0.073	−0.063 ± 0.073	0.539 ± 0.067
	50	0.470 ± 0.086	0.454 ± 0.083	0.469 ± 0.082	0.448 ± 0.037	−0.0246 ± 0.0194	0.531 ± 0.026
	100	0.531 ± 0.068	0.529 ± 0.059	0.531 ± 0.067	0.524 ± 0.027	0.014 ± 0.011	0.527 ± 0.018
WWH	10	0.411 ± 0.248	0.591 ± 0.174	0.433 ± 0.200	0.416 ± 0.148	−0.006 ± 0.103	0.569 ± 0.069
	50	0.352 ± 0.116	0.660 ± 0.099	0.360 ± 0.111	0.437 ± 0.083	0.003 ± 0.019	0.532 ± 0.016
	100	0.392 ± 0.137	0.619 ± 0.145	0.395 ± 0.134	0.442 ± 0.070	0.003 ± 0.009	0.522 ± 0.018
MI	10	0.557 ± 0.149	0.409 ± 0.241	0.539 ± 0.111	0.416 ± 0.115	−0.022 ± 0.111	0.534 ± 0.041
	50	0.470 ± 0.100	0.443 ± 0.081	0.470 ± 0.098	0.448 ± 0.069	−0.028 ± 0.043	0.569 ± 0.046
	100	0.468 ± 0.081	0.471 ± 0.069	0.468 ± 0.079	0.462 ± 0.046	−0.014 ± 0.020	0.544 ± 0.034
MIC	10	0.500 ± 0.051	0.636 ± 0.196	0.517 ± 0.042	0.547 ± 0.066	0.090 ± 0.121	0.573 ± 0.062
	50	0.515 ± 0.120	0.494 ± 0.084	0.515 ± 0.116	0.492 ± 0.070	0.003 ± 0.044	0.551 ± 0.031
	100	0.510 ± 0.058	0.502 ± 0.071	0.510 ± 0.057	0.501 ± 0.038	0.003 ± 0.017	0.531 ± 0.024
Wilks	10	0.522 ± 0.113	0.477 ± 0.087	0.517 ± 0.109	0.498 ± 0.098	0.0004 ± 0.13	0.592 ± 0.048
	50	0.536 ± 0.102	0.509 ± 0.120	0.536 ± 0.099	0.507 ± 0.073	0.014 ± 0.044	0.554 ± 0.027
	100	0.523 ± 0.050	0.502 ± 0.080	0.523 ± 0.049	0.508 ± 0.048	0.006 ± 0.021	0.538 ± 0.025
KCCA	10	0.472 ± 0.267	0.432 ± 0.202	0.467 ± 0.231	0.393 ± 0.168	−0.067 ± 0.219	0.623 ± 0.083
	50	0.442 ± 0.121	0.464 ± 0.119	0.442 ± 0.117	0.428 ± 0.070	−0.031 ± 0.037	0.541 ± 0.058
	100	0.453 ± 0.100	0.502 ± 0.090	0.454 ± 0.098	0.462 ± 0.058	−0.011 ± 0.024	0.541 ± 0.036
dCor	10	0.506 ± 0.061	0.545 ± 0.166	0.511 ± 0.069	0.520 ± 0.102	0.034 ± 0.140	0.573 ± 0.060
	50	0.529 ± 0.084	0.513 ± 0.103	0.529 ± 0.082	0.512 ± 0.067	0.014 ± 0.042	0.556 ± 0.031
	100	0.514 ± 0.061	0.510 ± 0.091	0.514 ± 0.060	0.505 ± 0.049	0.006 ± 0.021	0.538 ± 0.025
CMMD	10	0.573 ± 0.201	0.545 ± 0.129	0.569 ± 0.176	0.540 ± 0.112	0.085 ± 0.164	0.611 ± 0.066
	50	0.508 ± 0.081	0.491 ± 0.088	0.508 ± 0.079	0.494 ± 0.065	−0.00006 ± 0.041	0.547 ± 0.031
	100	0.512 ± 0.071	0.505 ± 0.068	0.512 ± 0.070	0.503 ± 0.044	0.004 ± 0.019	0.532 ± 0.028
RDC	10	0.522 ± 0.147	0.568 ± 0.227	0.528 ± 0.139	0.527 ± 0.143	0.062 ± 0.203	0.599 ± 0.038
	50	0.518 ± 0.085	0.522 ± 0.076	0.518 ± 0.083	0.515 ± 0.076	0.013 ± 0.039	0.551 ± 0.032
	100	0.517 ± 0.070	0.515 ± 0.042	0.517 ± 0.069	0.051 ± 0.04	0.007 ± 0.018	0.534 ± 0.026

