Running head
Gini correlation for transcriptome analysis

Corresponding author
Name: Xiangfeng Wang
Address: School of Plant Science, University of Arizona, Tucson, AZ, 85721, USA.
Tel: 520-626-4184.
E-mail: xwang1@cals.arizona.edu

Research category
Bioinformatics

Keywords
Gini correlation, correlation analysis, transcription regulation, microarray, RNA-Seq
Application of the Gini correlation coefficient to infer regulatory relationships in transcriptome analysis

Chuang Ma¹ and Xiangfeng Wang¹ *

¹School of Plant Sciences, University of Arizona, Tucson, AZ, 85721 USA.
FOOTNOTES

*Corresponding author: Xiangfeng Wang, e-mail: xwang1@cals.arizona.edu.
ABSTRACT

One of the computational challenges in plant systems biology is to accurately infer the transcriptional regulation relationships based on the correlation analyses of gene expression patterns. Despite several correlation methods are applied in biology to analyze microarray data, concerns regarding the compatibility of these methods to the gene expression data profiled by high-throughput RNA transcriptome sequencing (RNA-Seq) technology have been raised. These concerns are mainly due to the fact that the distribution of read counts in RNA-Seq experiments is different from that of fluorescence intensities in microarray experiments. Therefore, a comprehensive evaluation of the existing correlation methods and if necessary, introduction of novel methods into biology is appropriate. In this study, we compared four existing correlation methods used in microarray analysis and one novel method called Gini correlation coefficient, on previously published microarray-based and sequencing-based gene expression data in Arabidopsis and maize. The comparisons were performed on more than 11,000 regulatory relationships in Arabidopsis, including 8,929 pairs of transcription factors and target genes. Our analyses pinpointed the strengths and weaknesses of each method, and indicated that the Gini correlation can compensate for the shortcomings of the Pearson correlation, the Spearman correlation, the Kendall correlation and the Tukey’s biweight correlation. The Gini correlation method, with the other four evaluated methods in this study, was implemented as an R package named “rsgcc” that can be utilized as an alternative option for biologists to perform clustering analyses of gene expression patterns or transcriptional network analyses.

INTRODUCTION
One of the computational challenges in plant systems biology is to construct biological networks that aid in elucidating the functional relationships of genes during plant development and in response to environmental stimuli from genome-scale experiments (Long et al., 2008; Nakashima et al., 2009; Moreno-Risueno et al., 2010; Wellmer and Riechmann, 2010). Although biological networks encompass different types of physical interactions at the protein, RNA, DNA and even epigenetic levels, inference of the transcriptional regulation relationships from gene expression data remains the most common and efficient way to monitor dynamic biological processes (Ma et al., 2007; Long et al., 2008; Berri et al., 2009; Vandepoele et al., 2009). While microarray technology has been a dominant approach for gene expression profiling over the past decade, next-generation sequencing technology has emerged as a powerful platform to profile transcriptomes in a de novo manner without relying on the availability of genome sequences (Mortazavi et al., 2008; Wang et al., 2009). Compared to microarray data, in which gene expression levels are measured by fluorescence intensities, RNA-Seq experiments use short read counts to represent gene expression abundance, in which the discrete nature of read counts results in a Poisson or binomial distribution characterized by a long, heavy tail (Garber et al., 2011; Hu et al., 2012). Based on this presumption, computational biologists have developed new software, such as EdgeR and Cufflinks that use Poisson and binomial distributions to detect differentially expressed genes from RNA-Seq data (Robinson et al., 2010; Trapnell et al., 2010). Currently, most existing RNA-Seq tools focus on read mapping, expression measurement, differential expression detection and variation calls. Thus, novel bioinformatic tools and methodologies are expected for advanced statistical analyses of sequencing-based gene expression data, such as clustering and network analyses with the consideration of the properties of RNA-Seq data.

In gene expression analyses, the co-regulation relationship of two genes can be inferred by the correlation coefficients that are derived using multiple mathematical methods, such as the Pearson’s product-moment correlation coefficient (PCC), the Spearman’s rank correlation coefficient (SCC) and the Kendall’s rank correlation coefficient (KCC) (Rice et al., 2005; Scheinine et al., 2009; Ficklin and Feltus, 2011).
While the PCC infers the linear relationship between two genes based on the covariance and standard deviation from the expression values in a series of samples, the SCC and the KCC use the ranks of gene expression levels in the samples to compute correlations instead of directly using expression values. Although the SCC and the KCC are more robust on non-normal distributions compared to the PCC, they have not been favored by biologists because the information of expression levels is not considered. Of the many correlation methods in biology, the Pearson correlation is the most commonly used technique that has been often applied in clustering analyses and network constructions; however, disadvantages of this method have been frequently reported. For instance, although the PCC performs well in deriving global linear relationships between two variables, its performance is dramatically reduced on partial linear relationships or non-linear relationships (Hardin et al., 2007; Reshef et al., 2011). Moreover, the PCC is not stable to outlier data points representing the extreme values (either low or high) of a gene’s expression, which are substantially deviated from the median and/or average expression level in a series of samples (Hardin et al., 2007; Usadel et al., 2009).

