BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparison of co-expression measures: mutual information, correlation, and model based indices

Lin Song[1,2], Peter Langfelder[1] and Steve Horvath[1,2*]

Abstract

Background: Co-expression measures are often used to define networks among genes. Mutual information (MI) is often used as a generalized correlation measure. It is not clear how much MI adds beyond standard (robust) correlation measures or regression model based association measures. Further, it is important to assess what transformations of these and other co-expression measures lead to biologically meaningful modules (clusters of genes).

Results: We provide a comprehensive comparison between mutual information and several correlation measures in 8 empirical data sets and in simulations. We also study different approaches for transforming an adjacency matrix, e.g. using the topological overlap measure. Overall, we confirm close relationships between MI and correlation in all data sets which reflects the fact that most gene pairs satisfy linear or monotonic relationships. We discuss rare situations when the two measures disagree. We also compare correlation and MI based approaches when it comes to defining co-expression network modules. We show that a robust measure of correlation (the biweight midcorrelation transformed via the topological overlap transformation) leads to modules that are superior to MI based modules and maximal information coefficient (MIC) based modules in terms of gene ontology enrichment. We present a function that relates correlation to mutual information which can be used to approximate the mutual information from the corresponding correlation coefficient. We propose the use of polynomial or spline regression models as an alternative to MI for capturing non-linear relationships between quantitative variables.

Conclusion: The biweight midcorrelation outperforms MI in terms of elucidating gene pairwise relationships. Coupled with the topological overlap matrix transformation, it often leads to more significantly enriched co-expression modules. Spline and polynomial networks form attractive alternatives to MI in case of non-linear relationships. Our results indicate that MI networks can safely be replaced by correlation networks when it comes to measuring co-expression relationships in stationary data.

Background

Co-expression methods are widely used for analyzing gene expression data and other high dimensional "omics" data. Most co-expression measures fall into one of two categories: correlation coefficients or mutual information measures. MI measures have attractive information-theoretic interpretations and can be used to measure non-linear associations. Although MI is well defined for discrete or categorical variables, it is non-trivial to estimate the mutual information between quantitative variables, and corresponding permutation tests can be computationally intensive. In contrast, the correlation coefficient and other model based association measures are ideally suited for relating quantitative variables. Model based association measures have obvious statistical advantages including ease of calculation, straightforward statistical testing procedures, and the ability to include additional covariates into the analysis. Researchers trained in statistics often measure gene co-expression by the correlation coefficient. Computer scientists, trained in information theory, tend to use a mutual information (MI) based measure. Thus far, the

*Correspondence: SHorvath@mednet.ucla.edu
[1]Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, California, USA
[2]Biostatistics, School of Public Health, University of California, Los Angeles, California, USA

majority of published articles use the correlation coefficient as co-expression measure [1-5] but hundreds of articles have used the mutual information (MI) measure [6-12].

Several articles have used simulations and real data to compare the two co-expression measures when clustering gene expression data. Allen et al. have found that correlation based network inference method WGCNA [5] and mutual information based method ARACNE [9] both perform well in constructing global network structure [13]; Steuer et al. show that mutual information and the Pearson correlation have an almost one-to-one correspondence when measuring gene pairwise relationships within their investigated data set, justifying the application of Pearson correlation as a measure of similarity for gene-expression measurements [14]. In simulations, no evidence could be found that mutual information performs better than correlation for constructing co-expression networks [15]. However, MI continues to be used in recent publications. Some authors have argued that MI is more robust than Pearson correlation in terms of distinguishing various clustering solutions [10]. Given the debates, it remains an open question whether mutual information could be supplanted by standard model based association measures. We affirmatively answer this question by i) reviewing the close relationship between mutual information and likelihood ratio test statistic in the case of categorical variables, ii) finding a close relationship between mutual information and correlation in simulations and empirical studies, and iii) proposing polynomial and spline regression models as alternatives to mutual information for modeling non-linear relationships.

While previous comparisons involved the Pearson correlation, we provide a more comprehensive comparison that considers i) different types of correlation coefficients, e.g. the biweight midcorrelation (bicor), ii) different approaches for constructing MI based and correlation based networks, iii) different ways of transforming a network adjacency matrix (e.g. the topological overlap reviewed below [4,16-18]), and iv) 8 diverse gene expression data from yeast, mouse and humans. Our unbiased comparison evaluates co-expression measures at the level of gene pair relationships and at the level of forming co-expression modules (clusters of genes).

This article presents the following results. First, probably the most comprehensive empirical comparison to date is used to evaluate which pairwise association measure leads to the biologically most meaningful network modules (clusters) when it comes to functional enrichment with GO ontologies. Second, polynomial regression and spline regression methods are evaluated when it comes to defining non-linear association measures between gene pairs. Third, simulation studies are used to validate a functional relationship (cor-MI function) between correlation and mutual information in case that the two variables satisfy a linear relationship. Our comprehensive empirical studies illustrate that the cor-MI function can be used to approximate the relationship between mutual information and correlation in case of real data sets which indicates that in many situations the MI measure is not worth the trouble. Gene pairs where the two association measures disagree are investigated to determine whether technical artifacts lead to the incongruence.

Overall, we find that bicor based co-expression measure is an attractive co-expression measure, particularly when limited sample size does not permit the detection of non-linear relationships. Our theoretical results, simulations, and 8 different gene expression data sets show that MI is often inferior to correlation based approaches in terms of elucidating gene pairwise relationships and identifying co-expression modules. A signed correlation network transformed via the topological overlap matrix transformation often leads to the most significant functional enrichment of modules. Polynomial and spline regression model based statistical approaches are promising alternatives to MI for measuring non-linear relationships.

Association measure and network adjacency

An association measure is used to estimate the relationships between two random variables. For example, correlation is a commonly used association measure. There are different types of correlations. While the Pearson correlation, which measures the extent of a linear relationship, is the most widely used correlation measure, the following two more robust correlation measures are often used. First, the Spearman correlation is based on ranks, and measures the extent of a monotonic relationship between x and y . Second, "bicor" (refer to Materials and Methods for definition and details) is a median based correlation measure, and is more robust than the Pearson correlation but often more powerful than the Spearman correlation [19,20]. All correlation coefficients take on values between -1 and 1 where negative values indicate an inverse relationship. A correlation coefficient is an attractive association measure since i) it can be easily calculated, ii) it affords several asymptotic statistical tests (regression models, Fisher transformation) for calculating significance levels (p-values), and iii) the sign of correlation allows one to distinguish between positive and negative relationships. Other association measures, such as mutual information, will be introduced in the next sections.

Association measures can be transformed into network adjacencies. For n variables $v_1, \ldots, v_n$, an adjacency matrix $A = (A_{ij})$ is an $n \times n$ matrix quantifying the pairwise connection strength between variables.

An (undirected) network adjacency satisfies the following conditions:

$$\begin{aligned} 0 \le A_{ij} &\le 1, \\ A_{ij} &= A_{ji}, \\ A_{ii} &= 1. \end{aligned} \quad (1)$$

An association network is defined as a network whose nodes correspond to random variables and whose adjacency matrix is based on the association measure between pairs of variables [21]. Association networks describe the pair wise associations between variables (interpreted as nodes). For a given set of nodes, there is a one-to one relationship between the association network and the adjacency matrix. In order to build an association network for n variables $v = (v_1, \ldots, v_n)$, we start by defining an association measure $AssocMeasure(x, y)$ as a real valued function of two vectors x , y . We then apply this function on the set of $N = n^2$ variable pairs $\{Pair_1 = (v_1, v_1), Pair_2 = (v_1, v_2), \ldots, Pair_N = (v_N, v_N)\}$, resulting in an $n \times n$ dimensional matrix

$$S = (AssocMeasure(v_i, v_j)). \quad (2)$$

Then, one needs to specify how the association matrix S is transformed into an adjacency matrix. This involves three steps: 1) symmetrize S ; 2) transform (and/or threshold) S to $[0, 1]$; 3) set diagonal values to 1. As for step 1, many methods can be used to symmetrize S if it is non-symmetric, such as the following three ways:

$$S_{ij}^{min} = min(S_{ij}, S_{ji}) \quad (3)$$

$$S_{ij}^{ave} = \frac{S_{ij} + S_{ji}}{2} \quad (4)$$

$$S_{ij}^{max} = max(S_{ij}, S_{ji}). \quad (5)$$

As for step 2, if $LowerBounds(S)$ and $UpperBounds(S)$ denote symmetric matrices of element-wise lower and upper bounds for S , then a simple transformation can be defined as:

$$A = \left(\frac{S - LowersBound(S)}{UpperBounds(S) - LowerBound(S)} \right)^{\beta}, \quad (6)$$

where the power β is constant and denotes a soft threshold. As an example, assume that the association measure is given by a correlation coefficient, i.e. $S = (cor(\boldsymbol{x}_i, \boldsymbol{x}_j))$. Since each correlation has the lower bound -1 and upper bound $+1$, Eq. 6 reduces to the case of a signed weighted correlation network given by [4,22]:

$$A_{ij} = \left(\frac{1 + cor(\boldsymbol{x}_i, \boldsymbol{x}_j)}{2} \right)^{\beta}. \quad (7)$$

Additional details of correlation based adjacencies (unweighted or weighted, unsigned or signed) are described in Materials and Methods .

Network adjacency based on co-expression measures

When dealing with gene expression data, x_i denotes the expression levels of the i-th gene (or probe) across multiple samples. In this article, we assume that the m components of x_i correspond to random independent samples. Co-expression measures can be used to define co-expression networks in which the nodes correspond to genes. The adjacencies A_{ij} encode the similarity between the expression profiles of genes i and j . In practice, transformations such as the topological overlap measure (TOM) [4,16-18] are often used to turn an original network adjacency matrix into a new one. Details of TOM transformation are reviewed in Materials and Methods .

Mutual information networks based on categorical variables

Assume two random samples dx and dy of length m from corresponding discrete or categorical random variables DX and DY . Each entry of dx equals one of the following R levels $ldx_1, \ldots, ldx_R$. The mutual information (MI) is defined as:

$$MI(dx, dy) = \sum_{r=1}^{R_x} \sum_{c=1}^{R_y} p(ldx_r, ldy_c) \log \left(\frac{p(ldx_r, ldy_c)}{p(ldx_r) p(ldy_c)} \right) \quad (8)$$

where $p(ldx_r)$ is the frequency of level r of dx , and log is the natural logarithm. Note that the following **simple relationship exists between the mutual information (Eq. 8) and the likelihood ratio test statistic** (described in Additional file 1):

$$MI(dx, dy) = \frac{LRT.statistic(dx, dy)}{2m} \quad (9)$$

This relationship has many applications. First, it can be used to prove that the mutual information takes on non-negative values. Second, it can be used to calculate an asymptotic p-value for the mutual information. Third, it points to a way for defining a mutual information measure that adjusts for additional conditioning variables $z_1, z_2, \ldots$ Specifically, one can use a multivariate *multinomial regression model* for regressing dy on dx and the conditioning variables. Up to a scaling factor of $2m$, the likelihood ratio test statistic can be interpreted as a (non-symmetric) measure of mutual information between dx and dy that adjusts for conditioning variables. More detailed discussion of mutual information can be found in [14,23,24]. In Additional file 1, we describe association measures between categorical variables in detail, including LRT statistic and MI.

As discussed below, numerous ways have been suggested for construct an adjacency matrix based on MI. Here we describe an approach that results in a

weighted adjacency matrix. Consider n categorical variables $dx_1, dx_2, \ldots, dx_n$. Their mutual information matrix $MI(dx_i, dx_j)$ is a similarity matrix S whose entries are bounded from below by 0. To arrive at an upper bound, we review the relationship between mutual information and entropy (the following equation is text book knowledge):

$$MI(dx, dy) = Entropy(dx) + Entropy(dy) - Entropy(dx, dy) \quad (10)$$

where $Entropy(dx)$ denotes the entropy of dx and $Entropy(dx, dy)$ denotes the joint entropy (refer to Additional file 1). Using Eq. 10, one can prove that the mutual information has the following 3 upper bounds:

$$MI(dx, dy) \leq min(Entropy(dx), Entropy(dy)), \quad (11)$$

$$MI(dx, dy) \leq \frac{Entropy(dx) + Entropy(dy)}{2}, \quad (12)$$

$$MI(dx, dy) \leq max(Entropy(dx), Entropy(dy)). \quad (13)$$

Using Eq. 6 with $\beta = 1$, lower bounds of 0 and $UpperBounds_{ij} = (Entropy(dx_i) + Entropy(dx_j))/2$ (Eq. 12) results in the *symmetric uncertainty based mutual information adjacency matrix*:

$$A_{ij}^{MI,SymmetricUncertainty} = \frac{2MI(dx_i, dx_j)}{Entropy(dx_i) + Entropy(dx_j)}. \quad (14)$$

A transformation of $A^{MI,SymmetricUncertainty}$ leads to the *universal mutual information based adjacency matrix version 1* (denoted AUV1):

$$A_{ij}^{MI,UniversalVersion1} = \frac{A_{ij}^{MI,SymmetricUncertainty}}{2 - A_{ij}^{MI,SymmetricUncertainty}} \quad (15)$$

One can easily prove that $0 \leq A_{ij}^{MI,UniversalVersion1} \leq 1$. The term "universal" reflects the fact that the adjacency based dissimilarity $dissMI_{ij}^{UniveralVersion1} = 1 - A^{MI,UniversalVersion1}$ turns out to be a universal distance function [25]. Roughly speaking, the universality of $dissMI_{ij}^{UniveralVersion1}$ implies that any other distance measure between dx_i and dx_j will be small if $dissMI_{ij}^{UniveralVersion1}$ is small. The term "distance" reflects the fact that $dissMI^{UniveralVersion1}$ satisfies the properties of a distance including the triangle inequality.