to the noisy and outliers in gene expressions. In practical applications, the selection of suitable association measures could be subjectively determined by research purpose, experimental design, phenotypic condition and data quality. An ensemble and self-adaptive association measures selection strategy is desirable to be proposed for the co-existence of different gene regulatory relationships.

In real microarray data, we perform our comparative study of quantifying gene regulations during hepatitis C virus (HCV) infection on host Huh7 cells. The gene expression data are downloaded from NCBI GEO (accession ID GSE20948) (Edgar et al., 2002). There are 28 samples of 14 HCV infected Huh7 hepatoma cell samples and 14 corresponding mock-infected samples, originally designed three replicates at 6, 12, 18, 24, and 48 h post-infections, respectively. Two samples at 6 h have not been enrolled after quality control. The details can be accessed from Ref. (Blackham et al., 2010). We also download the hepatocellular carcinoma (HCC) gene set from KEGG (Kanehisa and Goto, 2000). The gene set contains 123 genes with 94 genes containing their expression profiles in GSE20948 (Edgar et al., 2002).

For evaluating the inference consistency of these association measures, we calculate the pairwise gene regulatory strengths in the HCC genes by the 14 association measures respectively. In the results of each association measure, the pairs with the top 5% association values are regarded as the identified gene regulations in the context of specific gene expression profiles after HCV infection.

Figure 4 demonstrates the inferred gene coexpression regulatory network in the HCC genes by PCC. There is no information about direction, so we annotate the known human TFs and display them by different color nodes (cyan) with the other genes (green). From **Figure 4**, we can figure out the regulatory information about positive and negative relationships during HCV infection. As in the former comparisons, we compare the overlapping status of these inferred coexpression relationships by the four association measures with top performances, i.e., Pearson, MI, KCCA and dCor. There exists only one pair of genes ("IFNA1" and "IFNA13") is identified by the four measures, and the relationship between the two genes can be detected by any of them. Interestingly, Pearson and dCor contain many overlaps (177 regulations). It provides direct evidence that dCor is mainly to extract the linear correlations between genes as that Pearson done in this case study. There are few overlaps (3 regulations) between Pearson and MI, which indicates the linear and non-linear information are inconsistent with each other, and different association measures might identify different gene associations. The selection of suitable association measures is again proved to be very important for inferring gene coexpression regulatory network. The few overlapping regulations also imply the complex and diversity of regulatory relationships underlying gene expressions. More advanced methods beyond association measures are urged for elucidating gene regulatory mechanism from high-throughput data. See Section Discussion for some already available methods.