Recent studies regarding the regulatory networks in *Escherichia coli* (*E. coli*) and *Saccharomyces cerevisiae* have shown that the current correlation methods are not adequate enough to infer all of the regulatory relationships (Marbach et al., 2010; Allen et al., 2012). For instance, the Pearson correlation can only detect 60% of the true positive regulatory relationships in *E. coli*, and > 40% of the predicted relationships are false positives (Allen et al., 2012). This result is attributed to the complexity of the biological systems, in which most regulatory relationships are not globally represented as linear. Because the expression levels of a transcription factor (TF) and target genes may greatly vary and the transcriptional regulation may occur transiently in specific conditions or tissues, the PCC is not sensitive enough to derive such relationships (Usadel et al., 2009). Specifically, with the exception of linear relationships, a considerable amount of gene regulation exists in non-linear relationships, such as inverted (negative regulation) or time-delayed (regulatory response lag) patterns (Yu et al., 2003).
With the recent availability of the protein interactome (Arabidopsis Interactome Mapping Consortium, 2011) and a TF-target interaction database (AtTFDB of AGRIS) for Arabidopsis (Yilmaz et al., 2011), a systematic evaluation of the commonly used correlation methods in biology on their power to infer regulatory relationships and their compatibility to RNA-Seq data analyses is in high demand. If necessary, novel correlation methods shall be introduced into biology. The Gini correlation coefficient (GCC) is a member of the family of the Gini methodologies that have been widely used in economics, sociology, physics, engineering and informatics to solve a series of mathematic problems, without having to hypothesize the forms of data distributions (Yitzhaki, 2003). In economics, the GCC is used to calculate correlations between sources of family income (e.g., salaries) and the total family income for a country (Schechtman and Yitzhaki, 1999). The robustness of the GCC has recently been demonstrated in biology in analyzing the connectivity of genes in transcriptional networks (Ma et al., 2011). Similar to other correlation coefficients, the GCC values range from -1.0 to 1.0. While 0 indicates the absolute independency between two variables, -1.0 and 1.0 indicate the absolutely monotonic decreasing and absolutely monotonic increasing relationships, respectively. Different from the PCC, SCC and KCC, the GCC can compute the correlation of two variables considering both rank and value information. In this way, the Gini correlation is more robust on non-normally distributed data and it’s more stable for the data containing outliers, compared to the correlation methods developed based on normal distributions. Additionally, the consideration of gene expression values provides higher accuracy than the correlation methods that only use rank information.

In this study, we propose the use of the Gini correlation to infer regulatory relationships of genes from transcriptomic data. Using a compiled dataset that includes ~11,000 regulatory relationships from Arabidopsis, we systematically evaluated the performance of the GCC method and four other correlation methods, including the PCC, Tukey’s biweight (BiWt), the SCC and the KCC. We also assessed the compatibility and the consistency of these methods on RNA-Seq data. Our analyses indicate that the GCC has multiple advantageous merits, such as independent of distribution forms, more
capable of detecting non-linear relationships, more tolerant to outliers and less
dependence on sample sizes. Finally, we implemented the GCC correlation as an R
package named “rsgcc” to perform clustering analyses of transcriptomic data.

RESULTS

Compilation of a gene set with known regulatory relationships in Arabidopsis

To evaluate the performances of the Gini and the above-mentioned correlation
methods, we first compiled a set of genes that have documented regulatory relationships
according to the recently released protein interactome (Arabidopsis Interactome Mapping
Consortium, 2011) and the most updated TF-target interaction database (AtTFDB) that
contains >11,000 direct interactions collected from single gene studies and large-scale
ChIP-Chip/ChIP-Seq experiments (Yilmaz et al., 2011). In addition to the direct
interactions between the TFs and targets, the proteins that are physically interacting with
the TFs were defined as cofactors. Therefore, the compiled gene interaction dataset
includes 8,929 interactions between the TFs and targets, 1,428 interactions between the
TFs and cofactors and an addition of 772 interactions between the cofactors and targets,
which together cover 822 TFs, 6,287 target genes and 823 cofactors. Among those genes,
34.3% (2,159 of 6,287) of the target genes are regulated by more than one TF, and 32.7%
(269/822) of the TFs have the experimentally validated targets. Among these 269 TFs,
51.7% (139 of 269) of the TFs cooperate with at least one cofactor to regulate their target
genes (Supplemental Table S1). To fully include all of the possible regulatory
relationships by gene expression profiles, we downloaded the Affymetrix array data from
the AtGenExpress database, which contains 79 samples that were collected during
Arabidopsis development and includes major organs, such as the root, stem, leaf, whole
plant, apex, flower, floral organs, and seed, as well as tissues from various developmental
stages of each organ (Schmid et al., 2005). The expression dataset of approximately
7,000 genes from the 79 samples is sufficient to evaluate the performance of the above-
mentioned correlation methods based on the fact that these regulatory relationships may
be covered by a comprehensive expression profile during Arabidopsis development.

Evaluation of the overall performances of the five correlation methods
Using the compiled dataset, we first evaluated the overall performances of the proposed GCC method and the other four methods, which included the PCC, SCC, KCC and BiWt. Similar to the PCC, the BiWt calculates the correlation with the covariance and standard deviation from the expression values that are firstly weighted by the Tukey’s biweight estimation (Hardin et al., 2007). The correlations of the PCC, SCC and KCC methods were computed with the `cor.test` function in R. The correlation of the BiWt method was calculated using the `biwt` package in R. The statistical significance (p value) of each computed correlation was derived from 2,000 permutation tests by randomly shuffling the gene expression data of the analyzed gene pairs (Materials and Methods). Because the GCC method calculates the correlation of two variables based on one gene’s rank information and the other gene’s actual expression value, the GCC can produce two correlation coefficients (GCC1 and GCC2) for one gene pair (Materials and Methods). The two calculated correlations by reciprocally using the rank and value information are usually similar, as well as their p values (Supplemental Fig. S1). Hence, we chose the coefficient with the lower p value as the final GCC correlation.