Another adjacency matrix is based on the upper bound implied by inequality 13. We define the *universal mutual information based adjacency matrix version 2*, or AUV2, as follows:

$$A^{MI,UniversalVersion2} = \frac{MI(dx_i, dx_j)}{max(Entropy(dx_i), Entropy(dx_j))}. \quad (16)$$

The name reflects the fact that $dissMI^{UniveralVersion2} = 1 - A^{MI,UniversalVersion2}$ is also a universal distance measure [25].

While $A^{MI,UniversalVersion1}$ and $A^{MI,UniversalVersion2}$ are in general different, we find very high Spearman correlations ($r > 0.9$) between their vectorized versions.

Many alternative approaches exist for defining MI based networks, e.g. ARACNE [9], CLR [26], MRNET [27] and RELNET [6,28] are described in Materials and Methods .

Mutual information networks based on discretized numeric variables

In its original inception, the mutual information measure was only defined for discrete or categorical variables, see e.g. [23]. It is challenging to extend the definition to *quantitative* variables. But, several strategies have been proposed in the literature [7,28,29]. In this article, we will only consider the following approach which is based on discretizing the numeric vector x by using the equal width discretization method. This method partitions the interval $[min(x), max(x)]$ into equal-width bins (subintervals). The vector $discretize(x)$ has the same length as x but its l-th component reports the bin number in which x_l falls:

$$dx_l = discretize(x)_l = r \text{ if} x_l \in bin_r. \quad (17)$$

The number of bins, $no.bins$, is the only parameter of the equal-width discretization method.

In our subsequent studies, we calculate an MI-based adjacency matrix using the following three steps. First, numeric vectors of gene expression profiles are discretized according to the equal-width discretization method with the default number of bins given by $no.bins = \sqrt{m}$. Second, the mutual information $MI_{ij} = MI(discretize(x_i), discretize(x_j))$ is calculated between the discretized vectors based on Eq. 10 and the Miller Madow entropy estimation method (detailed in Additional file 1). Third, the MI matrix is transformed into one of three possible MI-based adjacency matrices: $A^{MI,SymmetricUncertainty}$ (Eq. 14), $A^{MI,UniversalVersion1}$ (Eq. 15), $A^{MI,UniversalVersion2}$ (Eq. 16).

Results

An equation relating $MI(discretize(x), discretize(y))$ to $cor(x, y)$

As described previously, the mutual information $MI(discretize(x), discretize(y))$ between the discretized vectors can be used as an association measure. Note that $MI(discretize(x), discretize(y))$ is quite different from $cor(x, y)$ in the following aspects. First, the estimated mutual information depends on parameter choices, e.g. the number of bins used in the equal-width discretization step for defining $dx = discretize(x)$. Second, the mutual information aims to measure general dependence-relationships while the correlation only measures linear or monotonic relationships. Third, the equations for the two measures are very different. Given these differences, it is surprising that a simple approximate relationship holds

between the two association measures if x, y are samples from a bivariate normal distribution and the equal-width discretization method is used with $no.bins = \sqrt{m}$. Under these assumptions, we will show that $A^{MI,UniversalVersion2}$ can be accurately approximated as follows:

$$A^{MI,UniversalVersion2}(dx, dy) = \frac{MI(dx, dy)}{max(Entropy(dx), Entropy(dy))} \approx F^{cor-MI}(cor(x, y)), \quad (18)$$

where the "cor-MI" function [21]

$$F^{cor-MI}(s) = \frac{\log(1 + \epsilon - s^2)}{\log(\epsilon)}(1 - \omega) + \omega \quad (19)$$

depends on the following two parameters

$$\omega = 0.43m^{-0.30} \quad \epsilon = \omega^{2.2}. \quad (20)$$

In general, one can easily show that $F^{cor-MI}(s)$ is a monotonically increasing function that maps the unit interval [0,1] to [0,1] if the two parameters ω and ϵ satisfy the following relationship

$$0 < \epsilon \leq \omega < 1. \quad (21)$$

Eq. 18 was stated in terms of the Pearson correlation, but it also applies for bicor as can be seen from our simulation studies.

Simulations where *x* and *y* represent samples from a bivariate normal distribution

Here we use simulation studies to illustrate that F^{cor-MI} (Eq. 19) can be used for predicting or approximating $A^{MI,UniversalVersion2}$ from the corresponding correlation coefficients (Eq. 18). Specifically, we simulate 2000 pairs of sample vectors x and y from a bivariate normal distribution. Each pair of vectors x and y is simulated to exhibit different pairwise correlations. Figure 1 shows the relationships of the MI-based adjacency measures with the (observed) Pearson correlation (cor) or biweight midcorrelation (bicor) when each of the vectors contains $m = 1000$ components but the relationship has been confirmed for m ranging from 20 to 10000. As can be seen from Figures (1A, B), the cor-MI function (Eq. 18) with parameters specified in Eq. 20 provides a highly accurate prediction of $A^{MI,UniversalVersion2}$ (Eq. 16) on the basis of $cor(x, y)$ and m . Since x and y are normally distributed, the Pearson correlation and bicor are practically indistinguishable (Figure 1C). Thus, replacing cor by bicor leads to equally good predictions of $A^{MI,UniversalVersion2}$ (Figure 1D). Figure (1E) shows that $A^{MI,UniversalVersion2}$ is practically indistinguishable from $A^{MI,SymmetricUncertainty}$. This suggests that cor-MI function can also be used to predict $A^{MI,SymmetricUncertainty}$ on the basis of the correlation measure. Figure (1F) indicates that $A^{MI,UniversalVersion1}$ and $A^{MI,UniversalVersion2}$ are different from each other but satisfy a monotonically increasing relationship.

Empirical studies involving 8 gene expression data sets

Our simulation results show that both the robust biweight midcorrelation and the Pearson correlation can be used as input of F^{cor-MI} for predicting $A^{MI,UniversalVersion2}$ when the underlying variables satisfy pairwise bivariate normal relationships. However, it is not clear whether F^{cor-MI} can also be used to relate correlation and mutual information in real data applications. In this section, we report 8 empirical studies to study the relationship between MI and the robust correlation measure bicor. To focus the analysis on genes that are likely to reflect biological variation and to reduce computational burden, we selected the 3000 genes with highest variance across the microarray samples for each data set. Description of data sets can be found in Materials and Methods .

We first calculate bicor and $A^{MI,UniversalVersion2}$ for all gene pairs in each data set. The two co-expression measures show strong monotonic relationships in most data sets (Figure 2). Then, we predict $A^{MI,UniversalVersion2}$ from bicor based on F^{cor-MI} (Eq. 18). Our predictions are closely related to true $A^{MI,UniversalVersion2}$ values (Figure 3). These results indicate that most gene pairs satisfy linear relationships in real data applications. Among the 8 data sets, SAFHS shows the strongest association between bicor and $A^{MI,UniversalVersion2}$ (Spearman correlation 0.72) and also gives the most accurate $A^{MI,UniversalVersion2}$ prediction (Pearson correlation 0.92). A possible reason is that the large samples size ($m = 1084$) leads to more accurate estimation of mutual information, thus enhancing the association with bicor and the performance of the prediction function. In contrast, the small sample size ($m = 44$) of the yeast data set adversely affects the calculation of mutual information and hence the prediction performance of F^{cor-MI} . In summary, our examples indicate that for most gene pairs, $A^{MI,UniversalVersion2}$ (Eq. 16) is a monotonic function (cor-MI) of the absolute value of bicor. This finding likely reflects the fact that the vast majority of gene pairs satisfy straight line relationships. This approximation improves with increasing sample size m , possibly reflecting more accurate estimation of mutual information.

Although F^{cor-MI} reveals a close relationship between bicor and $A^{MI,UniversalVersion2}$ for most gene pairs, there are cases where the two association measures strongly disagree. In the following, we present scatter plots to visualize the relationships between pairs of genes where MI found a significant relationship while bicor did not and vice versa. To facilitate a comparison between bicor and MI, we standardized each association measure across pairs, which resulted in the Z scores denoted by $Z.MI_{ij} =$

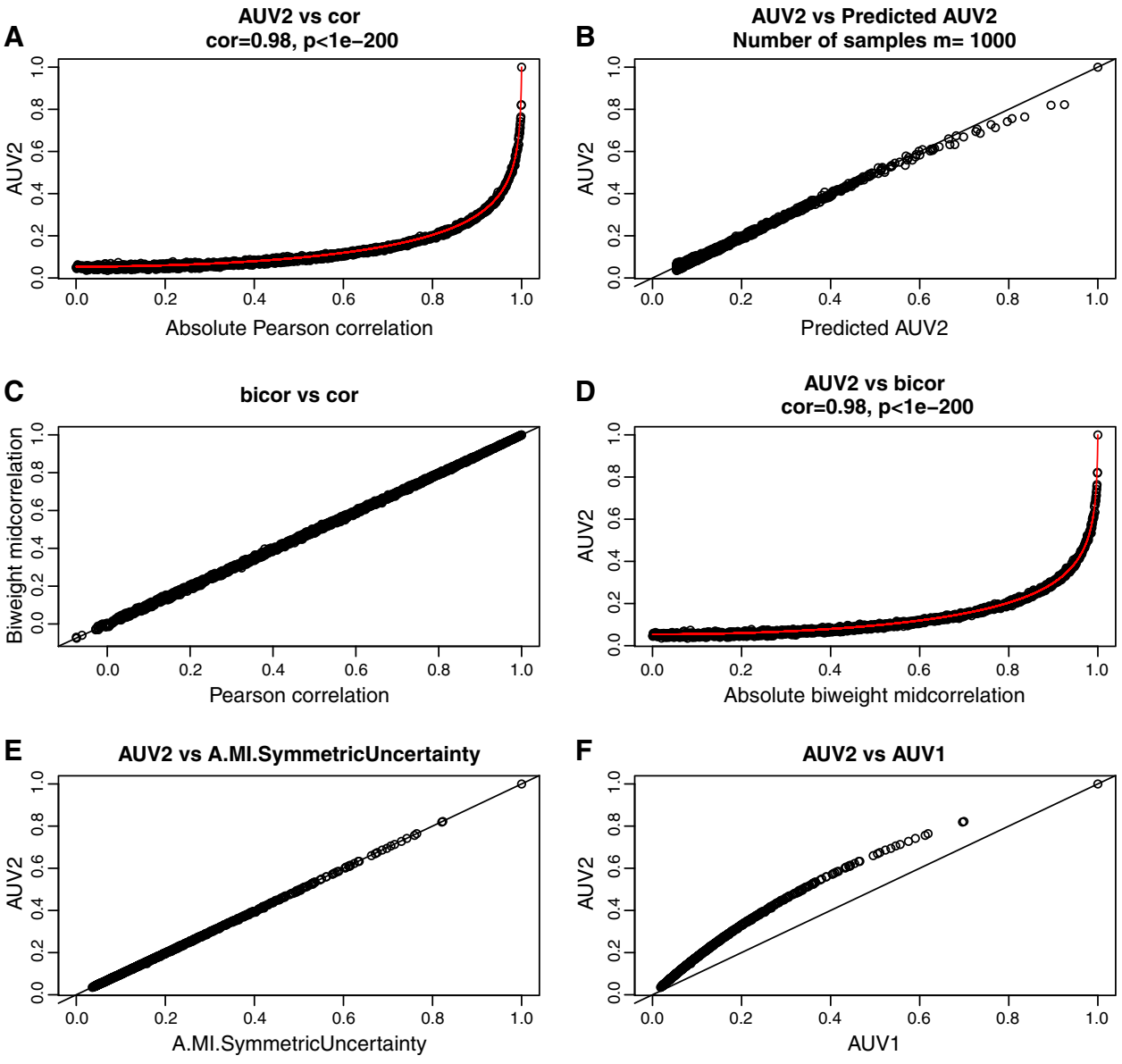

Figure 1 Relating mutual information based adjacencies to the Pearson correlation and biweight midcorrelation in simulation. Each point corresponds to a pair of numeric vectors x and y with length $m = 1000$. These pairs of vectors were simulated to exhibit different correlations. AUV1, AUV2, cor, bicor are abbreviations for $A^{MI,UniversalVersion1}$, $A^{MI,UniversalVersion2}$, Pearson correlation and biweight midcorrelation, respectively. **(A)** MI-based adjacency $A^{MI,UniversalVersion2}$ versus absolute Pearson correlation. Spearman correlation of the two measures and the corresponding p-value are shown at the top, implying a strong monotonic relationship. The red line shows the predicted $A^{MI,UniversalVersion2}$ according to F^{cor-MI} (Eq. 18). Note that the prediction function is highly accurate in simulation. **(B)** Observed $A^{MI,UniversalVersion2}$ versus its predicted value. The straight line has slope 1 and intercept 0. **(C)** Observed Pearson correlation (x-axis) and the corresponding bicor values (y-axis). The straight line has slope 1 and intercept 0. These 2 measurements are practically indistinguishable when x and y are normally distributed. **(D)** $A^{MI,UniversalVersion2}$ versus bicor. Spearman correlation and p-value of the 2 measurements are presented at the top, and predicted $A^{MI,UniversalVersion2}$ are shown as the red line. **(E)** $A^{MI,UniversalVersion2}$ versus $A^{MI.SymmetricUncertainty}$. **(F)** $A^{MI,UniversalVersion2}$ versus $A^{MI,UniversalVersion1}$.