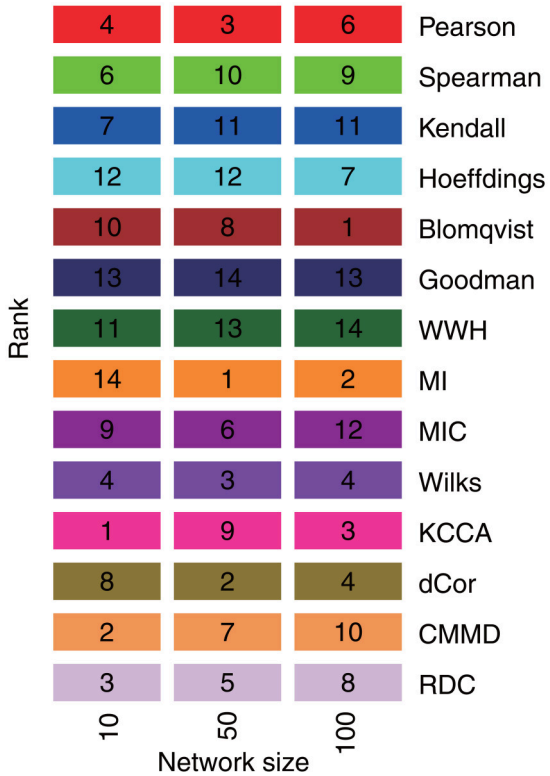

FIGURE 3 | The ranks of 14 association measures in the inferences of regulatory networks with different node sizes. The numbers in the color blocks refer to the ranks of corresponding association measures by the means of AUC in these benchmark networks.

DISCUSSION

It is known association is different from causality and correlation does not imply causation (Altman and Krzywinski, 2015). Detecting the causality between genes has been essential in gene regulatory network inference since the availability of high-throughput data (Opgen-Rhein and Strimmer, 2007). Gene association network indicates more general gene-gene relationship than regulation, and gene regulatory network indicates more general gene-gene relationship than causality. The gene causality network, that is to say, the causal regulations between genes are directed in the gene-gene interaction graph with the detailed information of which ones are upstream regulators, and which ones are downstream targets. In the direct regulations, TFs or signal transductors causally affect their target gene expressions. The information flow transits between genes will be revealed if a causal relationship exists. So far, there is no association measure has been defined for describing the causal relationship between genes (Zhang et al., 2014; Zhao et al., 2016), while more advanced methods based on conditional probability, model-based regression and differential equation have been proposed to address the evaluations of causality.

Based on conditional independence, some improved association measures, such as partial correlation coefficient

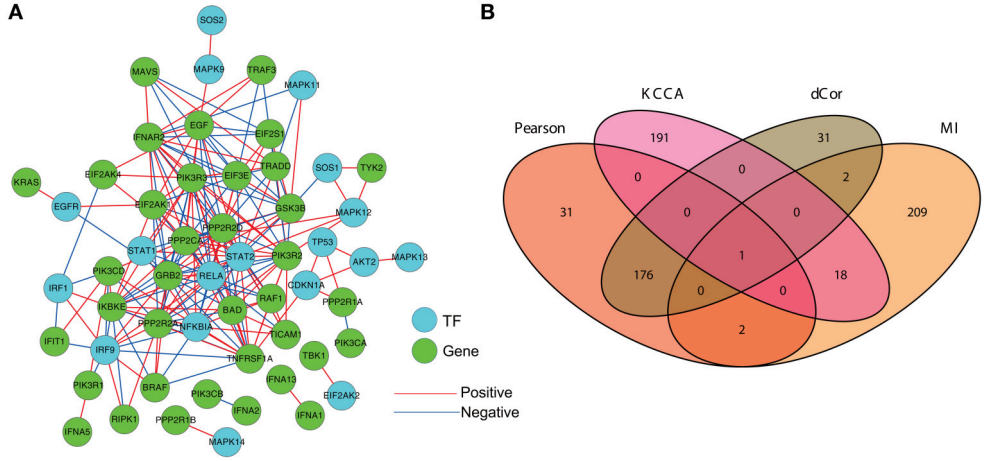

FIGURE 4 | The reconstructed gene coexpression regulatory network during HCV infection. **(A)** The gene association network constructed by the PCC-based method. Isolated genes are not shown. **(B)** The overlapping status of the inferred gene regulations by four association measures, i.e., Pearson, MI, KCCA, and dCor.

