


## Compositional correlation analysis of gene expression time series

\*<sup>1</sup>Fatih Dikbaş

<sup>1</sup>Pamukkale University, Civil Engineering Department, Denizli, Turkey

[f\\_dikbas@pau.edu.tr](mailto:f_dikbas@pau.edu.tr) 

### Abstract

Accurate determination of temporal dependencies among gene expression patterns is crucial in the assessment of functions of genes. The gene expression series generally show a periodic behavior with nonlinear curved patterns. This paper presents the determination of temporally associated budding yeast gene expression series by using compositional correlation method. The results show that the method is capable of determining real direct or inverse linear, nonlinear and monotonic relationships between all gene pairs. Pearson's correlation values between some of the gene pairs have shown negative or very weak relationships ( $r \approx 0$ ) even though they were found to be strongly associated. Inversely, a high positive  $r$  value was obtained even though the genes are inversely related as determined by the compositional correlation approach. Comparisons with Pearson's correlation, Spearman's correlation, distance correlation and the simulated annealing genetic algorithm maximal information coefficient (SGMIC) have shown that the presented compositional correlation method detects important associations which were not found by the compared methods. Supplementary materials containing the code of the used software together with some extended figures and tables are available online.

**Keywords:** Combinatorics, Compositions of  $n$ , Compositional correlation, Gene expression association, *Saccharomyces Cerevisiae*

### 1. INTRODUCTION

"The most merciful thing in the world, I think, is the inability of the human mind to correlate all its contents" wrote H. P. Lovecraft at the beginning of his cult story "The Call of Cthulhu" [1]. This was long before human mind managed to invent supercomputers to try to calculate correlations among data sets (huge or small) by using various correlation coefficients. Pearson's Correlation Coefficient (simply called correlation or  $r$ ) [2] - which was introduced when Lovecraft was only five years old - might still be the most widely used statistical measure for assessing relationships between data series.

In its nature, Pearson's correlation is a measure of linear association. Currently it is very hard to find a single issue of a scientific journal that does not include the word 'correlation'. Pearson's correlation is also used in the analysis of high-throughput data (such as genotype, genomic, imaging, and others) [3, 4], although the relationships are generally nonlinear. This tendency of using Pearson's correlation for non-linear associations still widely exists in literature despite clear warnings about its improper use: Correlation is misleading [5]; good correlation does not automatically imply good agreement [6]; risk of producing spurious correlations when analyzing non-independent variables is very large [7, 8]. The unintended and generally unnoticed misleading results are caused by the approach used in calculation of the correlation itself where the

averages of the whole series are used for assessing relationship. In fact, the average value of a data series is a single value which does not reflect the variations within the data series. In fact, the variations in data might have great importance in the determination of associations with other data series. Unfortunately, in association studies, there is still an inability in completely correlating the contents of data sets caused by inappropriate implementation of the currently used approaches or the inappropriate methodology of the used approach itself. Gene expression over time is a continuous process and can be considered as a continuous curve or function [9]. Genome-wide association studies try to determine associated gene pairs by comparing the expression series of each gene [10, 11]. Most of the studies use Pearson's correlation for attempting to find associations among the genes by comparing the expression series but Pearson's correlation is a measure of linear association and gene expression series generally show a periodic behavior with nonlinear curved patterns. Therefore, it is not surprising that the widely used Pearson method is generally reported to be less efficient than the compared methods in finding gene pairs of multiple relationships. It must be kept in mind that the efficiencies of different methods vary with the data properties to some degree and a pre-analysis is generally advised to identify the best performing method. A comprehensive comparison of gene association methods was provided by Kumari et al. [12].

The compositional correlation method used in this study takes its name from the term composition in number theory and combinatorics. The details of the method are presented

in the Materials and Methods section. Compositional correlation is based on the foundations of the two-dimensional correlation method developed by the author for assessing the degree and direction of relationships between matrices [13, 14]. These methods were developed when it was noticed by the author that, in some cases, the Pearson's correlation value decreases even though the estimations become closer to the observations in the hydrological modelling studies. Instead of considering the averages of the compared series, the compositional correlation approach considers the averages of all parts of all possible compositions of the data series. The comparison plots of calculated compositional variance, covariance and correlation values generate clouds that allow a comprehensive visual inspection of all obtained results on a single graph. The variance and covariance clouds also provide an opportunity for the comparison of the results with the Pearson's correlation. The purpose of this study is to present the implementation of the compositional correlation method in the determination of the yeast genes sharing similar temporal expression patterns. The aim was to provide strong clues for determining the functions of undefined yeast genes. The general trends of molecular events are directly associated with the timing of global gene expression patterns. Therefore, determination of the genes sharing similar temporal expression patterns is the first important step in the validation of functional implications of the inferred expression patterns [15].

Understanding the temporal relationships between the expression profiles of genes is crucial in determining the causes, functions and consequences of the biological processes like the cell cycle [16]; identifying the roles of genes in the stages of developmental processes of organisms [17, 18]; determination of genetic relatedness among various species [19]; investigating the functions of individual genes by exploring genetic interactions [20], and developing drugs to cure diseases by identifying genes that act in response to a certain disease [21]. Consequently, as the presented results also suggest, the compositional correlation method seems to be a very appropriate method for finding the associated genes by a complete comparison of the expression series, a task which is impossible to be made manually because of thousands of genes to be compared.

**2. MATERIALS AND METHODS**

This study presents for the first time in literature, the implementation of the compositional correlation method for gene expression time series data where the association levels of all yeast (*saccharomyces cerevisiae*) genes are determined. The details of the data used in the study are provided in the Results section below. The first introduction of the compositional correlation method was made in an association study between polynomial functions for which the Pearson's correlation failed because of nonlinearity of the polynomials [22]. The previous findings have shown that the compositional correlation method determines both the inversely and directly related portions of the examined functions. Therefore, gene expression series become very appropriate observations for the compositional correlation method because of their nonlinear structure which also

sometimes show an alternating (sometimes increasing and sometimes decreasing) behavior.

The main idea of the compositional correlation is that the association between two series might be better defined by the cumulative contribution of their parts but not the averages of the whole series especially when the series have varying (nonlinear, alternating, periodic etc...) behavior. If A is any set of positive integers, a composition of n with parts in A is an ordered collection of one or more elements in A whose sum is n [23]. The integer n is the number of observations in one of the compared series in the correlation case.

In number theory and combinatorics, a partition of a positive integer n, also called an integer partition, is an expression representing n as a sum of positive integers [24-27]. If order matters, which is also the case in the gene expression time series, the sum becomes a composition.

Each component of a composition is called a part of the composition. For example, the compositions of 3 are [1, 1, 1], [1, 2], [2, 1] and [3]. Similarly, if a sample data series has n (a positive integer) elements, the compositions of the series can be determined by dividing the series into parts. Each part should have at least two elements ( $m \geq 2$ ) for calculating compositional correlation. The compositions for  $2 \leq n \leq 10$  with parts  $\geq 2$  are shown in Table 1. The number of elements in each part are shown in brackets and the total number of compositions,  $t_n$ , is shown in the right column.

The number of possible compositions increases rapidly with n. The total numbers of compositions shown in the right column of Table 1 is a Fibonacci sequence ( $t_n = F_{n-1}$ ). The Fibonacci numbers are defined by  $F_{n+1} = F_n + F_{n-1}$  where the rate  $F_{n+1} / F_n$  rapidly tends to the golden ratio known as  $\phi = (1+\sqrt{5})/2 = 1.618...$  [28]. This means that there is golden ratio between the total number of compositions for two consecutive integers when the minimum number of observations in each part is equal to 2 ( $m = 2$ ) [29].