$(MI_{ij} - mean(MI))/\sqrt{(var(MI))}$ and $Z.bicor_{ij} = (bicor_{ij} - mean(bicor))/\sqrt{(var(bicor))}$. Next we selected gene pairs whose value of $Z.MI_{ij}$ was large but $Z.bicor_{ij}$ was low and vice versa. The resulting pairs correspond to the blue and red circles in Figures 2 and 3. To see what dependence patterns drives the discordant behavior of MI and bicor, we used scatter plots to visualize the relationship between the pairs of variables (Figure 4). Gene pairs in Figure (4A) have extreme $A^{MI,UniversalVersion2}$ but insignificant bicor values. Note that the resulting dependencies seem haphazard and may not reflect real biological dependencies. For example, the gene pair in the brain cancer data set

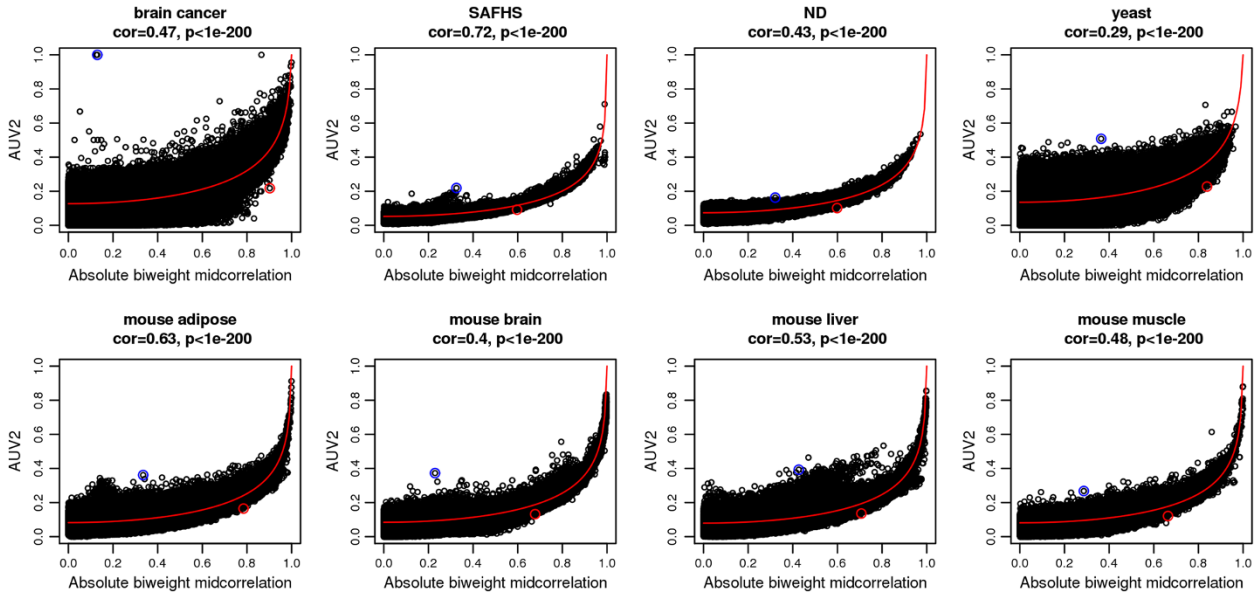

Figure 2 Comparison of correlation and mutual information based co-expression measures in 8 empirical data sets. Absolute value of bicor versus $A^{MI,UniversalVersion2}$ for all probe pairs in each data set. The Spearman correlation and corresponding p-value between the two measures are shown at the top. The two measures show different levels of monotonic relationships in data sets. The red curve predicts $A^{MI,UniversalVersion2}$ from bicor based on Eq. 18. The blue circle highlights the probe pair with the highest $A^{MI,UniversalVersion2}$ z-score among those with insignificant bicor z-scores (less than 1.9); the red circle highlights the probe pair with the highest bicor z-score among those with insignificant $A^{MI,UniversalVersion2}$ z-scores (less than 1.9).

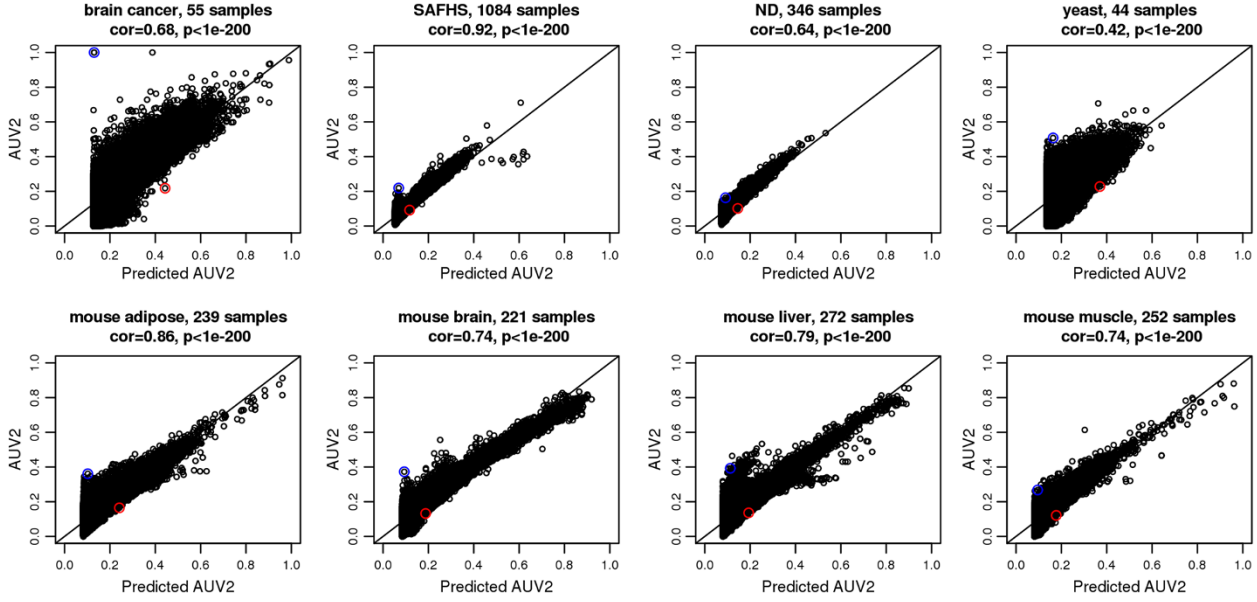

Figure 3 Comparison of predicted and observed $A^{MI,UniversalVersion2}$ in 8 empirical data sets. In all data sets, prediction from bicor based on Eq. 18 and observed $A^{MI,UniversalVersion2}$ are highly correlated (the Pearson correlation and corresponding p-value shown at top). Line y=x is added. Blue and red circles have the same meaning as in Figure 2.

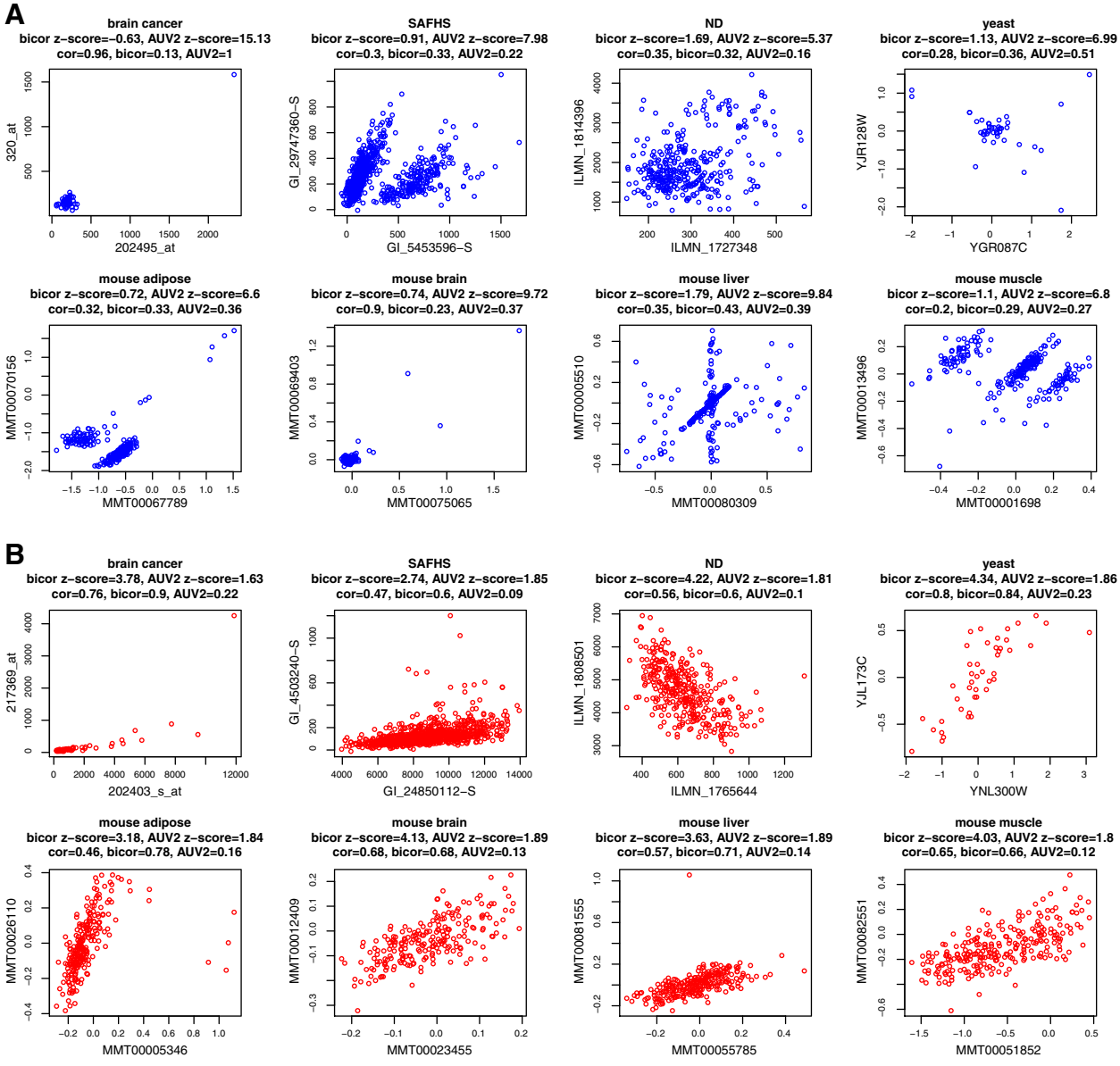

Figure 4 Gene expression of example probe pairs for which the correlation and mutual information based measures disagree. (A) Gene expression of probe pairs highlighted by blue circles in Figure 2. **(B)** Gene expression of probe pairs highlighted by red circles in Figure 2. The Pearson correlation, bicor, $A^{MI,UniversalVersion2}$ values and z-scores of the latter two measures are shown at the top. Mutual information is susceptible to outliers, sometimes detects unusual patterns that are hard to explain, and often misses linear relations that are captured by bicor.

exhibits no clear relationships as correctly implied by bicor, while the significant MI value is driven by an array outlier with extremely high expression for both genes. In the SAFHS data, the gene pair exhibits an unusual pattern that is more likely to be the result of batch effects rather than biological signals. The mouse liver data set displays a pairwise pattern that is neither commonly seen nor easily explained. The ND data set shows no obvious patterns at all, making mutual information less trustworthy. On the contrary, gene pairs with significant value of Z.bicor but insignificant *Z.MI* values show approximate linear relationships in all data sets (Figure 4B). Thus, bicor captures gene pairwise relationships more accurately and sensitively than the mutual information based adjacency $A^{MI,UniversalVersion2}$.