and conditional mutual information, have been proposed to eliminate false positive regulations from gene associations. The original association measures generate the footholds for detecting genuine relationships. Conditioning on another gene or gene set Z, partial correlation measure $r_{XY \cdot Z}$ between gene X and Y is to access the exact correlation between X and Y and that has no relationship with Z (de la Fuente et al., 2004). It is defined as

$$r_{XY \cdot Z} = \frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^2)(1 - r_{YZ}^2)}}.$$

Where r refers to PCC. In the similar philosophy of introducing other gene or gene set, the conditional mutual information (CMI; Liang and Wang, 2008) is defined as

$$I(X_i, Y_j | Z_k) = \sum_{X_i \in X,\ Y_j \in Y,\ Z_k \in Z} p(X_i, Y_j, Z_k) \log \frac{p(X_i, Y_j | Z_k)}{p(X_i | Z_k) p(Y_j | Z_k)}.$$

Based on CMI and the order of conditioned gene numbers, we proposed a gene regulatory inference method named PCA-CMI (Zhang et al., 2012, 2013), which detect out dedicate associations by removing undirect false positive regulations. For a pair of genes X and Y, Li proposed a conditional coexpression measure named liquid association (LA) between two genes by introducing a third gene Z (Li, 2002). Based on Z, the gene relationship of X and Y is defined as

$$LA(XY|Z) = E(XY|Z) = \sum_i \frac{X_i Y_i Z_i}{n}$$

where n is the sample size. The LA activity determines the functional associations of gene X and Y in the condition of Z.

Currently, the causality between genes is often quantified via Bayesian models (Friedman et al., 2000). According to data, the conditional probability of $P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$. The probability of gene X conditioned on gene Y, means Y have a causal effect on X because there exists a negative or positive values of the conditional probability. The structured model has been extended and formulated as diagrams using a graphical criterion known as d-separation (Bareinboim and Pearl, 2016). Bayesian network provides a model-based detection of causal regulatory relationships. Gene regulations are then identified from the graphical models (Liu et al., 2013).

Regression and other structured models often extract the effects of regulatory coefficients. The identification of model coefficients determines the global relationship of these individual genes (D'Haeseleer et al., 1999). Specifically, the regression models the response gene as the linear combinations of the other dependent genes, i.e., $Y = c_0 + c_1 X_1 + c_2 X_2 + \cdots + c_m X_m + \varepsilon$, m is the number of dependent genes in the regression and ε is the error variable. In generalized linear models, the response gene is changed to $\theta(Y)$, and $X_1, \cdots, X_m$ are replaced by $\phi_1(X_1), \cdots, \phi_m(X_m)$, respectively (Breiman and Friedman, 1985). In the special case of simple linear regression with $m = 1$, the model is to detect the linear relationship between the response gene and the only one dependent gene. The coefficient of determination denoted by r^2 is equal to the square of PCC (Altman and Krzywinski, 2016). The coefficient of determination, which represents the proportion of variation due to their linear relationship, generalizes the correlation coefficient for relationships beyond simple linear regression. Often, the regression equations often model the associations between response genes and dependent genes in an inter-coupled system. From a system biology perspective, regression models consider the genes in an integrated manner. Compared to the former pairwise associations, they identify more complicated relationships among genes. After determining the coefficients, the relationships in these genes are quantified correspondingly. How to determine crucial regulators and targets via statistical variable selections techniques, such as lasso (Tibshirani, 1996)

and elastic net (Zou and Trevor, 2005), are substantially important.

Similarly, ODE models the derivatives, i.e., $\frac{dY}{dt} = c_0 + c_1X_1 + c_2X_2 + \cdots + c_mX_m$, and so ODE quantifies the dynamics of the response as a function of the dependents in the system (Wu et al., 2014). The expression change rate of a response gene is modeled by the expressions of dependence genes. The Y might be another dependence gene and thus the system is closed. The system identification is to evaluate the coefficients in the right-hand side of the equation and the coefficient values refer to gene regulatory strengths. When the coefficient is 0, there is no relationship between the responding gene and the depending gene, otherwise the regulatory strength can be represented by positive or negative numeric values.