We adopted the receiver operating characteristic curve (ROC) analyses to evaluate the performance, which can graphically illustrates the power of the classifier in distinguishing positive samples from negative samples with the changes of significance thresholds. The x-axis in ROC represents the fraction of detected false positives from the negative dataset (false positive rate, FPR), and the y-axis represents the fraction of detected true positives from the positive dataset (true positive rate, TPR). Thus, for a pair of TF and target genes, their actual gene expression data across the 79 conditions was considered as the positive sample, and the negative sample was constructed from the randomly shuffled expression profiles (permutation) of the tested TF-target pair. The permutation was repeated for 2,000 times and an empirical distribution of the correlations for the permuted TF-target pairs was built, in which each correlation can be associated with a p value by considering its probability under the empirical distribution. Then, for all TF-target gene pairs, the positive and negative samples were combined as the positive and negative datasets, respectively. At each possible significance level (p value) of correlations for the samples in the positive and negative datasets, we were able to use the
p value as cutoff to determine the TPR from positive dataset and the FPR from the negative dataset. Then, the TPRs and FPRs were imported to the R package “pROC” to visualize the ROC curve, representing the TPR against FPR at different significance level. The area under the ROC curve (AUC) is then computed as a quantitative measure of the overall performance; this measure ranges from 0.0 to 1.0. A higher ROC curve results in a larger AUC value and indicates a better resolution to distinguish the positive samples from negatives samples.

The ROC curves of the GCC correlation were always beyond the curves from the other four methods, no matter the analysis was performed in the TF-target, TF-cofactor and cofactor-target datasets (Fig. 1, A to C). The BiWt ranks at the second position followed by the SCC and KCC methods, while the PCC always has the lowest AUC values (Fig. 1, A to C). To confirm this pattern, the ROC analysis was repeated for 2,000 times within each class of interactions. The order of the distributions of AUC values drawn in a boxplot was consistent with that in the ROC curves generated from the five correlation methods (Fig. 1D). Therefore, although the overall performances are not dramatically different among the five methods, the GCC method slightly outperform the other four methods in inferring the expected regulatory relationships.

In addition, at a significance level of \( p = 0.05 \), the GCC method detected 5,969 pairs of known TF-target interactions, which was 19.48%, 10.19% and 2.74% higher than that of the PCC, SCC and BiWt methods, respectively (Supplemental Fig. S2). The GCC method was able to identify 96.14% (4,803/4,996) of the linear correlations that were derived by the PCC method and 94.85% (5,138/5,417) of the monotonic correlations that were derived by the SCC method (Supplemental Fig. S2). Moreover, the GCC method identified 332 correlated expressions of the TF-target pairs that could not be detected by either the PCC or SCC methods at the same significance threshold \( (p = 0.05) \) (Supplemental Fig. S2). These analyses demonstrated the ability of the GCC method to derive both linear and non-linear (monotonic) relationships between a TF and target. Additionally, the GCC method may also be capable of detecting new forms of the regulatory relationships that have been overlooked by the value-only PCC or the rank-only SCC methods. Similar analyses that were performed on the TF-cofactor and
cofactor-target gene pairs show consistent results (Supplemental Fig. S2). We also examined the performance of the GCC method with different strategies of choosing the $p$ value for the final output. Compared to the selection of lower $p$ value of GCC1 and GCC2, the GCC method detected less significant interactions when using the higher one or the average $p$ value of GCC1 and GCC2, corresponding to the number of 4222 and 4619 respectively for the TF-target interactions.

We provide three examples to display the properties of the GCC in capturing various expression relationships below.

The Gini correlation is able to detect both linear and non-linear regulatory patterns

The feature allowing the Gini method to detect more regulatory relationships is attributed to its ability to consider both the value and rank information when calculating correlations, compared with other methods that only use value or rank information. To validate whether the significant correlations that are only detected by the GCC are genuine, we manually inspected the expression patterns of several pairs of TF-target genes. First, we examined whether the GCC can derive a similar correlation from a globally correlated TF-target pair that has a linear relationship and its correlation computed by the PCC and the SCC methods. Figure 2A demonstrates the expression profile of an E2F3 transcription factor (AT2G36010) and its target HAC7 (AT5G56740) that encodes a histone acetyltransferase. A global linear relationship was exhibited across all of the 79 samples. Except the BiWt generating a higher correlation of 0.94, the other three methods, the GCC, PCC and SCC yielded a similar correlation of 0.88 with $p$ values $\leq 0.001$, indicating the correlation calculated by the GCC is equivalent to either the PCC or the SCC methods to infer linear relationship. The higher correlation yielded by the BiWt is attributed to the down-weight of the outlier points in its algorithm, but seemingly the BiWt overestimated the correlation.

We then examined another pair of TF and target genes, a bHLH transcription factor (PYE, AT3G47640) and FRO3 (AT1G23020) that encodes a ferric chelate reductase, which demonstrated a linear relationship among the 74 samples, but were
obviously uncorrelated in only 5 tissues in Seed (Fig. 2B). While the rank-based SCC and the rank- and value-based GCC methods derived a similar correlation of 0.50 with significant $p$ values $\leq 0.001$, the correlation computed by the PCC was only 0.22, and the $p$ value of 0.06 was no longer considered a significant relationship. This case indicates that the performance and accuracy of the PCC can be greatly reduced due to only a few outlier samples if the PCC correlation is computed across all of the samples. Conversely, the GCC and SCC, which take the rank information to calculate the correlation, are more capable of tolerating outlier data. Again, the BiWt yielding a higher correlation of 0.68 likely overestimated the correlation.