**Table 1.** All possible compositions with parts  $\geq 2$  for  $2 \leq n \leq 10$

n	All possible compositions with parts $\geq 2$	$t_n$
2	[2]	1
3	[3]	1
4	[2, 2]; [4]	2
5	[2, 3]; [3, 2]; [5]	3
6	[2, 2, 2]; [2, 4]; [3, 3]; [4, 2]; [6]	5
7	[2, 2, 3]; [2, 3, 2]; [2, 5]; [3, 2, 2]; [3, 4]; [4, 3]; [5, 2]; [7]	8
8	[2, 2, 2, 2]; [2, 2, 4]; [2, 3, 3]; [2, 4, 2]; [2, 6]; [3, 2, 3]; [3, 3, 2]; [3, 5]; [4, 2, 2]; [4, 4]; [5, 3]; [6, 2]; [8]	13
9	[2, 2, 2, 3]; [2, 2, 3, 2]; [2, 2, 5]; [2, 3, 2, 2]; [2, 3, 4]; [2, 4, 3]; [2, 5, 2]; [2, 7]; [3, 2, 2, 2]; [3, 2, 4]; [3, 3, 3]; [3, 4, 2]; [3, 6]; [4, 2, 3]; [4, 3, 2]; [4, 5]; [5, 2, 2]; [5, 4]; [6, 3]; [7, 2]; [9]	21
10	[2, 2, 2, 2, 2]; [2, 2, 2, 4]; [2, 2, 3, 3]; [2, 2, 4, 2]; [2, 2, 6]; [2, 3, 2, 3]; [2, 3, 3, 2]; [2, 3, 5]; [2, 4, 2, 2]; [2, 4, 4]; [2, 5, 3]; [2, 6, 2]; [2, 8]; [3, 2, 2, 3]; [3, 2, 3, 2]; [3, 3, 2, 2]; [3, 2, 5]; [3, 3, 4]; [3, 4, 3]; [3, 5, 2]; [3, 7]; [4, 2, 2, 2]; [4, 2, 4]; [4, 3, 3]; [4, 4, 2]; [4, 6]; [5, 2, 3]; [5, 3, 2]; [5, 5]; [6, 2, 2]; [6, 4]; [7, 3]; [8, 2]; [10]	34

### 2.1. Calculation of Compositional Correlation

The name of the compositional correlation method is based on the term composition in number theory and its value is calculated by using compositional variance and compositional covariance [22]. Compositional variance is a cumulative measure of how far the numbers in a time series spread from the averages of the part they belong in a composition. The compositional variance of a scalar time series for any composition with  $k$  parts is defined by the following equation:

$$\text{Var}_c(A) = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (A_{i,j} - \bar{A}_i)^2}{n} \quad (1)$$

In the above equation:

A: A scalar vector;

$n$ : The number of data in A;

$\text{Var}_c(A)$ : The compositional variance of the vector [A] for the current composition;

$k$ : The number of parts in the current composition for which the correlation is being calculated;

$n_i$ : The number of data in part  $i$ ;

$A_{i,j}$ : The  $j^{\text{th}}$  data in  $i^{\text{th}}$  part of vector A;

$\bar{A}_i$ : The arithmetic mean of the  $i^{\text{th}}$  part of vector A;

Compositional covariance is a measure of how changes in the part averages of a time series are associated with changes in the part averages of a second time series. This approach enables a better consideration of the contribution of local associations among the observed series. It is based on the idea that any observation might be more related with the average of the neighboring values than it is related with the average of the whole series. The compositional covariance is negative when the relationship between the part averages is inverse and it is positive when the relationship is direct. Higher compositional covariance indicates a stronger association. The compositional covariance between scalar matrices A and B is defined by the following equation:

$$\text{Cov}_c(A, B) = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (A_{i,j} - \bar{A}_i)(B_{i,j} - \bar{B}_i)}{n} \quad (2)$$

where  $B_{i,j}$  is the  $j^{\text{th}}$  data in  $i^{\text{th}}$  part of vector B and  $\bar{B}_i$  is the arithmetic mean of the  $i^{\text{th}}$  part of vector B;

Covariance is a scale dependent dimensioned measure and its value increases when a variable is increased in scale. Correlation is a scaled and dimensionless version of covariance and it takes values between  $-1$  and  $1$ . A correlation of  $\pm 1$  indicates perfect linear association and  $0$  indicates no linear relationship. Based on the above definitions of compositional variance and covariance, the compositional correlation is defined as follows:

$$r_c = \frac{\text{Cov}_c(A, B)}{\sqrt{\text{Var}_c(A)\text{Var}_c(B)}} \quad (3)$$

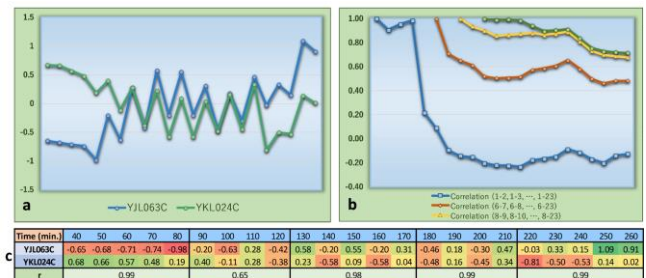
The following equation can also be used for calculating the compositional correlation directly:

$$r_c = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (A_{i,j} - \bar{A}_i) (B_{i,j} - \bar{B}_i)}{\sqrt{\left[ \sum_{i=1}^k \sum_{j=1}^{n_i} (A_{i,j} - \bar{A}_i)^2 \right] \left[ \sum_{i=1}^k \sum_{j=1}^{n_i} (B_{i,j} - \bar{B}_i)^2 \right]}} \quad (4)$$

## 3. RESULTS

### 3.1. Application on gene expression data

The gene expression dataset used in this study consists of the results of the *cdc15* experiment made by Spellmann et al. [30]. The expression data was provided by Reshef et al. [31]. Before explaining the implementation process of the compositional correlation, an example is presented for showing how unreliable the Pearson's correlation ( $r$ ) might be when comparing time series data (Figure 1). The selected gene pair is YJL063C and YKL024C. Both expression series have 23 observations ( $n = 23$ ) and  $r = -0.12$  between the whole time series of the genes. The correlation value is 1 for the first two pair (1-2) and the correlation gradually decreases to  $-0.12$  for the whole data range (observation 1-23) ( $r = 1$  for the range 1-2;  $r = 0.91$  for the range 1-3;  $r = 0.95$  for the range 1-4; . . . and  $r = -0.12$  for the range 1-23 as shown with the blue line in Figure 1b). The correlation between the first five observations is 0.99 and  $r$  suddenly decreases to 0.22 when the first six observations are considered. Value of  $r$  continues to have very low values for the remaining ranges and does not get positive values for the ranges from (1-7) to (1-23).



**Figure 1.** (a) The expression time series of budding yeast genes YJL063C and YKL024C, (b) the variation of Pearson's correlation with the selected data range and (c) the expression series and the correlations between the parts of the BCC of the genes YJL063C and YKL024C.