In summary, bicor usually detects linear relationships between gene pairs accurately while mutual information is susceptible to outliers, and sometimes identifies pairs that exhibit patterns unlikely to be of biological origin or that exhibit no clear dependency at all. We note that MI

results tend to be more meaningful when dealing with a large number of observations (say $m > 300$). Although we only consider 3000 genes with highest variances, our results are highly robust with respect to the number of genes. For example, in Additional file 2, we report results when considering all 23568 genes in the mouse adipose data set or considering 10000 randomly selected genes (rather than with high variance) in the ND data set. These results demonstrate that our findings do not depend on the number of genes.

Gene ontology enrichment analysis of co-expression modules defined by different networks

Gene co-expression networks typically exhibit modular structure in the sense that genes can be grouped into modules (clusters) comprised of highly interconnected genes (i.e., within-module adjacencies are high). The network modules often have a biological interpretation in the sense that the modules are highly enriched in genes with a common functional annotation (gene ontology categories, cell type markers, etc) [3,30,31]. In this section, we assess association measures (and network construction methods) by the gene ontology (GO) enrichment of their resulting modules in the 8 empirical data sets.

In order to provide an unbiased comparison, we use the same clustering algorithm for module assignment for all networks. Toward this end, we use a module detection approach that has been used in hundreds of publications: modules are defined as branches of the hierarchical tree that results from using $1 - Adjacency$ as dissimilarity measure, average linkage, and the dynamic tree cutting method [32]. An example of the module detection approach is illustrated in Figure 5. To provide an unbiased evaluation of GO enrichment of each module, we used the *GOenrichmentAnalysis* R function to test enrichment with respect to all GO terms [33,34] and retained the 5 most significant p-values for each module.

The 10 different adjacencies considered here are described in the last 2 columns of Table 1. We first compare modules based on $A^{MI,UniversalVersion2}$ with those resulting from 3 bicor based networks: unsigned adjacency (unsignedA, Eq. 29) , signed adjacency (signedA, Eq. 28) and Topological Overlap Matrix (TOM, Eq. 30) based on signed adjacency. GO enrichment p-values of modules in the 8 real data applications are summarized as barplots in Figure 6. Figure 6 indicates that, in terms of gene ontology enrichment, TOM is the best bicor based gene co-expression network construction method, and it is superior to $A^{MI,UniversalVersion2}$. Note that signed correlation network coupled with the topological overlap transformation exhibit the most significant GO enrichment p-values in all data sets, and the difference is statistically significant ($p < 0.05$) in 6 out of 8 comparisons. The effect of module size is discussed below. An obvious question is whether the performance of MI can be improved when using an alternative MI based network inference method. To address this, we compared the performance of the signed correlation network (with TOM) versus 4 commonly used mutual information: ARACNE, CLR, MRNET and RELNET (described in Materials and Methods). ARACNE allows one to choose a tolerance threshold ϵ ranging from 0 to 1. As ϵ increases, more edges of the ARACNE network will be preserved. We evaluated ARACNE ($\epsilon = 0$), ARACNE ($\epsilon = 0.2$) and ARACNE ($\epsilon = 0.5$) into our comparison. Similarly to Figures 6 and 7 summarizes the GO enrichment p-values of modules in the 8 real data applications. TOM leads to the highest enrichment p-values in 5 cases, and the difference is statistically significant in 4 of them. In two applications, ARACNE ($\epsilon = 0$) performs best, and MRNET performs best in one application. We need to point out that another mutual information based method, maximal information coefficient (MIC) [35], has been proposed recently. Although computational intensive, the MIC has clear theoretical advantages when it comes to capturing general dependence patterns. Additional file 3 compares the performance of MIC with that of TOM when it comes to GO ontology enrichment. TOM clearly outperforms MIC to identify GO enriched modules in 6 out of 7 data sets which may suggest that MIC tends to overfit the data in these applications. SAFHS data set is not included because the computation of MIC was time-consuming on this large data set.

Overall, these unbiased comparisons show that signed correlation networks coupled with the topological overlap transformation outperform the commonly used mutual information based algorithms when it comes to GO enrichment of modules.

Polynomial and spline regression models as alternatives to mutual information

A widely noted advantage of mutual information is that it can detect general, possibly non-linear, dependence relationships. However, estimation of mutual information poses multiple challenges ranging from computational complexity to dependency on parameters and difficulties with small sample sizes. Standard polynomial and spline regression models can also detect non-linear relationships between variables. While perhaps less general than MI, relatively simple polynomial and spline regression models avoid many of the challenges of estimating MI while adequately modeling a broad range of non-linear relationships. In addition to being computationally simpler and faster, regression models also make available standard statistical tests and model fitting indices. Thus, in this section we examine polynomial and spline regression as alternatives to MI for capturing non-linear relationships between gene expression profiles. We define association measures

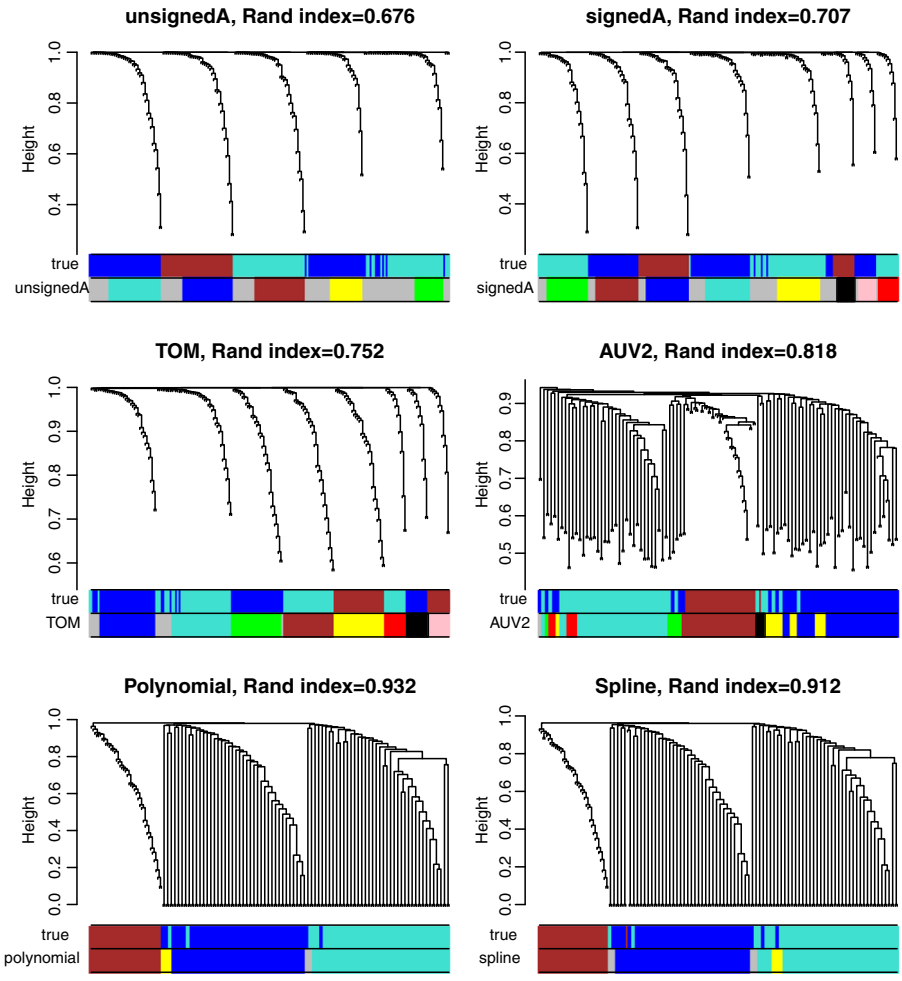

Figure 5 Module identification based on various network inference methods in simulation with non-linear gene-gene relationships. The data set is composed of 200 genes across 200 samples. 3 true modules are designed. Two of them, labeled with colors turquoise and blue, contain linear and non-linear (quadratic) gene-gene relationships. For each adjacency, the clustering tree and module colors are shown. True simulated module assignment is shown by the first color band underneath each tree. On top of each panel is the Rand index between inferred and simulated module assignments.

based on polynomial and spline regression models and study their performance.

Networks based on polynomial and spline regression models

Consider two random variables x and y and the following polynomial regression model of degree 3:

$$E(y|x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3. \quad (22)$$

The model fitting index $R^2(x, y)$ (described in Materials and Methods) can be used to evaluate the fit of the model. One can then reverse the roles of x and y to arrive at a model fitting index $R^2(y, x)$. In general, $R^2(x, y) \neq R^2(y, x)$.

Now consider a set of n variables $x_1, \ldots, x_n$. One can then calculate pairwise model fitting indices $R^2_{ij} = R^2(x_i, x_j)$ which can be interpreted as the elements of an $n \times n$ association matrix (R^2_{ij}). This matrix is in general non-symmetric and takes on values in $[0, 1]$, with diagonal values equal to 1. A large value indicates a close relationship between variables x_i and x_j . To define an adjacency matrix, we symmetrize (R^2_{ij}) through Eqs. 3, 4 or 5.

Spline regression models are also known as local polynomial regression models [36]. Local refers to the fact that these models amount to fitting models on subintervals of the range of x . The boundaries of subintervals are referred to as knots. In analogy to polynomial models, we build natural cubic spline model for all pairs of x_i, x_j. We use the following rule of thumb for the number of knots: if $m > 100$ use 5 knots, if $m < 30$ use 3 knots, otherwise use 4 knots. We then calculate model fitting indices and create corresponding network adjacencies. (Details of

Table 1 Types of networks and characteristics

Network type	Used here	Examples	Variable types	Ease of estimation	Utility for modeling						Adjacencies discussed this article	Used in GO enrichment analysis
					GRN	Reduce	Direct	Time	Nonlin.	Sign		
Correlation network	Yes	WGCNA [5]	Numeric	Easy	Yes	Yes	No	Maybe	No	Yes	unsignedA	Yes
											signedA	Yes
											TOM	Yes
Polynomial or Spline regression network	Yes	WGCNA [5]	Numeric	Moderate	Yes	Yes	No	Maybe	Yes	No	$polyR^2$	No
											$splineR^2$	No
Mutual information network	Yes	ARACNE [9], RELNET [6,28], CLR [26], MRNET [27], MIC [35]	Discretized numeric, categorical	Moderate	Yes	Not clear	No	Maybe	Yes	No	ASU	No
											AUV1	No
											AUV2	Yes
											ARACNE	Yes
											ARACNE0.2	Yes
											ARACNE0.5	Yes
											CLR	Yes
											MRNET	Yes
											RELNET	Yes
											MIC	Yes
Boolean network	No	Boolean network [71]	Dichotomized numeric	Moderate	Yes	Not clear	Yes	Yes	NA	NA	No	No
Probabilistic network	No	Bayesian network [72,73]	Any	Hard	Yes	Not clear	Yes	Yes	Yes	Yes	No	No

For each network method, the table reports what kinds of biological insights can be gained and what kind of data can be analyzed. Column "GRN" indicates whether the network has been (or can be) used for studying gene regulatory networks. Column "Reduce" indicates whether the method has been used for reducing high dimensional data (e.g. via modules and their representatives). Column "Direct" indicates whether the the network can encode directional information. Column "time" indicates whether the network method is suited for studying time series data. Column "Nonlin. " indicates whether the network can capture non-linear relationships between pairs of variables (represented as nodes). Column "Sign" indicates whether the network adjacency provides information on the sign of the relationship between two variables, e.g. a correlation coefficient can take on positive and negative values. The table entry "NA" stands for not applicable. Adjacencies discussed in this article: unsignedA: unsigned bicor; signedA: signed bicor; TOM: TOM transformed signed bicor; ASU: $A^{MI,SymmetricUncertainty}$; AUV1: $A^{MI,UniversalVersion1}$; AUV2: $A^{MI,UniversalVersion2}$; ARACNE: ARACNE, $\epsilon = 0$; ARACNE0.2: ARACNE, $\epsilon = 0.2$; ARACNE0.5: ARACNE, $\epsilon = 0.5$.