Compared to association measure, regression model and differential equation model regard gene regulatory network as an integrative system. The gene regulatory network inference is then transformed to a system identification problem of solving the coupled equations. The gene regulation strengths refer to the identified coefficients. From a sequential modeling perspective, the causality between regulators and targets can also be reflected by these system biology techniques.

In machine learning techniques such as clustering (Rui and Wunsch, 2005), there are some metrics have been developed for measuring the association between data points. The distances of Euclidean, cosine, Hamming, Manhattan are often used to measure gene relationships in gene expression clustering (D'Haeseleer, 2005). These distances evaluate the differences including dependences between genes, while these compared association measures focus on quantifying gene relationship such as regulation between genes. In gene expression data analyses of clustering and feature selection, distance metrics provide alternatives to define gene similarities. The distance metrics are not included in the comparative study for their diversity and case-intensity (Santini and Jain, 1999).

CONCLUSIONS

In this paper, we summarized and compared the main proximities and metrics for quantifying gene regulatory associations. Written in full, the definitions and descriptions of 14 association measures are summarized and their characteristics with applications in regulatory network inference have been presented. From the benchmark challenge data and real gene expression data, we compared their performances and consistencies in the network inferences. Furthermore, their advantages and limitations are also analyzed and discussed. Currently, developing causality measure is an urgent research topic from driving gene association to regulation causality (Bareinboim and Pearl, 2016). A powerful measure of causality will greatly benefit the discovery of important gene regulations. Moreover, the linear/non-linear regression and differential equation models regard many genes in dynamic systems and the parameters of these models represent the system in details. The model-based gene regulatory network inference methods seem to provide more powerful tools when compared to the association-based methods. However, the association measures contain their flexibility in sense, easy interpretation and large scope of applications.

In conclusion, gene association measures provide fundamental quantifications of detecting gene regulatory relationships from transcriptomic profiling data. The high-throughput technologies advance the measurements of thousands of genes in parallel manners. The association measures effectively accelerate the transformation processes from data to knowledge. Most of the proposed association measures are statistical techniques which focus only on the inter-relationships between genes, and they are very hard to get the causal gene relationships alone. With the improved conditional or joint association measures, such as partial correlation coefficient, conditional mutual information and liquid association, the causality between genes can be partially extracted out from data. The introduction of other genes in evaluating gene regulation provides promising alternatives to grasp the genuine regulations. For an entire system, many genes perform their functions coordinately and cooperatively. So more advanced models are extremely needed to describe the complex system of gene regulations. In such model as ODE, the time-varying regulations are exactly to quantify the gene regulatory interactions with temporal implications. For the model complexity and the data availability, the dynamics underlying the coefficients in regression and ODE will reveal much more complicated regulatory relationships.

AUTHOR CONTRIBUTIONS

ZL conceived and designed the study. ZL wrote the code and analyzed the data. ZL drafted the manuscript.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61572287 and 61533011; Natural Science Foundation of Shandong Province, China (ZR2015FQ001); the Fundamental Research Funds of Shandong University under Grant Nos. 2015QY001 and 2016JC007; the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China. The paper was also funded by a Pilot Research Grant from School of Control Science and Engineering at Shandong University.

REFERENCES

Altman, N., and Krzywinski, M. (2015). Points of significance: association, correlation and causation. *Nat. Methods* 12, 899–900. doi: 10.1038/nmeth.3587

Altman, N., and Krzywinski, M. (2016). Points of significance: simple linear regression. *Nat Methods* 12, 999–1000. doi: 10.1038/nmeth.3627

Bach, F. R., and Jordan, M. I. (2002). Kernel independent component analysis. *J. Mach. Learn. Res.* 3, 1–48. doi: 10.1109/ICASSP.2003.1202783

Bareinboim, E., and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proc. Natl. Acad. Sci. U.S.A.* 113, 7345–7352. doi: 10.1073/pnas.1510507113