When the first five observations are ignored (for which  $r = 0.99$ ) the  $r$  values vary between 1 (for the range 6-7) and 0.48 (for the range 6-23) (Figure 1b). The figure clearly shows that correlation only gets negative values when the first five values which nearly have a perfect positive correlation are included in the calculation of correlation. The expression time series of the sample genes always increase and decrease together for all smallest subsections ( $n = 2$ ) (Figure 1a) indicating a very strong quantitative relationship but the Pearson's correlation does not reflect this behavior.

The compositional correlation method calculates correlations for all compositions of the compared series and the values tend to be higher when the parts of the compositions are highly correlated. For example, the above gene pair is one of the numerous gene pairs determined to have a very low value of Pearson's correlation while most of

the compositional correlations are very high (up to 0.92). For this pair, the best correlated composition (BCC) is [5, 4, 5, 4, 5]. When the 23 observations are divided into five parts (which form the BCC) as shown in Figure 1c,  $r = 0.65$  for the second part and  $r \geq 0.98$  for the remaining parts. The high  $r$  values in all parts point out a strong direct relationship between the series as depicted by the time series graph (Figure 1a) while the  $r$  value for the whole series is interestingly a negative number close to zero indicating an inverse weak relationship. The high value of the compositional correlation for the gene pair indicates the apparent direct relationship. The above example shows the influence of sample size on the value of correlation.

The compositional correlation approach enables detection of this type of relationship by considering the cumulative influence of all possible sample compositions for the compared series. If there is no composition producing high correlations (positive or negative) for the compared parts, then the association between the compared series is definitely weak.

### 3.2. Compositional correlations between 4381 gene expression series

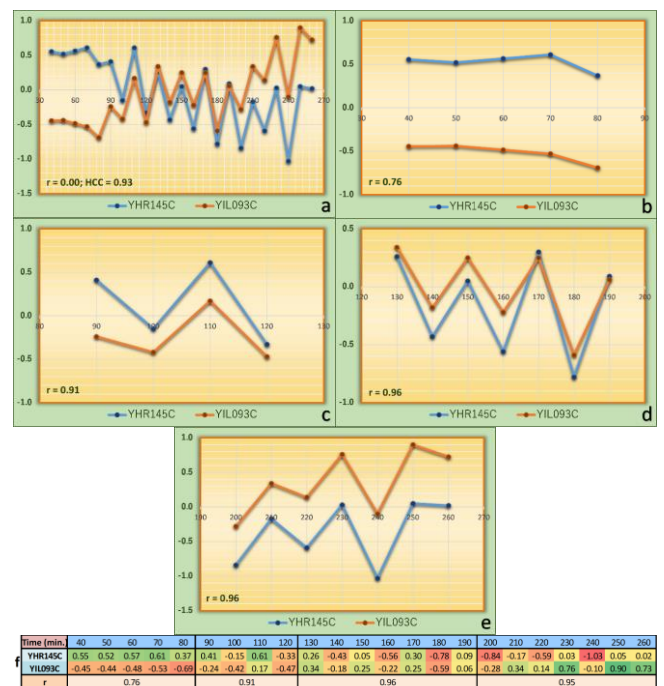
The compositional correlations between the available expression series of all pairs of 4381 budding yeast (*saccharomyces cerevisiae*) genes were calculated. Compositional correlations against time were also calculated for each gene. The expression time series for each gene consists of 23 observations ( $n = 23$ ) and the number of all possible gene pairs is 9,594,390. For each gene pair, all possible compositional correlations were calculated by considering the minimum number of observations in each part of a composition to be at least 4 (a total of 250 compositional correlations for each pair when  $m = 4$ ). The total number of calculated compositional correlations is over 2.3 billion. All of the calculated compositional correlations were not stored as output because this would significantly slow down the file generation process and increase the requirement of storage space. Only the highest compositional correlation (HCC),  $r$ , the lowest compositional correlation (LCC), best correlated composition (BCC) and the worst correlated composition (WCC) values for all possible data series pairs were written to the output file. The output file is provided for download via the following link for enabling further investigation. The file contains clues for determining the relationships and functions of the hundreds of yeast genes which are still unidentified.

<https://www.dropbox.com/s/lqqnmaf9h6g1rr/Compositional.Correlations.Spellman.m4.rar>

Among all the compared gene pairs, the highest compositional correlation (0.993) was obtained between the genes YDL003W and YDR097C for the composition [7, 4, 8, 4] (Table S1 in the Supplementary Material provides the complete list of the gene pairs with HCC values over 0.9). For this gene pair,  $r = 0.985$  and  $LCC = 0.942$  and the small difference between HCC and LCC indicates a very strong relationship all through the observed period (Figure S1 in the

Supplementary Material). LCC is higher than 0.9 for 777 gene pairs while it is over 0.85 for 5202 gene pairs. The lowest LCC value (-0.982) was obtained for the pair YIL141W and YMR031C for the composition [9, 4, 5, 5] (Table S2 provides the complete list of the gene pairs with lowest compositional correlation (LCC) values under -0.9). For this pair,  $r = -0.932$  and  $HCC = -0.791$  (Figure S2). HCC is lower than -0.9 for 146 gene pairs while it is less than -0.85 for 1355 gene pairs. The HCC values are over 0.9 for 31185 (0.325%) gene pairs (Table S1) while the Pearson's correlation is over 0.9 for only 2684 (0.0027%) of the gene pairs. For example,  $HCC = 0.93$  for the composition [5, 4, 7, 7] of the gene pair YHR145C and YIL093C while the composition [23] is the WCC and gives  $LCC = 0.00$  which is the Pearson's correlation (Figure 2).

This gene pair is one of the 58 gene pairs for which the  $HCC > 0.9$  while  $-0.1 < r < 0.1$  (Table S3). Figures 2b - 2e show the expressions of the genes for each part of the BCC and Figure 2f shows the expression values together with the Pearson's correlations for each part of the BCC. Even though the Pearson's correlation for the whole series is zero, the Pearson's correlations for each part of the BCC are very high (between 0.76 and 0.96; 3/4 of them being over 0.9). This shows that the Pearson's correlation for the gene pair is misleading because there seems to be a very strong direct relationship between the genes as shown by the time series graph and the very high compositional correlation value.



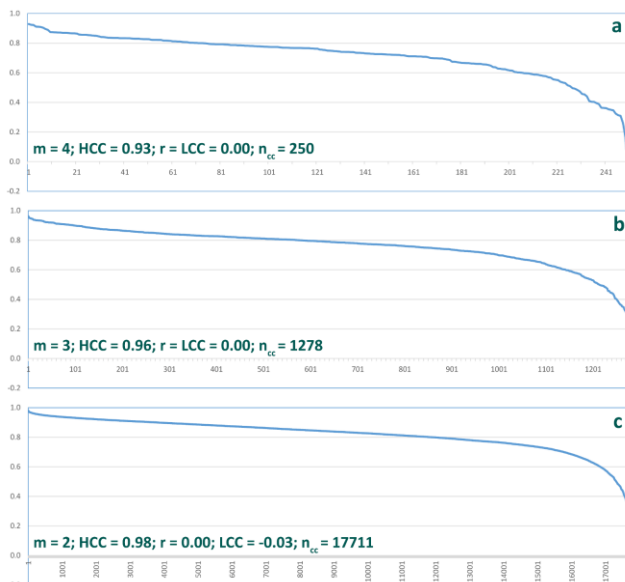
**Figure 2.** (a) Expression time series for the genes YHR145C and YIL093C; (b) first 5 expression pairs (40 to 80 minutes); (c) 4 expression pairs from 90 to 120 minutes; (d) 7 expression pairs from 130 to 190 minutes and (e) 7 expression pairs from 200 to 260 minutes and (f) the expression series and the correlations between the parts of the BCC of the genes YHR145C and YIL093C.