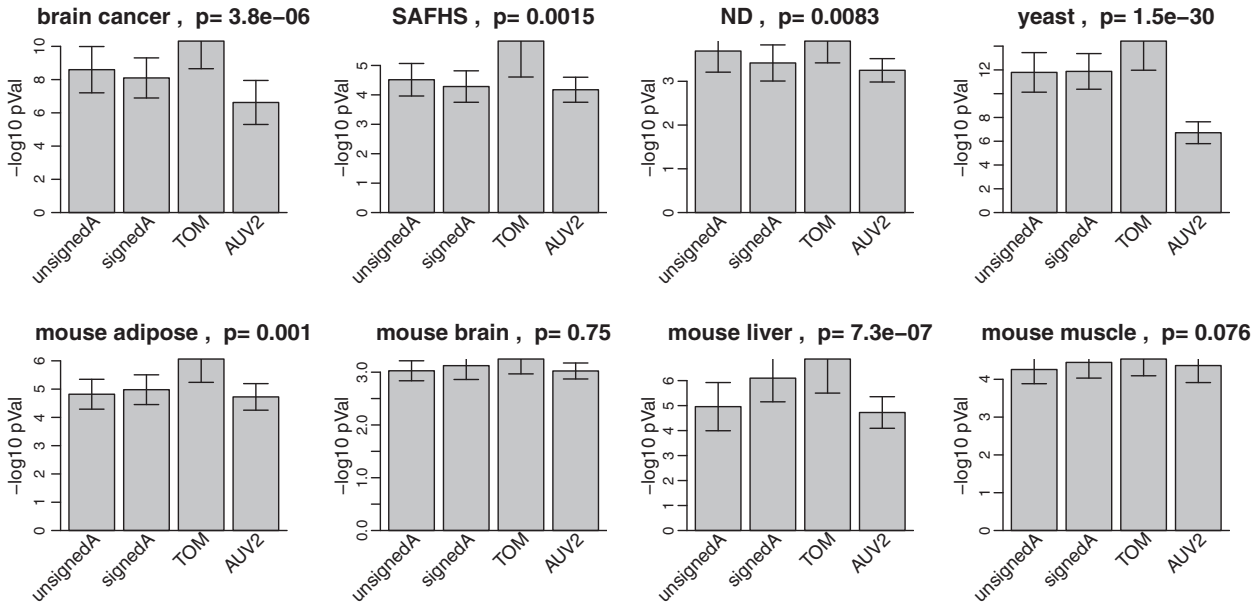

Figure 6 Gene ontology enrichment analysis comparing $A^{MI,UniversalVersion2}$ with bicor based adjacencies in 8 empirical data sets. 5 best GO enrichment p-values from all modules identified using each adjacency are log transformed, pooled together and shown as barplots. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. TOM outperforms the others in all 8 data sets.

spline model construction can be found in Materials and Methods .)

Compared to spline regression, polynomial regression models have a potential shortcoming: the model fit can be adversely affected by outlying observations. A single outlying observation (x_u, y_u) can "bend" the fitting curve into the wrong direction, i.e. adversely affect the estimates of the β coefficients. Spline regression alleviates this problem by fitting model on sub-intervals of the range of x.

Figure 8 (A-B) illustrates the use of regression models for measuring non-linear relationships. In simulation, polynomial and cubic spline regression can correctly capture non-linear trends.

Relationship between regression and MI based networks

Previously, we discussed the relationship between correlation and mutual information based adjacencies in simulations where x and y represent samples from a bivariate normal distribution. Here, we consider the performance of polynomial and spline association measure in the same scenario (Additional file 4). With all x, y pairs following linear relationships, both regression models reduce to simple linear models, and perform almost identically to correlation based measures (panel (A) and (C)). We find that the cor-MI function introduced previously also allows us to relate spline and polynomial regression based networks to the MI based network (panel (B) and (D)), e.g. $AUV2_{ij} \approx F^{cor-MI}(\sqrt{max(R^2(x_i, x_j), R^2(x_j, x_i))})$. Note that different symmetrization methods (Eq. 3) applied R^2 result in similar adjacencies in our applications (refer to Additional file 5), thus it's valid to use any of them.

In addition, our empirical data show that regression models and mutual information adjacency $A^{MI,UniversalVersion2}$ are highly correlated, and the relationship is stronger than that between bicor and $A^{MI,UniversalVersion2}$ (Figure 8 C-F). This indicates that $A^{MI,UniversalVersion2}$ and regression models discover some common gene pairwise non-linear relations that can not be identified by correlations. The Neurological Disease (ND) and mouse muscle sets are shown in Figure 8 as representatives. A detailed analysis of all data sets can be found in Additional file 5.

Simulations for module identification in data with non-linear relationships

Our empirical studies show that most gene pairs satisfy linear relationships, which implies that correlation based network methods perform well in practice. But one can of course simulate data where non-linear association measures (such as MI, spline R^2) outperform correlation measures when it comes to module detection. To illustrate this point, we simulated data with non-linear gene-gene relationships. Here we simulated 200 genes in 3 network modules across 200 samples. Two of the

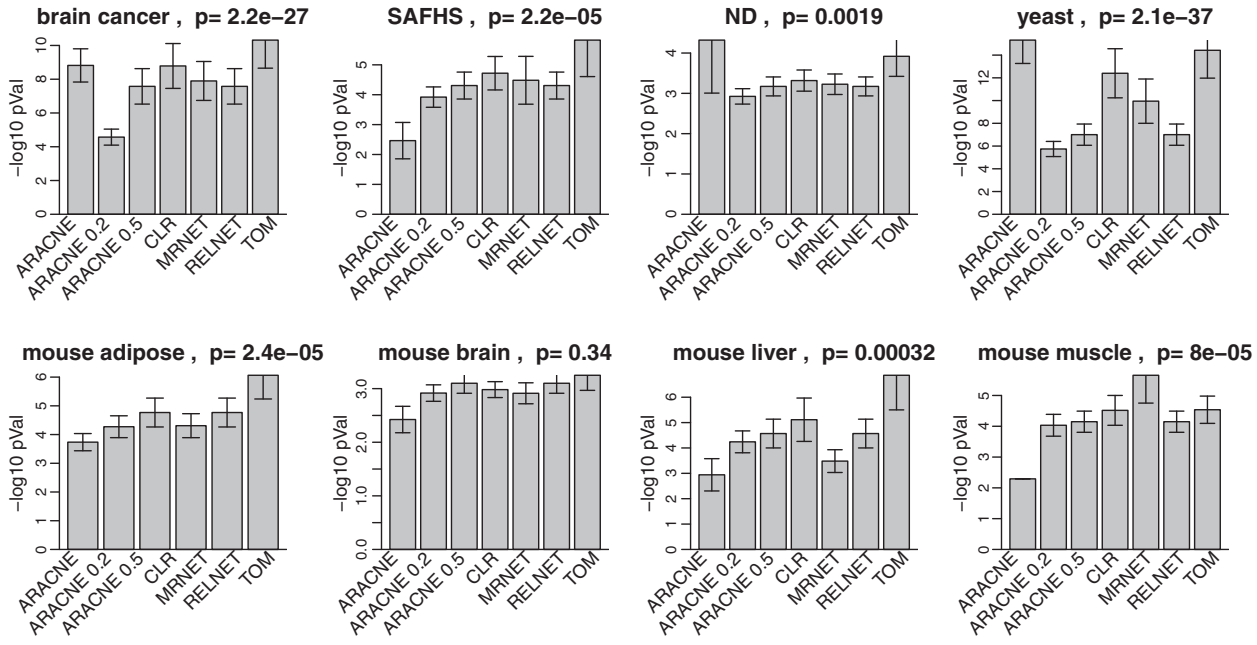

Figure 7 Gene ontology enrichment analysis comparing TOM with MI based adjacencies in 8 empirical data sets. 5 best GO enrichment p-values from all modules identified using each adjacency are log transformed, pooled together and shown as barplots. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. TOM outperforms the others in 5 data sets. ARACNE($\epsilon = 0$) wins in two data sets, making it the second best.

simulated modules, labeled for convenience by the colors turquoise and blue, contain linear and non-linear (quadratic) gene-gene relationships (Figure 5). We then use several different network inference methods to construct networks and define modules. To evaluate how well each network inference method recovers the simulated modules, we use the Rand index between the inferred and simulated module assignment. In this case, non-linear association measures, i.e. AUV2, polynomial and spline regression, identify modules more accurately than correlation based measures (Figure 5). In networks based on correlations, the simulated turquoise and blue modules are clearly divided into two separate ones, indicating that they miss the non-linear relationships within these two modules. In contrast, regression models capture non-linear gene pairwise relations and correctly assign these genes into the same modules. To study the effect of the number of observations, we repeated the analysis for m ranging from 10 to 500. Figure 9 shows that non-linear association measures, especially regression models, outperform correlation based measures as data sample size increases. Note that polynomial and spline regression based co-expression measures perform as well as MI based networks in this situation. Overall, our results validate the usage of polynomial and spline regression models as alternatives to mutual information for detecting non-linear relationships.

Overview of network methods and alternatives

A thorough review of network methods is beyond our scope and we point the reader to the many many review articles [37-40]. But Table 1 describes not only the methods used in this article but also alternative approaches. Table 1 also describes the kind of biological insights that can be gained from these network methods. As a rule, association networks (based on correlation or MI) are ill suited for causal analysis and for encoding directional information. While association networks such as WGCNA or ARACNE have been been successfully used for gene regulatory networks (GRNs) [13], a host of alternatives are available. For example, the DREAM (Dialogue for Reverse Engineering Assessments and Methods) project has repeatedly tackled this problem [41-43]. A limitation of our study is that we are focusing on undirected (as opposed to directed, causal models). Structural equation models, Bayesian networks, and other probabilistic graphical models are widely used for studying causal relationships. Many authors have proposed to use Bayesian networks for analyzing gene expression data [44-47] and for generating causal networks from observational data [48] or genetic data [49,50].

While it is beyond our scope to evaluate network inference methods for time series data (reviewed in [51]), we briefly mention several approaches. A (probabilistic) Boolean network [52] is a special case of a discrete

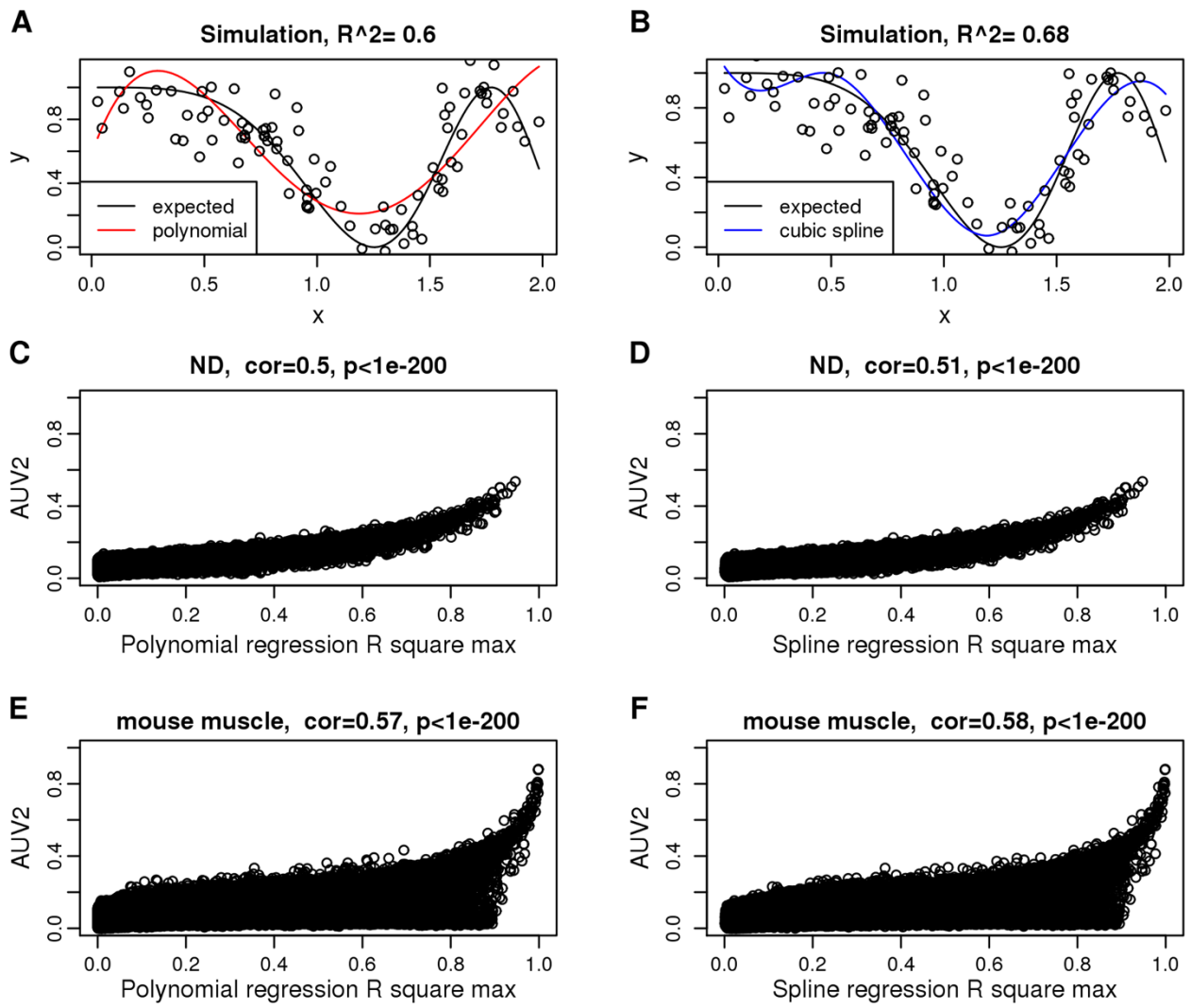

Figure 8 Fitting polynomial and spline regression models to measure non-linear relationships. (A-B) A pair of simulated data *x*, *y* (black dots) with the black curve illustrating the true expected value $E(y)$ given x, where $E(y|x) = cos(x^2)^2$. The red curve shows the fit of a polynomial regression model with degree $d = 4$. The blue curve shows the fit of a cubic spline regression model with 2 knots. Fitting indices of the two models are shown at the top. In simulation, polynomial and cubic spline regression models can properly discover non-linear relations. **(C-D)** Comparisons of regression models and mutual information based co-expression measures in the ND data set. Co-expression of probe pairs is measured with polynomial (d=3)/cubic spline regressions (x-axis) and mutual information $A_{MI,UniversalVersion2}$ (y-axis). The Spearman correlation and p-value of the two measures are shown at the top. **(E-F)** Comparisons in the mouse muscle data set. $A_{MI,UniversalVersion2}$ has a stronger correlation with regression models than with bicor, indicating that the first two measures can capture certain common non-linear patterns.

state space model that characterizes a system using dichotomized data. A Bayesian network is a graph-based model of joint multivariate probability distributions that captures properties of conditional independence between variables [45]. Such models are attractive for their ability to describe complex stochastic processes and for modeling causal relationships. Several articles describe the relationship between Boolean networks and dynamic Bayesian networks when it comes to models of gene regulatory relationships [47,53]. Finally, we mention that correlation network methodology can be adapted to model time series data, e.g. many authors have proposed to use a time-lagged correlation measure for inferring gene regulatory networks [54].