Because many transcription factors regulate their targets only in specific tissues or under specific environmental conditions, we were interested in whether the transient regulatory relationships could be detected using these four methods. A well-known transcription factor, LFY (AT5G61850), which controls flowering time and meristem development, is specifically expressed in Apex where a correlation with its target MUP24.5 (AT5G60630) was observed (Fig. 2C). While the PCC, SCC and BiWt computed an insignificant correlation of 0.14, 0.06 and -0.07, respectively, the GCC could still successfully detect this regulatory relationship with a significant correlation of 0.56 with a $p$ value $\leq 0.001$. More interestingly, because the GCC can compute two correlations (GCC1 and GCC2) reciprocally using the rank and value information of a pair of variables, one of the GCCs may be more significant than the other GCC; this is especially true in biology when the correlated expression of a TF and its target only exhibits in a small subset of samples, and in majority of the samples their expression patterns are not concordant and/or very different in expression levels. In the third case, there is a correlation between MUP24.5 and LFY in Apex, while the expression of MUP24.5 was slightly and significantly higher than that of LFY in Root and Seed, respectively (Fig. 2C). Hence, if using the rank information of LFY, the expression values of MUP24.5 are not in a full agreement with the LFY’s rank information, thus generating a low correlation (GCC1 = -0.19, $p$ value = 0.23). However, if using MUP24.5’s rank information, LFY’s expression values in Apex consecutively fit in the MUP24.5’s rank in Apex, thus generating a significant correlation (GCC2 = 0.56, $p$ value
This feature of the GCC method actually compensates for the shortcoming of the PCC and SCC that may only derive a global linear or monotonic relationship in the majority of samples, no matter using values or ranks to compute the correlation. The BiWt might have down-weighted the 10 correlated samples in Apex as outliers and generated the lowest correlation. Therefore, the GCC is more capable of detecting transient interactions (or partial concordances) that occur in a minority of samples, while the other methods require the majority of samples are correlated to derive a significant correlation between a TF and target.

Evaluation of the tolerance to outlier data points by the five correlation methods

The outlier data points in a gene’s expression profile refer to the extremely high or low expression values in a subset of samples. In reality, these genes are of more interest because of their tissue-specific expression behavior, and the correlated strength is expected to be persistent regardless of the number of samples in which the gene is specifically expressed. However, the value-based PCC method is not stable to outliers, and the existence of a small number of outliers may affect the derivation of accurate correlations, as illustrated in the pair of PYE and FRO3 (Fig. 2B). In this analysis, we tested the ability of the five correlation methods in terms of their consistency towards the number of outlier data points. First, we defined the outliers in the compiled dataset using the following criterion: outliers are classified as data points 1.5 times the interquartile range (IQR = Q3 - Q1, Q3 and Q1 represent the 75% quartile and 25% quartile, respectively) and above the 75% quartile or below the 25% quartile. There were about 83% (9,325/11,129) gene pairs in the compiled dataset contain outlier data points. Because the number of tested gene pairs significantly decreases with an increase in the number of outliers, we only tested the performance of the correlation methods on gene pairs with 0, 1~5, 6~10 and >10 outliers (Supplemental Fig. S3).

Within each range of outlier numbers, we again performed the ROC analysis for each type of interactions and repeated the test for 2,000 times to generate 2,000 AUC values for each correlation method. Distributions of these AUC values are shown in a boxplot (Fig. 3). Overall, the AUC values of all the five methods dropped with the
increase of outliers, suggesting these methods are all influenced by outliers. The PCC shows the most dramatically reduced performance. For the TF-target pairs without including any outliers, the average AUC value of the PCC was 0.88. However, when 6~10 outliers existed, the average AUC value decreased to 0.77 (Fig. 3 and Supplemental Table S2). The performance of another value-based correlation method BiWt was also greatly affected by the increase of outliers. The average AUC value of the BiWt dropped from 0.90 to 0.81 when the number of outliers increased from zero to more than 10 in TF-target gene pairs (Fig. 3 and Supplemental Table S2). Compared to the value-based methods, the rank-based methods (SCC and KCC), and the value- and rank-based GCC method are more tolerant to outliers on TF-target gene pairs (Fig. 3 and Supplemental Table S2). A similar analysis was also performed on the TF-cofactor and the cofactor-target gene pairs (Fig. 3). Overall, the GCC and BiWt showed better performance than other tested methods to infer regulatory relationships from microarray data, no matter the gene expression profiles contained outliers or not. Compared to the BiWt, the GCC could achieve higher AUC values and be more tolerant to outliers for most cases. The robustness of the GCC to tolerate outliers may be mainly attributed to its feature that uses the rank information if the expression levels of outliers are extremely deviated from the center of data distribution.