Table S4 presents the 250 compositional correlations between the genes YHR145C and YIL093C for  $m = 4$ . For the same gene pair, the 1278 compositional correlations for  $m = 3$  (Table S5) and 17711 (the maximum number of



possible compositions) compositional correlations for  $m = 2$  (Table S6) are calculated separately for investigating the variation of compositional correlation for the gene pair. The compositions for  $m = 3$  and  $m = 4$  are subsets of the compositions for  $m = 2$  and their compositional correlation values remain within the compositional correlation range obtained for  $m = 2$  (Figures 3a, 3b and 3c). The compositional correlation range is the difference between HCC and LCC and these values are available for each gene pair.

All compositional correlations are higher than  $r$  (which is 0.00) when  $m = 3$  and  $m = 4$ . Similarly, when  $m = 2$ , 17709 of the 17711 compositional correlations (99.99%) are higher than  $r$ . The results for  $m = 2$  and  $m = 3$  show that the obtained results for  $m = 4$  provide sufficient information on the compositional correlation structure of the investigated gene expression series. This indicates that there is a strong direct numerical relationship between the expressions all through the investigated period and might point out the existence of a functional relationship even though the Pearson's correlation is zero.



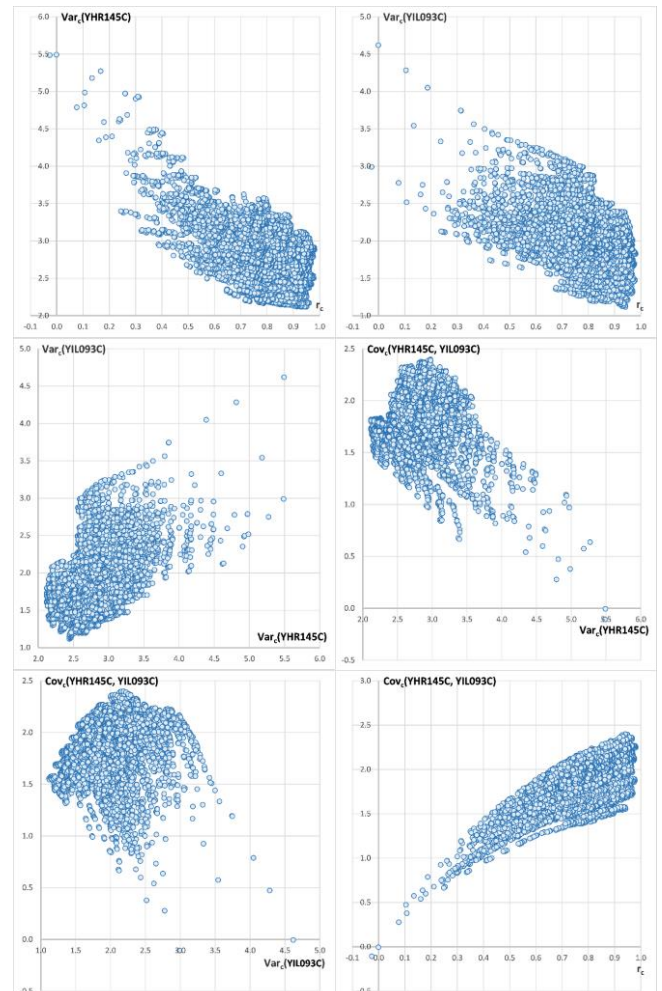
**Figure 3.** The compositional correlations obtained for the gene pair YHR145C and YIL093C when  $m = 4$  (a),  $m = 3$  (b) and  $m = 2$  (c). The HCC,  $r$ , LCC and the number of compositional correlations ( $n_{cc}$ ) are indicated on each figure.

The variance and covariance clouds of the genes YHR145C and YIL093C shown in Figure 4 also validate that the Pearson's correlation is far from representing the association between the genes. The Pearson's correlation is a point at the far end (the point at the origin) of the tail of the compositional covariance cloud shown in the bottom right panel. The Pearson's correlation does not indicate the strong direct (positive) relationship between the genes shown by all the panels in the figure.

### 3.3. Inverse relationships

As in all association studies, determination of inverse relationships between genes might also be as important as determining direct relationships for assessing expression balance of proteins [32]. Pearson's correlation is below -0.9

for 554 of the investigated gene pairs, while 12373 gene pairs have a LCC value below -0.9 indicating that there might be much more inversely related yeast gene pairs than the Pearson's correlation points out (Table S2). The genes YBR146W and YJR045C are one of the many inversely related gene pairs that Pearson's correlation fails to detect. For this pair,  $HCC = r = 0.00$  while  $LCC = -0.93$  for the WCC which is [4, 4, 6, 4, 5]. Figure 5 shows the expression time series for this pair together with the correlations for each part of the WCC.



**Figure 4.** The variance and covariance clouds obtained for the gene pair YHR145C and YIL093C when  $n = 23$  and  $m = 2$ .

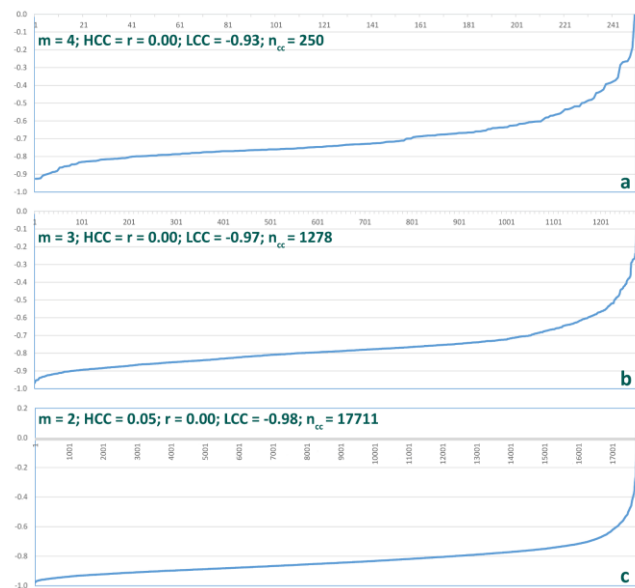
The negative correlations for each part are very close to -1 (ranging between -0.88 and -1.00) indicating a strong inverse relationship but the combined parts produce a zero Pearson's correlation falsely proposing that the genes have no relation. The inverse relationship is also apparent in the time series graph but it is practically impossible to generate graphs and determine these types of relationships manually when there are millions of pairs of genes. The computational procedure of the CompCorr software enables determination of these relationships by calculating compositional correlations for all possible compositions.

Tables S7, S8 and S9 respectively show in ascending order, the compositional correlations obtained for the genes YBR146W and YJR045C by taking  $m = 4$ ,  $m = 3$  and  $m = 2$ .

When  $m = 3$  and  $m = 4$ , all compositional correlations are lower than  $r$  (which is 0.00) and when  $m = 2$ , 17709 of the 17711 compositional correlations (99.99%) are lower than  $r$  (Figure 6a, 6b and 6c). These results imply that these genes might have a strong inverse functional relationship even though Pearson's correlation is zero.