A large part of GRN research focuses on the accurate assessment of individual network edges, e.g. [55-58] so many of these methods are not designed as data reduction methods. In contrast, correlation network methods, such as WGCNA, are highly effective at reducing high dimensional genomic data since modules can be represented by their first singular vector (referred to as module eigengene) [21,59].

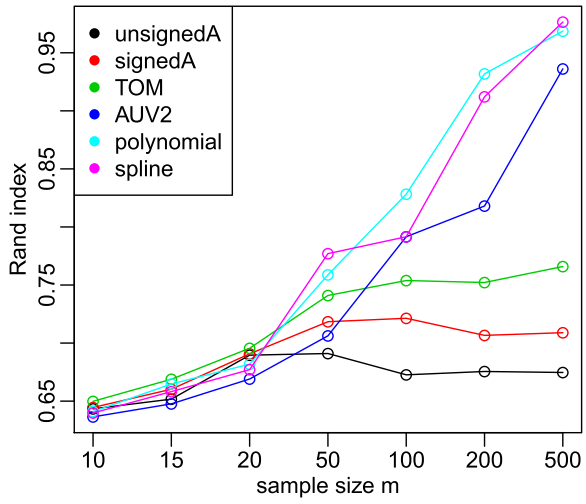

Figure 9 Rand indices in simulations with various number of observations. Simulation sample size versus Rand indices between inferred and simulated module assignments from different network inference methods. increase as the simulation data set contains more samples. Non-linear measures, especially polynomial and spline regression models, outperform other measures as sample size increases.

Discussion

This article presents the following theoretical and methodological results: i) it reviews the relationship between the MI and a likelihood ratio test statistic in case of two categorical variables, ii) it presents a novel empirical formula for relating correlation to MI when the two variables satisfy a linear relationship, and iii) it describes how to use polynomial and spline regression models for defining pairwise co-expression measures that can detect non-linear relationships.

Mutual information has several appealing information theoretic properties. A widely recognized advantage of mutual information over correlation is that it allows one to detect non-linear relationships. This can be attractive in particular when dealing with time series data [60]. But mutual information is not unique in being able to detect non-linear relationships. Standard regression models such as polynomial and spline models can also capture non-linear relationships. An advantage of these models is that well established likelihood based statistical estimation and testing procedures are available. Regression models allow one to calculate model fitting indices that can be used to define network adjacencies as well as flag possible outlying observations by analyzing residuals.

For categorical variables, mutual information is (asymptotically) equivalent to other widely used statistical association measures such as the likelihood ratio statistic or the Pearson chi-square test. In this case, all of these measures (including MI) are arguably optimal association measures. Interpreting MI as a likelihood ratio test statistic facilitates a straightforward approach for adjusting the association measure for additional covariates.

We and others [14] have found close relationships between mutual information and correlation based co-expression networks. Our comprehensive empirical studies show that mutual information is often highly related to the absolute value of the correlation coefficient. We observe that when robust correlation and mutual information disagree, the robust correlation findings appear to be more plausible statistically and biologically. We found that network modules defined using robust correlation exhibit on average higher enrichment in GO categories than modules defined using mutual information. Since our empirical studies involved expression data measured on a variety of platforms and normalized in different ways, we expect that our findings are broadly applicable.

The correlation coefficient is an attractive alternative to the MI for the following reasons. First, the correlation can be accurately estimated with relatively few observations and it does not require the estimation of the (joint) frequency distribution. Estimating the joint density needed for calculating MI typically requires larger sample sizes. Second, the correlation does not depend on hidden parameter choices. In contrast, MI estimation methods involve (hidden) parameter choices, e.g. the number of bins when a discretization method is being used. Third, the correlation allows one to quickly calculate p-values and false discovery rates since asymptotic tests are available (Additional file 1). In contrast, it is computationally challenging to calculate a permutation test p-value for the mutual information between two discretized vectors. Fourth, the sign of the correlation allows one to distinguish positive from negative relationships. Signed correlation networks have been found useful in biological applications [22] and our results show that the resulting modules tend to be more significantly enriched with GO terms that those of networks that ignore the sign information. Fifth, modules comprised of highly correlated vectors can be effectively summarized by the module eigennode (the first principal component of scaled vectors). Sixth, the correlation allows for a straightforward angular interpretation, which facilitates a geometric interpretation of network methods and concepts [59]. For example, intramodular connectivity can be interpreted as module eigennode based connectivity.

Our empirical studies show that a signed weighted correlation network transformed via the topological overlap matrix transformation often leads to the most significant functional enrichment of modules. The recently developed maximal information coefficient [35] has clear theoretical advantages when it comes to measuring general

dependence patterns between variables but our results show that the biweight midcorrelation coupled with the topological overlap measure outperforms the MIC when it comes to the GO ontology enrichment of resulting coexpression modules.

While defining mutual information for categorical variables is relatively straightforward, no consensus seems to exist in the literature on how to define mutual information for continuous variables. A major limitation of our study is that we only studied MI measures based on discretized continuous variables. For example, the corMI function for relating correlation to MI only applies when an equal width discretization method is used with $no.bins = \sqrt{m}$.

A second limitation concerns our gene ontology analysis of modules identified in networks based on various association measures in which we found that the correlation based topological overlap measure (TOM) leads to co-expression modules that are more highly enriched with GO terms than those of alternative approaches. A potential problem with our approach is that the enrichment p-values often strongly depend on (increase with) module sizes, and TOM tends to lead to larger modules. To address this concern, in Additional file 6 we show the enrichment p-values as a function of module size for modules identified by TOM and by AUV2. It turns out that in most studies, the enrichment of modules defined by TOM is better than that of comparably sized modules defined by AUV2.

A third limitation concerns our use of the bicor correlation measure as opposed to alternatives (e.g. Pearson or Spearman correlation). In our study we find that all 3 correlation measures lead to very similar findings (Additional file 7).

Conclusions

Our simulation and empirical studies suggest that mutual information can safely be replaced by linear regression based association measures (e.g. bicor) in case of stationary gene expression measures (which are represented by quantitative variables). To capture general monotonic relationships between such variables, one can use the Spearman correlation. To capture more complicated dependencies, one can use symmetrized model fitting statistics from a polynomial or spline regression model. Regression based association measures have the advantage of allowing one to include covariates (conditioning variables). In case of categorical variables, mutual information is an appropriate choice since it is equivalent to an association measure (likelihood ratio test statistic) of a generalized linear regression model but categorical variables rarely occur in the context of modeling relationships between gene products.

Materials and Methods

Empirical gene expression data sets description

Brain cancer data set. This data set was composed of 55 microarray samples of glioblastoma (brain cancer) patients. Gene expression profiling were performed with Affymetrix high-density oligonucleotide microarrays. A detailed description can be found in [61].

SAFHS data set. This data set [62] was derived from blood lymphocytes of randomly ascertained participants enrolled independent of phenotype in the San Antonio Family Heart Study. Gene expression profiles of 1084 samples were measured by Illumina Sentrix Human Whole Genome (WG-6) Series I BeadChips.

ND data set. This blood lymphocyte data set consisted of 346 samples from patients with neurological diseases. Illumina HumanRef-8 v3.0 Expression BeadChip were used to measure their gene expression profiles.

Yeast data set. The yeast microarray data set was composed of 44 samples from the Saccharomyces Genome Database (http://db.yeastgenome.org/cgi-bin/SGD/expression/expressionConnection.pl). Original experiments were designed to study the cell cycle [63]. A detailed description of the data set can be found in [64].

Tissue-specific mouse data sets. This study uses 4 tissue-specific gene expression data from a large F_2 mouse intercross (B × H) previously described in [65,66]. Specifically, the surveyed tissues include adipose (239 samples), whole brain (221 samples), liver (272 samples) and muscle (252 samples).

Definition of Biweight Midcorrelation

Biweight midcorrelation (bicor) is considered to be a good alternative to Pearson correlation since it is more robust to outliers [67]. In order to define the biweight midcorrelation of two numeric vectors $x = (x_1, \ldots, x_m)$ and $y = (y_1, \ldots, y_m)$, one first defines u_i, v_i with $i = 1, \ldots, m$:

$$\begin{aligned} u_i &= \frac{x_i - med(x)}{9mad(x)} \\ v_i &= \frac{y_i - med(y)}{9mad(y)} \end{aligned} \tag{23}$$

where $med(x)$ is the median of x , and $mad(x)$ is the median absolute deviation of x . This leads us to the definition of weight w_i for x_i , which is,

$$w_i^{(x)} = (1 - u_i^2)^2 I(1 - |u_i|) \tag{24}$$

where the indicator $I(1-|u_i|)$ takes on value 1 if $1-|u_i| > 0$ and 0 otherwise. Therefore, $w_i^{(x)}$ ranges from 0 to 1. It decreases as x_i gets away from $med(x)$, and stays at 0 when x_i differs from $med(x)$ by more than $9mad(x)$. An analogous weight $w_i^{(y)}$ can be defined for y_i . Given the weights, we can define biweight midcorrelation of x and y as:

$$bicor(x,y) = \frac{\sum_{i=1}^{m}(x_i - med(x))w_i^{(x)}(y_i - med(y))w_i^{(y)}}{\sqrt{\sum_{j=1}^{m}\left[(x_j - med(x))w_j^{(x)}\right]^2}\sqrt{\sum_{k=1}^{m}\left[(y_k - med(y))w_k^{(y)}\right]^2}}. \quad (25)$$

A modified version of biweight midcorrelation is implemented as function *bicor* in the WGCNA R package [5,20]. One major argument of the function is "maxPOutliers", which caps the maximum proportion of outliers with weight $w_i = 0$. Practically, we find that $maxPOutliers = 0.02$ detects outliers efficiently while preserving most data. Therefore, 0.02 is the value we utilize in this study.

Types of correlation based gene co-expression networks

Given the expression profile x ,the co-expression similarity s_{ij} between genes i and j can be defined as:

$$s_{ij} = |cor(\boldsymbol{x}_i, \boldsymbol{x}_j)|.$$

An **unweighted network adjacency** A_{ij} between gene expression profiles $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be defined by hard thresholding the co-expression similarity s_{ij} as follows

$$A_{ij} = \begin{cases} 1 & \text{if} \quad s_{ij} \geq \tau \\ 0 & \text{otherwise,} \end{cases} \quad (26)$$

where τ is the 'hard' threshold parameter. Hard thresholding of the correlation leads to simple network concepts (e.g., the gene connectivity equals the number of direct neighbors) but it may lead to a loss of information.

To preserve the continuous nature of the co-expression information, we define the **weighted network adjacency** between 2 genes as a power of the absolute value of the correlation coefficient [4,61]:

$$A_{ij} = s_{ij}^{\beta}, \quad (27)$$

with $\beta \geq 1$. This soft thresholding approach emphasizes strong correlations, punishes weak correlations, and leads to a weighted gene co-expression network.