Bar-Joseph, Z., Gitter, A., and Simon, I. (2012). Studying and modelling dynamic biological processes using time-series gene expression data. *Nat. Rev. Genet.* 13, 552–564. doi: 10.1038/nrg3244

Blackham, S., Baillie, A., Al-Hababi, F., Remlinger, K., You, S., Hamatake, R., et al. (2010). Gene expression profiling indicates the roles of host oxidative stress, apoptosis, lipid metabolism, and intracellular transport genes in the replication of hepatitis C virus. *J. Virol.* 84, 5404–5414. doi: 10.1128/JVI.02529-09

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Ann. Math. Stat.* 21, 593–600. doi: 10.1214/aoms/1177729754

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H. P., Scholkopf, B., and Smola, A. J. (2006). Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* 22, e49–e57. doi: 10.1093/bioinformatics/btl242

Breiman, L., and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* 80, 580–598. doi: 10.1080/01621459.1985.10478157

Butte, A. J., and Kohane, I. S. (2000). Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. *Pac. Symp. Biocomput.* 5, 418–429. doi: 10.1142/9789814447331_0040

Conover, W. J., and Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *Am. Stat.* 35, 124–129. doi: 10.1080/00031305.1981.10479327

de la Fuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* 20, 3565–3574. doi: 10.1093/bioinformatics/bth445

D'Haeseleer, P. (2005). How does gene expression clustering work? *Nat. Biotechnol.* 23, 1499–1501. doi: 10.1038/nbt1205-1499

D'Haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. *Pac. Symp. Biocomput.* 4, 41–52. doi: 10.1142/9789814447300_0005

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *J. Comput. Biol.* 7, 601–620. doi: 10.1089/106652700750050961

Goodman, L. A., and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Am. Stat. Assoc.* 49, 732–764. doi: 10.1080/01621459.1954.10501231

Hoeffding, W. (1948). A non-parametric test of independence. *Ann. Math. Stat.* 19, 546–557. doi: 10.1214/aoms/1177730150

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika* 30, 81–93. doi: 10.1093/biomet/30.1-2.81

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat.* 22, 79–86. doi: 10.1214/aoms/1177729694

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Li, K. C. (2002). Genome-wide coexpression dynamics: theory and application. *Proc. Natl. Acad. Sci. U.S.A.* 99, 16875–16880. doi: 10.1073/pnas.252466999

Liang, K. C., and Wang, X. (2008). Gene regulatory network reconstruction using conditional mutual information. *EURASIP J. Bioinform. Syst. Biol.* 2008:253894. doi: 10.1155/2008/253894

Liu, Z. P. (2015). Reverse engineering of genome-wide gene regulatory networks from gene expression data. *Curr. Genomics* 16, 3–22. doi: 10.2174/1389202915666141110210634

Liu, Z. P., Wang, Y., Zhang, X. S., and Chen, L. (2012). Network-based analysis of complex diseases. *IET Syst. Biol.* 6, 22–33. doi: 10.1049/iet-syb.2010.0052

Liu, Z. P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database* 2015:bav095. doi: 10.1093/database/bav095

Liu, Z. P., Wu, H., Zhu, J., and Miao, H. (2014). Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. *BMC Bioinformatics* 15:336. doi: 10.1186/1471-2105-15-336

Liu, Z. P., Zhang, W., Horimoto, K., and Chen, L. (2013). Gaussian graphical model for identifying significantly responsive regulatory networks from time course high-throughput data. *IET Syst. Biol.* 7, 143–152. doi: 10.1049/iet-syb.2012.0062

Lopez-Paz, D., Hennig, P., and Scholkopf, B. (2013). The randomized dependence coefficient. *Adv. Neural Inf. Process. Syst.* 26, 1–9.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. U.S.A.* 107, 6286–6291. doi: 10.1073/pnas.0913357107

Nelsen, R. B. (2006). *An Introduction to Copulas*. New York, NY: Springer-Verlag.