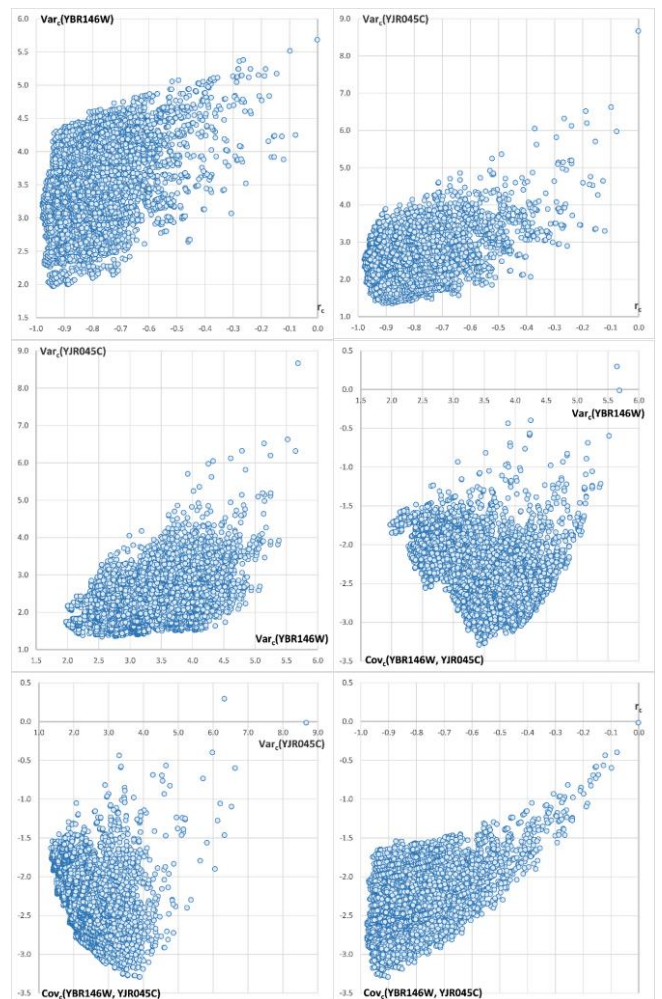


**Figure 5.** (a) Expression time series for the genes YBR146W and YJR045C; (b) first 4 expression pairs (40 to 70 minutes); (c) 4 expression pairs from 80 to 110 minutes; (d) 6 expression pairs from 120 to 170 minutes; (e) 4 expression pairs from 180 to 210 minutes and (f) 5 expression pairs from 220 to 260 minutes and (g) the expression series and the correlations between the parts of the WCC of the genes YBR146W and YJR045C.



**Figure 6.** The compositional correlations obtained for the gene pair YBR146W and YJR045C when  $m = 4$  (a),  $m = 3$  (b) and  $m = 2$  (c). The HCC,  $r$ , LCC and the number of compositional correlations ( $n_{cc}$ ) are indicated on each figure.

The variance and covariance clouds of the genes YBR146W and YJR045C in Figure 7 show that the Pearson's correlation does not correctly point out the association between the genes. The Pearson's correlation is a point at the far end (the point at the origin) of the tail of the compositional covariance cloud shown in the bottom right panel and it is far from representing the strong direct (positive) relationship between the genes shown by all the panels in the figure. The strong indirect relationship between the gene expression series through the whole observation period is also validated by the covariance clouds in Figure 7.



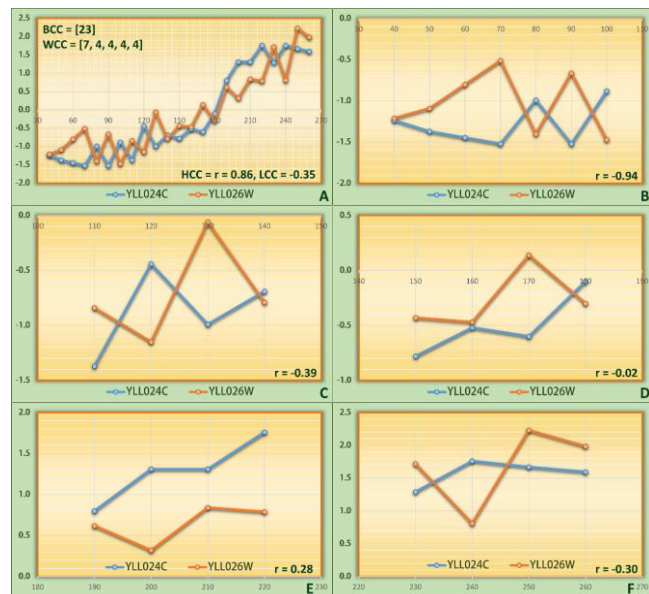
**Figure 7.** The variance and covariance clouds obtained for the genes YBR146W and YJR045C when  $n = 23$  and  $m = 2$ .

The above gene pairs are only two examples among the thousands of pairs with probable functional relationships determined by the compositional correlation method. Some other selected examples with high compositional correlation but significantly lower  $r$  values are presented in Figure S3. Each graph in the figure includes HCC,  $r$  and LCC values together with the BCC's and WCC's. Table S10 shows the expression values of the gene pairs and the correlations for all parts of the BCC's for each pair in Figure S3.

### 3.4. Directly or Inversely Related?

Another feature of compositional correlation is its ability to determine the existence of inverse relationships when the series have a general direct relationship (or vice versa). An

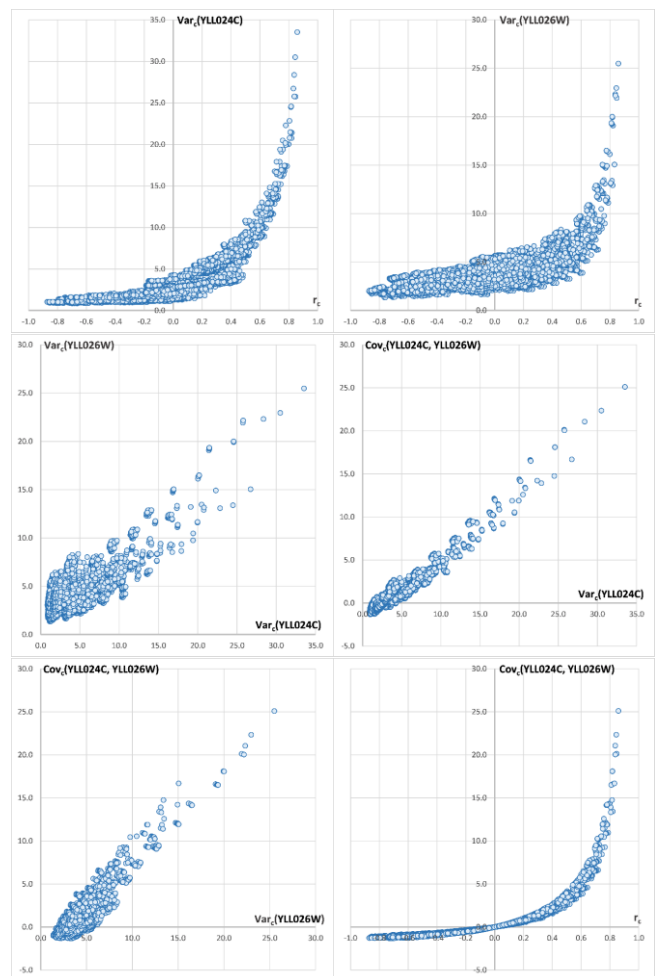
example of this feature is the gene pair YLL024C and YLL026W. Figure 8a shows that the genes in this pair both have a similarly increasing expression trend and the value of Pearson's correlation (0.86) also validates this behavior, but the expression values are always inversely related between all observation points (when one is increasing, the other is decreasing) only except for the minutes between 180 and 190 where they both increase and between the minutes 250 and 260 where they both decrease. This inverse expression behavior is determined by looking at the differences between the HCC and LCC values. For this pair, the LCC value is -0.35 indicating that there is an inverse relationship for some compositions of the gene expression series.