An important choice in the construction of a correlation network concerns the treatment of strong negative correlations. In **signed networks** negatively correlated nodes are considered unconnected. In contrast, in **unsigned networks** nodes with high negative correlations are considered connected (with the same strength as nodes with high positive correlations). As detailed in [4,22], a signed weighted adjacency matrix can be defined as follows

$$A_{ij} = (0.5 + 0.5\text{cor}(x_i, x_j))^{\beta} \quad (28)$$

and an unsigned adjacency by

$$A_{ij} = |cor(x_i, x_j)|^{\beta}. \quad (29)$$

β is default to 6 for unsigned adjacency and 12 for signed adjacency. The choice of signed vs. unsigned networks depends on the application; both signed [22] and unsigned [30,61,65] weighted gene networks have been successfully used in gene expression analysis.

Adjacency function based on topological overlap

The topological overlap matrix (TOM) based adjacency function A_{TOM} maps an original adjacency matrix $A^{original}$ to the corresponding topological overlap matrix, i.e.

$$A_{TOM}(A^{original})_{ij} = \frac{\sum_{l\neq i,j} A_{il}^{original} A_{l,j}^{original} + A_{ij}^{original}}{min(\sum_{l\neq i} A_{il}^{original}, \sum_{l\neq j} A_{jl}^{original}) - A_{ij}^{original} + 1}. \quad (30)$$

The TOM based adjacency function A_{TOM} is particularly useful when the entries of $A^{original}$ are sparse (many zeroes) or susceptible to noise. This replaces the original adjacencies by a measure of interconnected that is based on shared neighbors. The topological overlap measure can serve as a filter that decreases the effect of spurious or weak connections and it can lead to more robust networks [17,18,68].

Mutual-information based network inference methods

There are 4 commonly used mutual-information based network inference methods: RELNET, CLR, MRNET and ARACNE. In order to identify pairwise interactions between numeric variables x_i, x_j, all methods start by estimating mutual information $MI(x_i, x_j)$.

RELNET

The relevance network (RELNET) approach [6,28] thresholds the pairwise measures of mutual information by a threshold τ . However, this method suffers from a significant limitation that vectors separated by one or more intermediaries (indirect relationships) may have high mutual information without implying a direct interaction.

CLR

The CLR algorithm [26] is based on the empirical distribution of MI. It first defines a score z_i given the mutual information $MI(x_i, x_j)$ and the sample mean μ_i and standard deviation σ_i of the empirical distribution of mutual information $MI(x_i, x_k), k = 1, \ldots, n$:

$$z_i = \max\left(0, \frac{MI(x_i, x_j) - \mu_i}{\sigma_i}\right). \quad (31)$$

z_j can be defined analogously. In terms of z_i, z_j, the score used in CLR algorithm can be expressed as $z_{ij} = \sqrt{z_i^2 + z_j^2}$.

MRNET

MRNET [27] infers a network by repeating the maximum relevance/minimum redundancy (MRMR) feature selection method for all variables. The MRMR method starts by selecting the variable x_i having the highest mutual information with target y. Next, given a set S of selected variables, the criterion updates S by choosing the variable x_k that maximizes $u_j - r_j$ where u_j is a relevance term and r_j is a redundancy term. In particular,

$$u_j = MI(x_k, y) \tag{32}$$

$$r_j = \frac{1}{|S|} \sum_{x_i \in S} MI(x_k, x_i) \tag{33}$$

The score of each pair x_i and x_j will be the maximum score of the one computed when x_i is the target and the one computed when x_j is the target.

ARACNE

The ARACNE [9] (Algorithm for the Reconstruction of Accurate Cellular Networks) developed by Andrea Califano's group is an extension of RELNET. Given the limitation of RELNET, ARACNE removes the vast majority of indirect candidate interactions using a well-known information theoretic property, the **data processing inequality** (DPI). The DPI applied to association networks states that if variables x_i and x_j interact only through a third variable x_k , then

$$MI(x_i, x_j) \leq min(MI(x_i, x_k), MI(x_k, x_j)) \tag{34}$$

ARACNE starts with a network graph where each pair of nodes with $MI_{ij} > \tau$ is connected by an edge. The weakest edge of each triplet, e.g. the edge between i and j , is interpreted as an indirect interaction and is removed if the difference between $min(MI(x_i, x_k), MI(x_k, x_j))$ and $MI(x_i, x_j)$ lies above a threshold ϵ , i.e. the edge is removed if

$$MI(x_i, x_j) \leq min(MI(x_i, x_k), MI(x_k, x_j)) - \epsilon. \tag{35}$$

The tolerance threshold ϵ could be chosen to reflect the variance of the MI estimator and should decrease with increasing sample size m . Using a non-zero tolerance $\epsilon > 0$ can lead to the persistence of some 3-vector loops.

The outputs from RELNET, CLR, MRNET or ARACNE are association matrices. They can be transformed into corresponding adjacencies based on the algorithm discussed in Introduction.

MIC

Another mutual information based method is the recently proposed the maximal information coefficient (MIC) [35]. The MIC is a type of maximal information-based nonparametric exploration (MINE) statistics [35]. In our empirical evaluations, we calculate the MIC using the *minerva* R package [69].

Fitting indices of polynomial regression models

While networks based on the Pearson correlation can only capture linear co-expression patterns there is clear evidence for non-linear co-expression relationships in transcriptional regulatory networks [70]. The following classical regression based approaches can be used for studying non-linear relationships. The polynomial regression model:

$$\begin{aligned} E(y) &= \beta_0 1 + \beta_1 x + \beta_2 x^2 \ldots + \beta_d x^d \\ &= M\beta\,, \end{aligned} \tag{36}$$

where

$$M = [\,1, x, \ldots, x^d\,]\,. \tag{37}$$

One can show that the least squares estimate of the parameter vector $\hat{\beta}$ is

$$\hat{\beta} = (M^\tau M)^- M^\tau\, y,$$

where $^-$ denotes the (pseudo) inverse, and $^\tau$ denotes the transpose of a matrix.

Given $\hat{\beta}$, we can calculate the fitting index R^2 as:

$$R^2 = cor(y, \hat{y})^2 = cor(y, M\hat{\beta})^2 \tag{38}$$

In the context of a regression model, R^2 is also known as the proportion of variation of y explained by the model.

Spline regression model construction

To investigate the relationship between variable x and y , one can use another textbook method from the arsenal of statisticians: spline regression models. Here knots are used to decide boundaries of the sub-intervals. They are typically pre-specified, e.g. based on quantiles of x . The choice of the knots will affect the model fit. It turns out that the values of the knots (i.e. their placement) is not as important as the number of knots. We use the following rule of thumb for the number of knots: if $m > 100$ use 5 knots, if $m < 30$ use 3 knots, otherwise use 4 knots.

To ensure that fit between y and x satisfies a continuous relationship, we review the **hockey stick function** $()_+$ to transform x :

$$(s)_+ = \begin{cases} s & \text{if } s \geq 0 \\ 0 & \text{if } s < 0. \end{cases} \tag{39}$$

This function can also be applied to the components of a vector, e.g. $(x)_+$ denotes a vector whose negative components have been set to zero. So $(x - knot1)_+$ is a vector whose u-th component equals $x[\,u] - knot1$ if $x[\,u] - knot1 \geq 0$ and 0 otherwise.

We are now ready to describe **cubic spline regression model**, which fits polynomial of degree 3 to sub-intervals. The general form of a cubic spline with 2 knots is as follows

$$E(y) = \beta_0 1 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - knot_1)_+^3 + \beta_5 (x - knot_2)_+^3. \quad (40)$$

The knot parameters (numbers) $knot_1, knot_2, \ldots$ are chosen before estimating the parameter values. Analogous to polynomial regression, R^2 can be calculated as the association measure between x and y . This method guarantees the smoothness of the regression line and restrict the influence of each observation to its local sub-interval.

Other networks

Boolean network [71] and Probabilistic network [72,73] are briefly mentioned in Table 1.

Availability of software

Project name: Adjacency matrix for non-linear relationships
Project home page: http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/Rpackages/WGCNA
Operating system(s): Platform independent
Programming language: R
Licence: GNU GPL 3

The following functions described in this article have been implemented in the WGCNA R package [5]. Function *adjacency.polyReg* and *adjacency.splineReg* calculate polynomial and spline regression R^2 based adjacencies. Users can specify the R^2 symmetrization method. Function *mutualInfoAdjacency* calculates the mutual information based adjacencies $A^{MI,SymmetricUncertainty}$ (Eq. 14), $A^{MI,UniversalVersion1}$ (Eq. 15) and $A^{MI,UniversalVersion2}$ (Eq. 16). Function *AFcorMI* implements the F^{cor-MI} prediction function 18 for relating correlation with mutual information.

Additional files

Additional file 1: Detailed methods descriptions. In this document, we provide detail information of entropy, mutual information, likelihood ratio test statistics and p-value calculation of correlation coefficients.

Additional file 2: Empirical analysis using large number of genes in the mouse adipose and ND data sets. Page one is an empirical analysis using all 23568 genes without restricting to 3000 genes for the mouse adipose data set. (A) Absolute value of bicor versus $A^{MI,UniversalVersion2}$. One million randomly sampled gene pairs are plotted to reduce computational burden. The two measures show good monotonic relationship. The red curve predicts $A^{MI,UniversalVersion2}$ from bicor. The blue circle highlights the probe pair with the highest $A^{MI,UniversalVersion2}$ z-score among those with insignificant bicor z-scores (less than 1.9); the red circle highlights the probe pair with the highest bicor z-score among those with insignificant $A^{MI,UniversalVersion2}$ z-scores (less than 1.9). Red and blue circles are selected based on all gene pairs rather than sampled ones. (B) Prediction from bicor based on Eq. 18 and observed $A^{MI,UniversalVersion2}$ are highly correlated. As in (A), one million randomly sampled gene pairs are plotted. Line y=x is added. (C) Gene expression of probe pairs highlighted by blue circles. (D) Gene expression of probe pairs highlighted by red circles.
Page two is the same analysis for ND data set using 10000 randomly selected genes rather than 3000 genes with highest variance.

Additional file 3: Comparison of MIC and correlation based co-expression measures. Comparison of MIC and correlation in our empirical gene expression data sets except SAFHS. This is an extension of Figure 6. 5 best GO enrichment p-values from all modules identified using MIC and TOM are log transformed, pooled together and shown as barplots. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. TOM outperforms MIC in all data sets except the mouse brain data.

Additional file 4: Compare polynomial and spline regression models to correlation or mutual information based co-expression measures in simulation. Each point corresponds to a pair of numeric vectors x and y with length $m = 1000$. Data is simulated as in Figure 1. (A) Square root of R^2 from polynomial regression symmetrized by Eq. 5 versus absolute Pearson correlation values. The two measures are indistinguishable since the data is simulated to exhibit linear relationships. (B) R^2 from polynomial regression symmetrized by Eq. 5 versus $A^{MI,UniversalVersion2}$. The red line predicts $A^{MI,UniversalVersion2}$ from R^2. (C-D) Same plots for spline regression models.

Additional file 5: Polynomial and spline regression models for estimating non-linear relationships in real data application. In this document, we use polynomial and spline regression models to estimate non-linear relationships in real data applications.

Additional file 6: The relationship between module size and gene ontology enrichment p-values in 8 real data applications. In each panel, module size (x-axis) is plotted against − log 10 GO enrichment p-values (y-axis)in dots. Loess regression lines are provided to show the trend. Red and black color represent network modules constructed using TOM and $A^{MI,UniversalVersion2}$ based measures, respectively. In most data sets, the enrichment of modules defined by TOM is better than that of comparably sized modules defined by $A^{MI,UniversalVersion2}$.

Additional file 7: Comparison of bicor, Pearson correlation and Spearman correlation based signed adjacency in 8 empirical data sets. Each panel show the − log 10 transformed 5 best gene ontology enrichment p-values of all modules identified using each type of adjacency. Error bars stand for 95% confidence intervals. On top of each panel is a p-value based on multi-group comparison test. All three types of correlation are similar in terms of GO enrichment.

Abbreviations

MI: Mutual information; Bicor: Biweight midcorrelation; MIC: Maximal information coefficient; ARACNE: Algorithm for the reconstruction of accurate cellular networks; GO: Gene ontology; LRT: Likelihoood ratio test; TOM: Topological overlap matrix; WGCNA: Weighted correlation network analysis.

Competing interests

We declare no conflict of interest.

Authors' contributions

LS and SH performed the research; LS, SH, and PL wrote the paper and developed R software functions. SH designed the research. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge grant support from 1R01 DA030913-01, P50CA092131, R01NS058980, and the UCLA CTSI.