The influence of sample size on the performance of the five correlation methods

The number of samples (sample size) is another critical issue that may greatly affect the power of many statistical methods, such as differential expression call (Jorstad et al., 2007). While a relatively small sample size may lead to a higher false positive rate of detecting differentially expressed genes, a larger sample size is usually required to perform sound tests. Inspired by the concern that a rank-based method might need a minimum amount of variables to derive a correct rank order to calculate correlations, we investigated whether the sample size would affect the performance of the correlation methods studied. To conduct this analysis, we firstly selected a TF and target pair, AGL9 (AT1G24260) and Hsp40 (AT3G04960), which demonstrated a global linear relationship with a Spearman correlation of 0.93 (p value ≤ 0.001) and a Pearson correlation of 0.94 (p value ≤ 0.001) (Fig. 4A and 4B). We then computed the correlation coefficients on the
simulated gene pairs with 5 to 75 samples randomly selected from the real gene pairs using the five correlation methods. This process was repeated 1,000 times and an average correlation coefficient was calculated for each sample size. The reason that we selected a globally correlated TF-target pair was to minimize the chance that the non-correlated data points were selected and biased the evaluation. The PCC, GCC and BiWt methods that consider value information could derive similar correlations regardless of the increase in the sample size from 5 to 75 (Fig. 4C). Conversely, the correlations calculated by the rank-only SCC method gradually increase from 0.84 to 0.94, which indicates a strong dependence of the SCC method on the sample size to derive the expected correlation (Fig. 4C). The result from another rank-based method, the KCC, is much lower than the expected value of 0.94 (Fig. 4C).

To further confirm this pattern derived from the study of one gene pair, we then performed a similar analysis on 75 gene pairs with the Pearson correlations higher than 0.80. For each gene pair, the differences between the average correlations on simulated gene pairs and the Pearson correlation on the real gene pair were computed in order to estimate the influence of sample size on the performance of the five methods. The boxplots of these differences for 5~75 randomly selected samples are shown in Figure 4D. Consistent with the results on gene pair AGL9 and Hsp40, the accuracy of the SCC method was largely dependent on the sample size to derive a meaningful rank order and to properly calculate the correlations (Fig. 4D). In contrast to the SCC, the dependence of the BiWt on the sample size is relatively small (Fig. 4D). The GCC, KCC and PCC could yield stable correlations when the sample size was increased from 5 to 75, indicating that their dependences on the sample size are minimal. Noting that the KCC correlations are much lower than the expected values, this may be caused by the fact that the KCC method calculates the difference between the probability of concordance and discordance obtained from the rank information of all possible pairs of data points.

The compatibility of the five correlation methods on RNA-Seq data

The RNA-Seq technology has greatly accelerated the production of transcriptomic data in biology, without relying on whole-genome sequences or pre-collected cDNA sequences. However, concerns have been raised regarding whether the current analytic
methods and tools developed for microarray platforms can be directly applied to RNA-
Seq data, because the data properties between microarray and RNA-Seq are naturally
different (Wang et al., 2009). Therefore, we further evaluated the five correlation methods
on RNA-Seq data, performed on both read counts per gene and FPKM (fragments per
kilobase per million reads) values – the two popular measurements of gene expression
abundance. Compared to read counts, the FPKM produced from the Cufflinks RNA-Seq
analyses pipeline is generally considered as a more reasonable measure to quantify gene
expression levels because the bias caused by the gene length and sequencing depth is
normalized to perform comparable between-sample analysis (Trapnell et al., 2010;
Garber et al., 2011).

We first plotted the distributions of the read counts and the FPKM values from the
RNA-Seq datasets from Arabidopsis (GEO accession: GSM838184 and GSM764078), in
which both forms of data were not normally distributed featuring with long heavy tails
(Fig. 5A). From each distribution, we generated 2,000 pairs of genes for each number of
simulated samples \(5 \leq N \leq 100\) with an expected correlation coefficient of 0.7 using the
copulas function in MATLAB. Then, the correlation coefficients from the 2,000 gene
pairs were computed using the five correlation methods. The average correlation
coefficient within a -0.05 to +0.05 deviation range was considered as the expected
correlation. When computing the correlations on the read counts or the FKPMs, the GCC
yielded an average correlation coefficient within the expected range, followed by the
correlations computed by the SCC (Fig. 5B). While the average correlations computed by
the PCC and BiWt methods were below the expected range (Fig. 5B). Moreover, we also
found the performances of the BiWt and PCC decreased with the increase of sample size.
We speculate that the low effectiveness of the PCC and BiWt may be due their strict
dependence on a normal distribution to derive a correct mean and standard deviation. On
the contrary, the rank-based GCC and SCC have better performances on non-normally
distributed RNA-Seq data.

Considering that most methods in microarray analyses use log transformation to
scale the expression intensities to a proximal normal distribution, we next evaluated the
five methods on the log2-transformed read counts and the FPKMs. As for the read counts, the average correlations calculated by the GCC, PCC and BiWt were all close to the expected 0.7 value (Fig. 5C). The performance of the SCC was not improved, since log transformation does not change the ranks of gene expression (Fig. 5C). When computing the correlations on log2-transformed FPKM values, the results from the GCC method is approximate to the expected value, while the average correlations from the PCC and BiWt was slightly below 0.7 (Fig. 5C). We speculate that calculating FPKM values from genes with very few reads may generate FPKM values below 1, and the log transformation resulted in negative values that influenced the accuracy of correlations. Collectively, our analyses showed that the GCC, PCC and BiWt methods were equally effective on log2-transformed read counts which can be recommend as the most optimal data form in RNA-Seq analysis.