**Figure 8.** The genes YLL024C and YLL026W have a similar expression trend with  $r=0.86$  (a) but their expression patterns have inverse relationship (b-f).

Calculation of correlations for the parts of the WCC ([7, 4, 4, 4, 4]) of the pair also points out the inverse relationship between the genes (Figures 8b-8f). The  $r$  values of the parts of the WCC varies between -0.94 and 0.28 (4/5 being negative). In fact, 10585 of the 17711 compositions (59.8%) have a negative compositional correlation when  $m = 2$  (Figure 9). The dispersion of the variance and the covariance clouds are very small for these genes. This shows that both genes are closely related. The variation of the compositional correlation between 0.859 and -0.871 indicates that the genes have both a positive and a negative relationship. The compositions containing long parts produce positive compositional correlations as expected (the genes have an overall similar pattern) and the compositions composed of shorter parts produce negative compositional correlations because the smaller parts are inversely related.

Some other selected examples of gene pairs with apparent inverse relationships which cannot be detected by calculating  $r$  are shown in Figure S4. The expression values of the sample pairs and the  $r$  values for each part of the BCC for each pair are presented in Table S11. The  $r$  values of the parts (the majority being between -0.95 and -1.00) validate that the pairs might have a strong functional inverse relationship even though the  $r$  values obtained for the whole series vary between -0.01 and -0.53.



**Figure 9.** The variance and covariance clouds obtained for the genes YLL024C and YLL026W when  $n = 23$  and  $m = 2$ .

#### 4. COMPARISONS WITH OTHER CORRELATION METHODS

Determination of genes with similar or concordant (and also inversely related) expression profiles is crucial in the determination of the functions of the genes. If a method fails to find some of the harmonious gene pairs, then it will be much more difficult to make decisions on the functions of the genes as in the case of *saccharomyces cerevisiae* for which there are still hundreds of unidentified genes even though its genome was sequenced nearly 25 years ago. In this section, the performance of the compositional correlation method is compared with four widely used correlation methods which are the Pearson's correlation, Spearman's correlation, distance correlation and SGMIC which was proposed as an algorithm for the precise calculation of the maximal information coefficient [33].

Genes generally exhibit varying expression behaviors through time. Detecting this behavior by using standard correlation approaches is generally not possible because the series are considered as a whole and the different behaviors in subsections are not detected and considered. For example, the expressions of the gene pair YMR296C and YOL032W (Figure S5ac) first slightly decrease together between minutes 40 and 70. Then they begin to fluctuate together with a very similar pattern until minute 170.



Finally, after minute 180, YOL032W shows an increasing and YMR296C shows a decreasing trend while they still have a very similar expression profile. Even though the time series graph for this pair shows that there might be a strong relationship between these two genes, this behavior was not detected by Pearson's (0.048), Spearman's (0.082) and distance correlations (0.386) which pointed out a weak or no relation. The SGMIC value for this gene pair is 0.566 which seems to be a better estimate but might easily be ignored among the millions of SGMIC values for all the possible gene pairs as it also does not point out a significant relationship.

The compositional correlation method successfully determined the relationship between these two genes with an HCC value of 0.943. For another 58 gene pairs, the HCC values are higher than 0.9 for which the Pearson's correlation remains between -0.1 and 0.1 (Figure S5). The HCC, Pearson's correlation, LCC, Spearman's correlation, distance correlation and SGMIC values for these 58 gene pairs are presented in Table 2. The minimum and maximum values for each statistic are indicated in bold.

The obtained compositional correlations are much higher than the correlations of the compared methods only except for the SGMIC value of the pair YBR082C and YOR262W (Figure S5av). The distance correlation value (0.505) for this pair is the second highest distance correlation among the 58 pairs. It is known that the distance correlation is zero if and only if the random variables are statistically independent but it cannot be claimed that the distance correlation always exactly determines the strength of statistical dependence. The distance correlations for none of the compared pairs are close to zero showing that all of the 58 gene pairs are statistically dependent (Figure S5).

The Pearson's and Spearman's correlations fail to detect the statistical dependence between the genes and they produce values close to zero for all the compared pairs in Table 2. This result is caused by the fact that both methods generally fail to detect the relationship when the trend line for one series is increasing while the trend line for the other series is decreasing with nearly the same angle even though the series have a strong relationship. The presented gene pairs are good examples of this deficiency of Pearson's and Spearman's correlation approaches.

The second-best performance after the compositional correlation is shown by the SGMIC method with an average SGMIC value of 0.511 but all SGMIC values except for 0.932 obtained for the pair YBR082C and YOR262W are under 0.8 showing that the SGMIC results for these gene pairs are not sufficiently maximal to be noticed.

The presented comparative results and the supporting figures provide satisfactory proof for the statistical dependence between the compared gene pairs. There are many more strongly related gene pairs detected by the compositional correlation method with very low Pearson's correlation values close to zero. For example, the number of gene pairs with a compositional correlation value over 0.8 but Pearson's correlation between -0.2 and 0.2 is 5999. Consequently, the

results of this study provide the yeast researchers a very narrowed down target for defining the functions of the genes, a task which seems to be impossible by using conventional correlation measures that fail to detect real relationships between genes showing alternating but dependent behavior through the course of time.

## 5. COMPARISON OF THE RESULTS WITH EXISTING LITERATURE

A genome-wide association analysis on *Saccharomyces Cerevisiae* is required for both identifying new genes and exploring the extent to which genetic background influences mechanism, because the majority of functional studies on *Saccharomyces Cerevisiae* are carried out in a small number of laboratory strains that do not represent the rich diversity found in this species [34]. Global gene expression of *Saccharomyces cerevisiae* is also investigated in order to identify the correlation between redox potential profiles and gene expression patterns and enables locating genes that could be modulated by altering culture redox potential during VHG ethanol fermentation [35]. Another potential use of the whole-genome sequences is mapping the genetic basis of phenotypic variation through genome-wide association (GWA) studies, with the benefit that associated variants can be studied experimentally with greater ease [36]. Genome-wide comparative analysis are primarily based on genomic sequence information although differences among organisms are often attributed to differential gene expression [37].

Genes whose expression varies differentially and periodically over the cell cycle might be identified by both experimental and computational methods. Aside from the aforementioned uses of genome-wide gene expression analysis, principal-oscillation-pattern (POP) analysis which is a multivariate and systematic technique for identifying the dynamic characteristics of a system from time-series data, can be used to infer oscillation patterns in gene expression [38]. The gene YDL003W is reported by many researches as one of the cell-cycle genes in *Saccharomyces Cerevisiae*. The results obtained in this paper indicate that there are 31 genes determined by the compositional correlation method to have HCC values with the gene YDL003W which are higher than 0.9. These genes are listed in Table 3 together with the HCC,  $r$  and LCC values. The obtained results show that, all genes listed in Table 3 are also reported as cell-cycle genes by several studies. A detailed summary on these methods were provided by de Lichtenberg et al. [39] and the whole lists of the cell-cycle genes reported in these studies is provided by Wang et al. [38]. The findings of the compositional correlation method as presented in this paper show that the compositional correlation method both compares well with existing computational methods and experiments, and it also determines complementary knowledge in addition to information provided by other approaches because it also detected cell-cycle genes not determined by all compared methods. This comparison proves that the compositional correlation method can be used reliably not only in the determination of cell-cycle genes but also other genome wide association studies, but still, the users must be warned against type I errors which should be



checked as always before making final decisions on their association studies.