Received: 14 March 2012 Accepted: 30 November 2012
Published: 9 December 2012

References

1. Eisena M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci U S A* 1998, **95**(25):14863–14868.
2. Zhou X, Kao M, Wong W: **Transitive Functional Annotation By Shortest Path Analysis of Gene Expression Data.** *Proc Natl Acad Sci U S A* 2002, **99**(20):12783–12788.
3. Stuart JM, Segal E, Koller D, Kim SK: **A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules.** *Science* 2003, **302**(5643):249–255.
4. Zhang B, Horvath S: **General framework for weighted gene coexpression analysis.** *Stat Appl Genet Mol Biol* 2005, **4:**17.
5. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9:**559.
6. Butte A, Tamayo P, Slonim D, Golub T, Kohane I: **Discovering Functional Relationships Between RNA Expression and Chemotherapeutic Susceptibility Using Relevance Networks.** *Proc Natl Acad Sci U S A* 2000, **97:**12182–12186.
7. Daub C, Steuer R, Selbig J, Kloska S: **Estimating mutual information using B-spline functions - an improved similarity measure for analysing gene expression data.** *BMC Bioinformatics* 2004, **5:**118.
8. Basso K, Margolin A, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human B cells.** *Nat Genet* 2005, **37**(4):382–390.
9. Margolin A, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera R, Califano A: **ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.** *BMC Bioinformatics* 2006, **7**(Suppl 1):S7.
10. Priness I, Maimon O, Ben-Gal I: **Evaluation of gene-expression clustering via mutual information distance measure.** *BMC Bioinformatics* 2007, **8:**111. [http://www.biomedcentral.com/1471-2105/8/111]
11. Meyer P, Lafitte F, Bontempi G: **minet: A R/Bioconductor Package for Inferring Large Transcriptional Networks Using Mutual Information.** *BMC Bioinformatics* 2008, **9:**461.
12. Cadeiras M, Bayern MV, Sinha A, Shahzad1 K, Lim WK, Grenett H, Tabak E, Klingler T, Califano A, Deng MC: **Drawing networks of rejection - a systems biological approach to the identification of candidate genes in heart transplantation.** *J Cell Mol Med* 2010, **15**(4):949–956.
13. Allen JD, Xie Y, Chen M, Girard L, Xiao G: **Comparing Statistical Methods for Constructing Large Scale Gene Networks.** *PLoS ONE* 2012, **7:**e29348. [http://dx.doi.org/10.1371]
14. Steuer R, Kurths J, Daub CO, Weise J, Selbig J: **The mutual information: Detecting and evaluating dependencies between variables.** *Bioinformatics* 2002, **18**(Suppl 2):S231–S240.
15. Lindlof A, Lubovac Z: **Simulations of simple artificial genetic networks reveal features in the use of Relevance Networks.** *In Silico Biology* 2005, **5**(3):239–250.
16. Ravasz E, Somera A, Mongru D, Oltvai Z, Barabasi A: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297**(5586):1551–1555.
17. Yip A, Horvath S: **Gene Network Interconnectedness and the Generalized Topological Overlap Measure.** *BMC Bioinformatics* 2007, **8**(8):22.
18. Li A, Horvath S: **Network neighborhood analysis with the multi-node topological overlap measure.** *Bioinformatics* 2007, **23**(2):222–231.
19. Hardin J, Mitani A, Hicks L, VanKoten B: **A robust measure of correlation between two genes on a microarray.** *BMC Bioinformatics* 2007, **8:**220.
20. Langfelder P, Horvath S: **Fast R Functions For Robust Correlations And Hierarchical Clustering.** *J Stat Softw* 2012, **46**(i11):1–17.
21. Horvath S: *Weighted Network Analysis. Applications in Genomics and Systems Biology*. New York: Springer Book; 2011.
22. Mason M, Fan G, Plath K, Zhou Q, Horvath S: **Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells.** *BMC Genomics* 2009, **10:**327.
23. Cover T, Thomas J: *Elements of information theory*. New York: John Wiley Sons; 1991.
24. Paninski L: **Estimation of entropy and mutual information.** *Neural Computation* 2003, **15**(6):1191–1253.
25. Kraskov A, Stögbauer H, andrzejak R, Grassberger P: **Hierarchical Clustering Using Mutual Information.** *EPL (Europhysics Letters)* 2007, **70**(2):278.
26. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS: **Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles.** *PLoS Biol* 2007, **5:**e8. [http://dx.doi.org/10.1371]
27. Meyer PE, Kontos K, Lafitte F, Bontempi G: **Information-Theoretic Inference of Large Transcriptional Regulatory Networks.** *EURASIP J Bioinforma Syst Biol* 2007, **2007:**79879.
28. Butte A, Kohane I: **Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurments.** *Pac Symp Biocomput* 2000:418–429.
29. Moon YI, Rajagopalan B, Lall U: **Estimation of mutual information using kernel density estimators.** *Phys Rev E* 1995, **52**(3):2318–2321.
30. Oldham M, Konopka G, Iwamoto K, Langfelder P, Kato T, Horvath S, Geschwind D: **Functional organization of the transcriptome in human brain.** *Nat Neurosci* 2008, **11**(11):1271–1282.
31. Wolfe C, Kohane I, Butte A: **Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks.** *BMC Bioinformatics* 2005, **6:**227.
32. Langfelder P, Zhang B, Horvath S: **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut library for R.** *Bioinformatics* 2007, **24**(5):719–720.
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Sherlock GMRG: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25:**25–29.
34. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini A, Sawitzki G, Smith C, Smyth G, Tierney L, Yang Y, Zhang J: **Bioconductor: Open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5:**R80.
35. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC: **Detecting Novel Associations in Large Data Sets.** *Science* 2011, **334**(6062):1518–1524. [http://www.sciencemag.org/content/334/6062/1518.abstract]
36. Faraway J: **Practical Regression and Anova using R.** *R pdf file at http://cranr-projectorg/doc/contrib/Faraway-PRApdf* 2002.
37. D'Haeseleer P, Liang S, Somogyi R: **Genetic network inference: from co-expression clustering to reverse engineering.** *Bioinformatics* 2000, **16**(8):707–726. [http://dx.doi.org/10.1093/bioinformatics/16.8.707]
38. Markowetz F, Spang R: **Inferring cellular networks–a review.** *BMC bioinformatics* 2007, **8**(Suppl 6):S5+. [http://dx.doi.org/10.1186/1471-2105-8-S6-S5]
39. Bansal M, Belcastro V, Ambesi-Impiombato A, di Bernardo, D: **How to infer gene networks from expression profiles.** *Molecular Systems Biology* 2007, **3:**78. [http://dx.doi.org/10.1038/msb4100120]
40. De Smet, R, Marchal K: **Advantages and limitations of current network inference methods.** *Nat Rev Micro* 2010, **8**(10):717–729. [http://dx.doi.org/10.1038/nrmicro2419]
41. Stolovitzky G, MONROE D, Califano A: **Dialogue on Reverse-Engineering Assessment and Methods.** *Ann NY Acad Sci* 2007, **1115**(1):1–22.
42. Stolovitzky G, Prill RJ, Califano A: **Lessons from the DREAM2 Challenges.** *Ann NY Acad Sci* 2009, **1158:**159–195.
43. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G: **Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges.** *PLoS ONE* 2010, **5**(2):e9202.
44. Friedman N, Linial M, Nachman I, Pe'er D: **Using Bayesian networks to analyze expression data.** *J Comput Biol* 2000, **7**(3):601–620.
45. Perrin B, Ralaivola L: **Gene networks inference using dynamic Bayesian networks.** *Bioinformatics* 2003, **19**(Suppl 2):II138–II148.
46. Friedman N: **Inferring cellular networks using probabilistic graphical models.** *Science* 2004, **303**(5659):799–805.
47. Li P, Zhang C, Perkins E, Gong P, Deng Y: **Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks.** *BMC Bioinformatics* 2007, **8**(Suppl 7):S13. [http://www.biomedcentral.com/1471-2105/8/S7/S13]

48. Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED: **Advances to Bayesian network inference for generating causal networks from observational biological data.** *Bioinformatics* 2004, **20**(18):3594–3603. [http://bioinformatics.oxfordjournals.org/content/20/18/3594.abstract]
49. Zhu J, Lum P, Lamb J, HuhaThakurta D, Edwards S, Thieringer R, Berger J, Wu M, Thompson J, Sachs A, Schadt E: **An integrative genomics approach to the reconstruction of gene networks in segregating populations.** *Cytogenet Genome Res* 2004, **105:**363–374.
50. Schadt E, Lamb J, Yang X, Zhu J, Edwards J, GuhaThakurta D, Sieberts S, Monks S, Reitman M, Zhang C, Lum P, Leonardson A, Thieringer R, Metzger J, Yang L, Castle J, Zhu H, Kash S, Drake T, Sachs A, Lusis A: **An integrative genomics approach to infer causal associations between gene expression and disease.** *Nature Genetics* 2005, **37**(7):710–717.
51. Sima C, Hua J, Jung S: **Inference of Gene Regulatory Networks Using Time-Series Data: A Survey.** *Curr Genomics* 2009, **10**(6):416–429.
52. Shmulevich I, Dougherty ER, Kim S, Zhang W: **Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks.** *Bioinformatics* 2002, **18**(2):261–274. [http://bioinformatics.oxfordjournals.org/content/18/2/261.abstract]
53. Lahdesmki H, Hautaniemi S, Shmulevich I, Yli-Hrja O: **Relationships between probabilistic Boolean networks and dynamic Bayesian networks as models of gene regulatory networks.** *Signal Processing* 2006, **86**(4):814–834.
54. Schmitt WA, Raab RM, Stephanopoulos G: **Elucidation of Gene Interaction Networks Through Time-Lagged Correlation Analysis of Transcriptional Data.** *Genome Research* 2004, **14**(8):1654–1663. [http://genome.cshlp.org/content/14/8/1654.abstract]
55. Fernandes JS, Sternberg PW: **The tailless Ortholog nhr-67 Regulates Patterning of Gene Expression and Morphogenesis in the C. elegans Vulva.** *PLoS Genet* 2007, **3**(4):e69. [http://dx.plos.org/10.1371]
56. Yan J, Wang H, Liu Y, Shao C: **Analysis of Gene Regulatory Networks in the Mammalian Circadian Rhythm.** *PLoS Comput Biol* 2008, **4**(10):e1000193. [http://dx.doi.org/10.1371]
57. Altay G, Emmert-Streib F: **Revealing differences in gene network inference algorithms on the network-level by ensemble methods.** *Bioinformatics* 2010, **26**(14):1738–1744.
58. Chaitankar V, Ghosh P, Perkins E, Gong P, Zhang C: **Time lagged information theoretic approaches to the reverse engineering of gene regulatory networks.** *BMC Bioinformatics* 2010, **11**(Suppl 6):S19.
59. Horvath S, Dong J: **Geometric interpretation of Gene Co-expression Network Analysis.** *PloS Comput Biol* 2008, **4**(8):e1000117.
60. Wiggins C, Nemenman I: **Process pathway inference via time series analysis.** *Experimental Mechanics* 2003, **43**(3):361–370.
61. Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, Laurance M, Zhao W, Shu Q, Lee Y, Scheck A, Liau L, Wu H, Geschwind D, Febbo P, Kornblum H, TF C, Nelson S, Mischel P: **Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target.** *Proc Natl Acad Sci U S A* 2006, **103**(46):17402–7.
62. Goring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J: **Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes.** *Nat Genet* 2007, **39:**1208–1216.
63. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization.** *Mol Biol Cell* 1998, **9**(12):3273–3297.
64. Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson SF: **Gene Connectivity, Function, and Sequence Conservation: Predictions from Modular Yeast Co-expression Networks.** *BMC Genomics* 2006, **7**(7):40.
65. Ghazalpour A, Doss S, Zhang B, Plaisier C, Wang S, Schadt E, Thomas A, Drake T, Lusis A, Horvath S: **Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight.** *PloS Genetics* 2006, **2**(2):8.
66. Fuller T, Ghazalpour A, Aten J, Drake T, Lusis A, Horvath S: **Weighted gene coexpression network analysis strategies applied to mouse weight.** *Mamm Genome* 2007, **18**(6-7):463–472.
67. Wilcox R: *Introduction to Robust Estimation and Hypothesis Testing*. San Diego: Academic Press; 1997.
68. Dong J, Horvath S: **Understanding Network Concepts in Modules.** *BMC Syst Biol* 2007, **1:**24.
69. Albanese D, Filosi M, Visintainer R, Riccadonna S, Jurman G, Furlanello C: **cmine, minerva and minepy: a C engine for the MINE suite and its R and Python wrappers.** *ArXiv e-prints* 2012 , **1**(24).
70. Li H, Zhan M: **Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data.** *Bioinformatics* 2008, **24**(17):1874–1880.
71. Kauffman S: **Metabolic stability and epigenesis in randomly connected nets.** *J.Theoret.Biol.* 1969, **22:**437–467.
72. Chen X, Chen M, Ning K: **BNArray: an R package for constructing gene regulatory networks from microarray data by using Bayesian network.** *Bioinformatics* 2006. [http://view.ncbi.nlm.nih.gov/pubmed/17005537]
73. Werhli AV, Grzegorczyk M, Husmeier D: **Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks.** *Bioinformatics* 2006, **22**(20):2523–2531. [http://dx.doi.org/10.1093/bioinformatics/btl391]

doi:10.1186/1471-2105-13-328
Cite this article as: Song *et al.*: Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* 2012 **13**:328.