Evaluation of the five methods using 29 TF-target genes in maize RNA-Seq data

Finally, we evaluated the five methods using a recently published RNA-Seq dataset in maize, containing 13 samples that included eight male and female reproductive tissues, four tissues from developing seeds, and one leaf tissue (Davidson et al., 2011). Analyses were performed on three transcription factors (OC1, WRI and MRP1) functioning in these reproductive tissues and regulating 29 known target genes; this information was collected from a literature search (Supplemental Table S3). Correlations using these 29 pairs of TFs and targets were computed on the FPKM values using the five correlation methods (Supplemental Fig. S4) and the statistical significance (p value) was determined by the 2,000 permutation tests. The BiWt failed to calculate the correlations for six pairs of TF and target genes, since a number of samples had zero reads for these genes. The BiWt totally identified only 7 TF-target interactions, which were much fewer than the numbers of TF-target interactions detected by the SCC (12), KCC (13), PCC (13) and GCC (17) methods with p=0.05 cutoff (Supplemental Fig. S4). All these five methods failed to detect 12 pairs of TF-target interactions, including ten pairs co-expressed in only one sample in either “Endosperm” or “Anthers”, and two pairs exhibiting a lagged co-expression pattern (Supplemental Fig. S5). These results
indicated that all these five methods require the regulatory relationships to concordantly present in at least two samples to be detectable.

Implementation of the Gini correlation as an R package for gene expression clustering analyses of transcriptomic data

Clustering analyses of gene expression patterns are an essential part of large-scale transcriptome analyses, which are usually performed with hierarchical clustering methods that measure the distance between two genes based on correlation coefficients (D’Haeseleer, 2005). We implemented the Gini correlation method and other four correlation methods evaluated in this study in an R package, named “rsgcc” to perform clustering analyses of both microarray data and RNA-Seq data based on either read counts or FPKM values and to visualize the clustered gene expression pattern using a heatmap. This package is also capable of performing parallel computing to increase the speed of the calculation of correlation coefficients for thousands of genes via the implementation of the snowfall package in the R environment. Additionally, we also provided a user-friendly interface using the gWidgetsRGtk2 package in R, which allows the users to perform analyses via a series of mouse actions without command-line based R programming (Fig. 6). The “rsgcc” package allows a user to easily select different correlation and clustering methods, to specify the number of CPUs for parallel computing and to choose the color-scales for heatmap visualization (Fig. 6). In the current version of “rsgcc”, three types of distance measurements (raw correlation: 1-coef; absolute correlation: 1-|coef|; squared correlation: 1-|coef|^2. coef: correlation coefficient) and seven clustering methods (complete-linkage, average-linkage, median-linkage, centroid-linkage, McQuitty-linkage, single-linkage and ward-linkage) are provided for users to select a variety of clustering methods.

Although the “rsgcc” is efficient in performing clustering analyses on all of the genes in a genome by taking advantage of parallel computing, preselecting a group of differentially expressed genes identified by Cufflinks or EdgeR or a group of tissue-specifically expressed genes to generate the clustered heatmap is highly recommended. Therefore, we provided a function in “rsgcc” to select tissue-specific genes by calculating
a tissue-specificity (ts) score for each gene. The detailed ts algorithm is described in the online manual of “rsgcc”. To demonstrate the function of “rsgcc”, we first used the “find ts-gene” function to select a group of 2,279 tissue-specifically expressed genes out of the 39,456 genes from the RNA-Seq data profiled in 13 reproductive samples in maize (Davidson et al., 2011). Then, using the GCC-based similarity measure, a heatmap of the 2,279 clustered, tissue-specific genes was generated by “rsgcc” in which the genes specifically expressed in the same tissue were successfully clustered in one group (Fig. 6). The clustered gene expression pattern can be saved in a standard output of hierarchical clustering result: the “CDT” format, which can also be visualized and analyzed using the TreeView program (Saldanha, 2004). The “rsgcc” package and manual documents can be freely accessed from the CRAN (The Comprehensive R Archive Network) at http://cran.r-project.org/web/packages/rsgcc.

CONCLUSION

In this study, we compared the five correlation methods including the Pearson correlation, the Spearman correlation, the Kendall correlation, the Tukey’s biweight correlation and the Gini correlation, in terms of the effectiveness in inferring regulatory relationships from gene expression data. The evaluation analyses were performed based on known TF and target interactions collected from Arabidopsis and maize. Among these methods, the Gini correlation coefficient was first time introduced in plants to analyze the transcriptomic data produced from microarray and RNA-Seq platforms. Compared to the other four methods, one of the unique features for the Gini correlation is that its algorithm reciprocally considers the value and rank information of a pair of TF and target, making the Gini correlation less dependent on the form of data distribution. This merit allows the Gini correlation to identify non-linear relationships between TFs and targets, and transient interactions occurred in a small subset of samples, which might be missed by the methods that only globally consider value or rank information from all the samples. The robustness of the Gini correlation is also reflected from its higher tolerance of outlier data points and less dependence on sample size.
Application of the Gini correlation provides an alternative for biologists to analyze gene expression data. We implemented the Gini correlation as an R package to perform clustering analyses based on microarray data and RNA-Seq data. Additionally, this package can be also applied to construct gene co-expression networks (GCNs) and to perform network analyses on other types of interaction data. For instance, the “rsgcc” package is available to be called by the WGCNA (weighted gene co-expression network analysis) package in the R environment (Langfelder and Horvath, 2008). This package can also be incorporated to the Cytoscape software as a plugin for broader utilization in network visualization and network analysis in biology (Smoot et al., 2011). Moreover, the Gini-based methodologies are a system of mathematical solutions including the Gini correlation, Gini mean difference, Gini index, Gini covariance and Gini regression that can be used for a variety of purposes when analyzing the data that is not distributed normally and is widely used in other disciplines, such as economics, physics, informatics and sociology. Therefore, the Gini methodological systems have a promising prospect to model the complexity of biological systems.