**Table 2** Comparisons of correlation methods for gene pairs with compositional correlations over 0.9 while  $-0.1 < r < 0.1$

Gene 1	Gene 2	HCC	Pearson	LCC	Spearman	Dist.Cor.	SGMIC
YCL063W	YPL203W	0.926	<b>0.098</b>	<b>0.098</b>	-0.031	0.357	0.546
YGR244C	YKR072C	0.908	0.095	0.095	-0.038	0.363	0.528
YLR109W	YLR441C	0.913	0.094	0.094	-0.028	0.274	0.445
YDR375C	YDR449C	0.906	0.093	0.093	0.026	0.326	0.548
YLL034C	YPL118W	0.908	0.091	0.091	0.029	0.297	0.652
YMR093W	YNL252C	0.905	0.089	0.089	0.126	0.320	0.441
YPR060C	YPR158W	0.904	0.084	0.084	0.124	0.338	0.510
YKL150W	YNL007C	0.941	0.080	-0.028	0.138	0.352	0.667
YKL122C	YLR109W	0.913	0.079	0.079	-0.052	0.301	0.520
YJR054W	YOL052C	0.903	0.079	0.077	0.059	0.295	0.398
YGR244C	YLR075W	0.913	0.078	0.078	0.066	0.357	0.586
YIL093C	YLR197W	0.914	0.078	0.078	0.001	0.419	0.791
YFR050C	YGL189C	0.919	0.078	0.078	0.053	0.350	0.544
YNL005C	YOR095C	0.903	0.076	0.076	0.091	0.302	0.423
YJL063C	YOL097C	0.910	0.073	0.073	0.006	0.353	0.423
YIL070C	YLR344W	0.907	0.072	0.072	-0.036	0.291	0.545
YLR203C	YPL048W	0.916	0.069	0.069	0.072	0.380	0.464
YGL049C	YMR186W	0.905	0.068	0.068	0.048	0.252	0.361
YGL120C	YJL063C	0.901	0.068	0.068	0.015	0.400	0.559
YJL125C	YNL252C	0.936	0.064	0.064	0.019	0.329	0.360
YDR489W	YPR158W	0.903	0.063	0.063	0.065	0.270	0.418
YLR354C	YNL007C	0.912	0.063	0.033	0.133	0.317	0.586
YEL039C	YER027C	0.905	0.058	-0.112	0.192	0.289	0.455
YGR244C	YOR300W	0.902	0.051	0.051	0.062	0.309	0.456
YGR244C	YHR145C	0.936	0.050	0.050	-0.027	0.447	0.735
YNL135C	YOR325W	0.921	0.050	0.050	0.076	0.308	0.490
YER156C	YLR109W	0.920	0.050	0.050	-0.056	<b>0.208</b>	0.316
YBL066C	YDR231C	0.902	0.048	0.048	0.198	0.409	0.464
YMR296C	YOL032W	<b>0.943</b>	0.048	0.048	0.082	0.386	0.566
YCR056W	YOL032W	<b>0.900</b>	0.044	0.044	0.088	0.330	0.453
YLR203C	YPR085C	0.922	0.040	0.040	0.012	0.367	0.321
YIL093C	YLR175W	0.908	0.035	0.035	-0.055	0.386	0.697
YGL219C	YNL007C	0.903	0.031	<b>-0.179</b>	0.132	0.403	0.436
YBL101W-B	YLR069C	0.923	0.028	0.028	0.029	0.366	0.482
YLL026W	YML008C	0.909	0.024	0.024	-0.005	0.288	0.510
YJL063C	YPR062W	0.902	0.023	0.023	-0.146	0.393	0.761
YDL022W	YJL029C	0.919	0.016	0.016	-0.047	0.353	0.667
YDR231C	YLR185W	0.924	0.016	0.016	<b>0.200</b>	0.386	0.588
YCR056W	YPL118W	0.913	0.012	0.012	-0.008	0.290	0.351
YJL063C	YOR300W	0.940	0.007	0.007	-0.052	0.341	0.493
YHR145C	YIL093C	0.929	-0.001	-0.001	-0.135	0.407	0.588
YLL034C	YLR203C	<b>0.900</b>	-0.008	-0.008	-0.029	0.380	0.592
YJL063C	YMR102C	0.912	-0.022	-0.022	-0.037	0.341	0.367
YBR183W	YHR216W	0.911	-0.030	-0.030	0.064	<b>0.508</b>	0.775
YHL035C	YNL007C	0.905	-0.036	-0.177	0.177	0.367	0.586
YGL221C	YNL007C	0.921	-0.040	-0.123	0.113	0.332	0.618
YER049W	YGR048W	0.901	-0.041	-0.041	0.003	0.309	0.463
YBR082C	YOR262W	0.917	-0.044	-0.105	0.044	0.505	<b>0.932</b>
YLR109W	YPR110C	0.925	-0.047	-0.047	-0.115	0.241	0.348
YCL014W	YDL110C	0.901	-0.060	-0.090	<b>-0.204</b>	0.398	0.649
YGR097W	YLL026W	0.904	-0.062	-0.062	-0.002	0.274	0.332
YLR109W	YOR300W	0.911	-0.063	-0.063	-0.087	0.256	0.426
YGR228W	YOR310C	0.904	-0.073	-0.073	-0.110	0.251	<b>0.259</b>
YLR138W	YPL118W	0.917	-0.076	-0.076	-0.116	0.332	0.324
YGR244C	YLR293C	0.924	-0.081	-0.081	-0.144	0.350	0.559
YDR509W	YGR254W	0.908	-0.082	-0.082	-0.187	0.371	0.499
YKL024C	YLR109W	0.907	-0.089	-0.089	-0.176	0.301	0.495
YPL118W	YPR048W	0.906	<b>-0.094</b>	-0.094	-0.127	0.293	0.288
	<b>Max:</b>	0.943	0.098	0.098	0.200	0.508	0.932
	<b>Min:</b>	0.900	-0.094	-0.179	-0.204	0.208	0.259
	<b>Average:</b>	0.913	0.024	0.010	0.009	0.340	0.511