Opgen-Rhein, R., and Strimmer, K. (2007). From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Syst. Biol.* 1:37. doi: 10.1186/1752-0509-1-37

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* 58, 240–242. doi: 10.1098/rspl.1895.0041

Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. *Ann. Math. Stat.* 26, 117–121. doi: 10.1214/aoms/1177728599

Poczos, B., Ghahramani, Z., and Schneider, J. (2012). "Copula-based kernel dependency measures," in *Proceedings of International Conference on Machine Learning* (Edinburgh), 775–782.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., et al. (2011). Detecting novel associations in large data sets. *Science* 334, 1518–1524. doi: 10.1126/science.1205438

Rui, X., and Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Trans. Neural Netw.* 16, 645–678. doi: 10.1109/TNN.2005.845141

Santini, S., and Jain, R. (1999). Similarity measures. *IEEE Trans. Pattern Anal. Mach. Intell.* 21, 871–883. doi: 10.1109/34.790428

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470. doi: 10.1126/science.270.5235.467

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Universite de Paris* 8, 229–231.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *Am. Sociol. Rev.* 27, 799–811. doi: 10.2307/2090408

Spearman, C. C. (1904). The proof and measurement of association between two things. *Am. J. Psychol.* 15, 72–101. doi: 10.2307/1412159

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255. doi: 10.1126/science.1087447

Szekely, G. J., and Rizzo, M. L. (2009). Brownian distance covariance. *Ann. Appl. Stat.* 1236–1265. doi: 10.1214/09-AOAS312

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

Wang, Y. X., Waterman, M. S., and Huang, H. (2014). Gene coexpression measures in large heterogeneous samples using count statistics. *Proc. Natl. Acad. Sci. U.S.A.* 111, 16371–16376. doi: 10.1073/pnas.1417128111

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi: 10.1038/nrg2484

Wilks, S. S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica* 3, 309–326. doi: 10.2307/1905324

Wu, S., Liu, Z. P., Qiu, X., and Wu, H. (2014). Modeling genome-wide dynamic regulatory network in mouse lungs with influenza infection using high-dimensional ordinary differential equations. *PLoS ONE* 9:e95276. doi: 10.1371/journal.pone.0095276

Yule, G. U. (1900). On the association of attributes in statistics: with illustrations from the material of the childhood society, & c. *Philos. Trans. R. Soc. Lond. Ser. A* 194, 257–319. doi: 10.1098/rsta.1900.0019

Zar, J. H. (1972). Significance testing of the spearman rank correlation coefficient. *J. Am. Stat. Assoc.* 67, 578–580. doi: 10.1080/01621459.1972.10481251

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4:17. doi: 10.2202/1544-6115.1128

Zhang, X., Liu, K., Liu, Z. P., Duval, B., Richer, J. M., Zhao, X. M., et al. (2013). NARROMI: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference. *Bioinformatics* 29, 106–113. doi: 10.1093/bioinformatics/bts619

Zhang, X., Zhao, J., Hao, J. K., Zhao, X. M., and Chen, L. (2014). Conditional mutual inclusive information enables accurate quantification of associations in gene regulatory networks. *Nucleic Acids Res.* 43, e31. doi: 10.1093/nar/gku1315

Zhang, X., Zhao, X. M., He, K., Lu, L., Cao, Y., Liu, J., et al. (2012). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics* 28, 98–104. doi: 10.1093/bioinformatics/btr626

Zhao, J., Zhou, Y., Zhang, X., and Chen, L. (2016). Part mutual information for quantifying direct associations in networks. *Proc. Natl. Acad. Sci. U.S.A.* 113, 5130–5135. doi: 10.1073/pnas.1522586113

Zou, H. H., and Trevor, H. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. B* 67, 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

Zou, K. H., Tuncali, K., and Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology* 227, 617–628. doi: 10.1148/radiol.2273011499

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.