**MATERIALS AND METHODS**

**Microarray and RNA-Seq data sets**

The microarray gene expression data was downloaded from the AtGenExpress database (http://www.weigelworld.org/resources/microarray/AtGenExpress/), which includes 79 samples that were collected during Arabidopsis development. The microarray data was generated with the Affymetrix ATH1 array platform and has been normalized with the gcRMA method. More details about this microarray dataset can be referred in (Schmid et al., 2005).

The maize RNA-Seq data was obtained from (Davidson et al., 2011), which contains 13 samples from eight reproductive tissues, four tissues from developing seeds, and one leaf tissue. The sequence reads of these tissue samples were firstly generated using the Illumina sequencing platform and then aligned to the maize genome (B73) by using the Bowtie and TopHat alignment tools with the limit of intron length ranged from 5bp to 60,000 bp. The normalized gene expression levels in the FPKM format were
finally calculated with the Cufflinks software. The detailed information about this RNA-Seq dataset can be found in the reference (Davidson et al., 2011).

Computation of the Gini correlation coefficient

The Gini correlation coefficient is a well-defined measure to quantify the correlation between two variables following normal and/or non-normal distributions (Schechtman and Yitzhaki, 1999; Yitzhaki, 2003). As the GCC method reciprocally utilizes the value information of one variable and the rank information of the other variable, it can produce two correlation coefficients. For a given gene pair \((X, Y)\), one Gini correlation coefficient is defined as

\[
GCC(X, Y) = \frac{\sum_{i=1}^{n} (2i-n-1) \cdot x(i, Y)}{\sum_{i=1}^{n} (2i-n-1) \cdot x(i, X)},
\]

where \(n\) is the sample size (i.e., number of observed gene expression values), \(x(i, X)\) is the \(i^{th}\) value of gene expression profile \(X\) sorted in an increasing order, here \(x(1, X) \leq x(2, X) \leq \ldots \leq x(n, X)\). \(x(i, Y)\) is the corresponding value of \(X\) in the gene pair \((X, Y)\) for the \(i^{th}\) value of gene expression profile \(Y\) sorted in an increasing order.

The other GCC value can be given as

\[
GCC(Y, X) = \frac{\sum_{i=1}^{n} (2i-n-1) \cdot y(i, X)}{\sum_{i=1}^{n} (2i-n-1) \cdot y(i, Y)},
\]

Where \(y(i, X)\) and \(y(i, Y)\) are respectively defined similarly as \(x(i, Y)\) and \(x(i, X)\) in formula (1).

According to formula (1) and (2), correlations of GCC method can be interpreted as the differences between two curves weighted by the information deriving from the rank order of gene expression data. The \(x(i, Y)\) and \(y(i, X)\) were represented as the red or blue curves in the last two columns, whereas \(x(i, X)\) and \(y(i, Y)\) were represented as the black curves in these columns in Figure 2.

Determining statistical significance

The statistical significance \((p\text{-value})\) of the correlation was computed with the permutation test method (Qian et al., 2001; Wang et al., 2008). For one given gene pair...
and correlation method, the p value was calculated as follow: (1) Computing the
correlation r on the real paired expression values. (2) Constructing a permuted gene pair
by randomly shuffling gene expression data in different samples and re-computing the
correlation on the permuted gene pair. (3) Repeating step (2) for a large number of times
(N = 2,000), an empirical distribution (H₀) of the correlations on the permuted gene pairs
is then generated. (4) Calculating the statistical significance of the correlation r under the
empirical distribution H₀ with the formula: p value= 2× m/N, here m denotes the times
that the absolute value of the correlation on the shuffled data greater than that of
correlation on the real data.

REFERENCES


Berri S, Abbruscato P, Faivre-Rampant O, Brasileiro AC, Fumasoni I, Satoh K,
of WRKY co-regulatory networks in rice and Arabidopsis. BMC Plant Biol 9:
120

1499-1501

Davidson RM, Hansey CN, Gowda M, Childs KL, Lin HN, Vaillancourt B, Sekhon
sequencing for analysis of maize reproductive transcriptomes. The Plant Genome
4: 13

Ficklin SP, Feltus FA (2011) Gene coexpression network alignment and conservation of
gene modules between two grass species: maize and rice. Plant Physiol 156:
1244-1256


**FIGURE LEGENDS**

**Figure 1.** Assessment of the overall performances of the five correlation methods evaluated by the ROC analyses. The ROC curves were plotted for the GCC, PCC, SCC, KCC and BiWt using the datasets of TF-target (A), TF-cofactor (B), and cofactor-target (C) gene pairs with a 1:1 ratio of positive and negative samples. (D) The boxplot of the AUC values derived from the ROC analysis repeated for 2,000 times. The ‘cof.’ represents cofactor.

**Figure 2.** The GCC can detect the regulatory relationships missed by the PCC, SCC and BiWt methods. (A) The GCC can detect a linear relationship with similar correlation values to the PCC and SCC correlations. (B) The PCC failed to infer the relationship in the samples containing outliers, which was detected by the GCC, SCC and BiWt. The 5 outlier samples were represented by the red circles in the gray region in Seed. (C) The GCC was able to identify transient interactions that were overlooked by the PCC, SCC and BiWt. The expression values of TF and target are only correlated in 9 samples in Apex out of the 79 samples (red circles in the gray region). Two correlations, GCC1 and GCC2 are produced by the GCC reciprocally using the rank and value information of the two genes’ expression data. In the last two columns, the expression data of genes sorted with their own rank information are displayed as black dashed curves, while the expression data of genes sorted with the other gene’s rank information are shown as blue or red solid curves. The Gini correlation can be explained as the difference between the solid and dashed curves weighted by the rank information. ‘Value’ and ‘Rank’ denote the value and the rank information of the gene expression data, respectively.

**Figure 3.** Assessment of the influences from the outlier data points on the five methods. The performances of the correlation methods were evaluated on the pairs of TF-target, TF-cofactor and cofactor-target gene sets with influences of 0, 1-5, 6-10 and >10 outliers by the ROC analyses.
Figure 4. Assessment of the influences from sample size on the five methods. (A and B) A significant linear relationship exists between the gene expression profiles of the TF-target (AGL-Hsp40) gene pair. (C) The average correlation of the different correlation methods for 1,000 gene pairs with 5~75 samples randomly selected from the gene expression profiles of the AGL-Hsp40 gene pair. (D) The differences between the correlations computed on simulated gene pairs (sCor) and the Pearson correlations on the real gene pairs (rCor), computed with the formula (sCor-rCor)/rCor. ‘Value’ denotes the value of the gene expression data.

Figure 5. The compatibility of the five correlation methods on RNA-Seq data. (A) The kernel density estimation of the read counts and the RPKM values were generated from the real Arabidopsis RNA-Seq data (GEO accession: GSM838184 and GSM764078). (B) The average correlation coefficients of each method were calculated from 2,000 random gene pairs with an expected correlation coefficient of 0.70 across the 5 to 100 simulated samples. The ‘Ave correlation’ represents the average correlation coefficients.

Figure 6. A screen shot of the GCC-based R package (rsgcc) for correlation and clustering analyses of gene expression data. The rsgcc was applied to cluster ~2,800 tissue-specifically expressed genes in maize RNA-Seq data. ‘DAP’ represents the days after pollination, ‘pre-em’ is pre-emergence, ‘post-em’ indicates post-emergence, ‘endo’ is endosperm and ‘ts-genes’ are tissue-specific genes.
Supplemental data

**Supplemental Figure S1.** Scatter plots of GCC1 versus GCC2 for TF-target, TF-cofactor, cofactor-target gene pairs. The dots in gray represent the non-significant correlations of GCC1 and GCC2 ($p > 0.05$). The dots in yellow indicate only the GCC1 or GCC2 with significant correlation ($p \leq 0.05$). The dots in blue denote that both the GCC1 and GCC2 with significant correlation and the correlations of the PCC or SCC are also significant. The dots in red represent the correlations of both the GCC1 and GCC2 are significant but the correlations of the PCC and SCC are not significant.

**Supplemental Figure S2.** The Venn diagrams of detected regulatory relationships with GCC, SCC, PCC and BiWt at the significance level of $p=0.05$.

**Supplemental Figure S3.** Number of gene pairs versus number of outliers.

**Supplemental Figure S4.** Correlations of 29 TF-target gene pairs computed by the five correlation methods. The red colors indicate the correlations with the significance of $p \leq 0.05$. The stars denote that two correlations of the Gini method (GCC1 and GCC2) with significance of $p \leq 0.05$. The arrows represent the correlations of gene pairs failed to be calculated by the BiWt method. The annotation information of these 29 gene pairs were given in Supplemental Table S3. The GCC can detect the TF-target interaction for GP11, as the co-expression occurred in more than 10 samples.

**Supplemental Figure S5.** Gene expression profiles of 12 TF-target gene pairs missed by all tested correlation methods. The TF and target information of GP4-GP10, GP12-GP16 were given in Supplemental Table S3. DAP, days after pollination.

**Supplemental Table S1.** Statistics of the complied dataset including known transcriptional regulation relationships in Arabidopsis collected from AtTFDB database.
Supplemental Table S2. The influence of outliers on the performance of different correlation methods.

Supplemental Table S3. List of the 29 TF-target interactions from maize.
A. 

BiWt=0.94 (P≤0.001)
SCC=0.88 (P≤0.001)

log2 value

---

B. 

BiWt=0.68 (P≤0.001)
SCC=0.50 (P≤0.001)

log2 value

---

C. 

BiWt=-0.07 (P=0.46)
SCC=0.06 (P=0.56)

log2 value
A

SCC = 0.93 (P ≤ 0.001)

B

PCC = 0.94 (P ≤ 0.001)

C

Average correlation

D

Differences between correlations on simulated gene pairs and Pearson correlation on real gene pairs

GCC

PCC

KCC

GCC2

SCC

BiWt

5 to 75 randomly selected samples

5 to 75 randomly selected samples
**A**

Read counts FPKM

**B**

Reads counts FPKM

**C**

log2(Reads counts) log2(FPKM)

---

GCC1 GCC2 PCC

SCC KCC BiWt
Step 1: Load gene exp data
- Click here to load

- Display loaded data
- Find and cluster ts-genes

Step 2: Select a correlation method
- Gini correlation
- Spearman correlation
- Kendall correlation
- Pearson correlation
- Turkey's biweight

Step 3: Specify a distance measure
- raw correlation (1-coef)
- absolute correlation (1-|coef|)
- squared correlation (1-|coef|^2)

Step 4: Choose a cluster method
- complete linkage

Step 5: CPUs for correlation calculation
- 1

Start to run

Adjust colors for heat map
Color for max GE value: 16776001
Color for median GE value: 1
Color for min GE value: 16776001

Save correlations and cluster data