**Table 3** The genes determined to have HCC values with the gene YDL003W higher than 0.9

GENE	HCC	r	LCC	BCC	WCC
YDR097C	0.9928	0.9851	0.9422	7, 4, 8, 4	4, 5, 5, 5, 4
YJL115W	0.9723	0.9454	0.8718	6, 4, 9, 4	4, 5, 5, 5, 4
YGR044C	0.9692	0.8275	0.7788	7, 5, 7, 4	4, 5, 5, 5, 4
YOL017W	0.9618	0.9565	0.8425	7, 6, 5, 5	4, 5, 6, 4, 4
YGR152C	0.9554	0.9220	0.8242	10, 9, 4	4, 5, 5, 5, 4
YKL045W	0.9514	0.9266	0.7459	6, 6, 6, 5	4, 5, 5, 5, 4
YDL101C	0.9453	0.9246	0.7959	7, 4, 8, 4	4, 5, 5, 9
YOL007C	0.9440	0.9234	0.6680	7, 4, 8, 4	4, 5, 5, 5, 4
YLL002W	0.9423	0.9337	0.6919	8, 4, 6, 5	4, 5, 5, 5, 4
YHR110W	0.9416	0.9296	0.7608	7, 4, 8, 4	4, 5, 5, 5, 4
YJR148W	0.9388	0.8685	0.5928	8, 5, 5, 5	4, 5, 5, 5, 4
YDL211C	0.9387	0.7795	0.5522	4, 4, 11, 4	4, 5, 5, 9
YIL066C	0.9374	0.8455	0.6212	7, 5, 7, 4	4, 5, 5, 9
YIL076W	0.9361	0.7787	0.5249	8, 5, 4, 6	4, 5, 14
YLR049C	0.9358	0.7527	0.5078	7, 4, 8, 4	4, 5, 5, 5, 4
YLR194C	0.9343	0.8638	0.7242	10, 9, 4	4, 4, 6, 5, 4
YDL103C	0.9338	0.8097	0.7082	6, 6, 6, 5	4, 5, 5, 5, 4
YJL074C	0.9321	0.8373	0.6543	5, 8, 6, 4	4, 5, 6, 4, 4
YPL256C	0.9321	0.9066	0.6296	6, 4, 7, 6	4, 5, 5, 9
YLL022C	0.9310	0.8986	0.6626	8, 5, 5, 5	4, 5, 5, 5, 4
YOR114W	0.9305	0.8335	0.5588	8, 5, 6, 4	4, 5, 5, 9
YGR151C	0.9291	0.8723	0.6938	8, 10, 5	4, 5, 5, 5, 4
YLR121C	0.9275	0.8503	0.4697	8, 11, 4	4, 5, 5, 9
YML027W	0.9264	0.8891	0.5626	11, 8, 4	4, 5, 5, 9
YDL127W	0.9248	0.7899	0.6389	4, 7, 7, 5	5, 4, 5, 5, 4
YLR183C	0.9214	0.9156	0.6761	13, 5, 5	4, 5, 5, 5, 4
YOL090W	0.9202	0.9060	0.6185	8, 11, 4	4, 5, 5, 5, 4
YPR175W	0.9188	0.8733	0.7006	5, 5, 4, 4, 5	4, 5, 6, 4, 4
YGR189C	0.9063	0.8866	0.5842	10, 9, 4	4, 5, 5, 5, 4
YLR286C	0.9049	0.7052	0.6415	4, 10, 9	4, 5, 5, 5, 4
YMR029C	0.9042	0.5445	0.4129	6, 4, 4, 4, 5	4, 5, 14

## 6. THE COMPCORR SOFTWARE

The CompCorr software developed in Python for implementing the compositional correlation method is freely provided together with this manuscript. The software accepts an Excel file containing the data series as input and generates a text file as output containing the compositional correlations. The number of compositional correlations calculated for each pair varies according to the length of the data series and the minimum number of accepted values in each part of the compositions. The compositions were determined by using the ruleGen function which generates all interpart restricted compositions of  $n$  by using restriction function  $\sigma$  [40]. For each composition, the compositional correlation is determined by using Equation 4.

## 7. CONCLUSION

The results obtained in this study have shown that the compositional correlation method is very successful in determining linear, nonlinear, direct and indirect relationships between gene expression series and that the method has a great potential of being applied in all areas of science. Comparisons with widely used and well-established correlation methods also validated the results of the study. Taken together, the presented findings could be applied quite reliably in studies aimed at determining the functions of

specific yeast genes for which the functions are still undefined. However, the current results were obtained by using available expressions of 4381 of the yeast genes which are estimated to be around 6000. Therefore, future research might include the remaining genes for finding more relationships.

In the light of the findings on the gene expression data series, it is evident that the method may enable possibilities for numerous important discoveries and will contribute to the improvement of our understanding of correlation as a new way of finding associations. The usefulness and benefit of the compositional correlation method lies in the approach that the variation of average through the observations is considered instead of considering the average of the whole series. The author also hopes that the method and the results presented in this manuscript will also provide important clues and the tools to the biologists trying to find the functions for the genes of many other organisms. The software code developed for implementing the compositional correlation method is freely provided as a supplement together with the manuscript.

## 8. SUPPLEMENTARY MATERIAL

The online Supplementary Material contains all the Figures S1 to S5, the Tables S1 to S12 and the Python code of the CompCorr software. The software is provided under the terms of the GNU Free Documentation License, Version 1.3. The Supplementary Material is available for download in the following link:

<https://www.dropbox.com/s/ai8r590sz2e8aw6/Supplementary.Material.pdf?dl=0>

**Author contributions:** Concept – F.D.; Data Collection &/or Processing – F.D.; Literature Search – F.D.; Writing – F.D.

**Conflict of Interest:** No conflict of interest was declared by the author.

**Financial Disclosure:** The author declared that this study has received no financial support.

## REFERENCES

- [1] H. P. Lovecraft. (1928, February) The Call of Cthulhu. *Weird Tales*. 159-178.
- [2] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents," *Proceedings of the Royal Society of London*, vol. 58, no. 347-352, pp. 240-242, January 1, 1895 1895, doi: 10.1098/rspl.1895.0041.
- [3] J.-L. Magnard et al., "Biosynthesis of monoterpene scent compounds in roses," *Science*, vol. 349, no. 6243, pp. 81-83, 2015, doi: 10.1126/science.aab0696.
- [4] Y. X. R. Wang, K. Jiang, L. J. Feldman, P. J. Bickel, and H. Huang, "Inferring gene-gene interactions and functional modules using sparse canonical correlation analysis," (in en), *Ann. Appl. Stat.*, vol. 9, no. 1, pp. 300-323, 2015/03 2015, doi: 10.1214/14-AOAS792.
- [5] J. M. Bland and D. G. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet*, vol. 1, no. 8476, pp. 307-310, 1986. [Online]. Available:

- <http://www.scopus.com/inward/record.url?eid=2-s2.0-0022624332&partnerID=40&md5=7814d6e99afa1a58edebf08387536f8c>.
- [6] M. B. I. Lobbes and P. J. Nelemans, "Good correlation does not automatically imply good agreement: The trouble with comparing tumour size by breast MRI versus histopathology," *European Journal of Radiology*, vol. 82, no. 12, pp. e906-e907, 2013, doi: 10.1016/j.ejrad.2013.08.025.
- [7] M. T. Brett, "When is a correlation between non-independent variables "spurious"?", *Oikos*, vol. 105, no. 3, pp. 647-656, 2004, doi: 10.1111/j.0030-1299.2004.12777.x.
- [8] L. Duan, W. N. Street, Y. Liu, S. Xu, and B. Wu, "Selecting the Right Correlation Measure for Binary Data," *ACM Trans. Knowl. Discov. Data*, vol. 9, no. 2, p. Article 13, 2014, doi: 10.1145/2637484.
- [9] N. Coffey and J. Hinde, "Analyzing time-course microarray data using functional data analysis - A review," *Statistical Applications in Genetics and Molecular Biology*, Review vol. 10, no. 1, 2011, Art no. 23, doi: 10.2202/1544-6115.1671.
- [10] J. Zhang et al., "Genome-wide association study for flowering time, maturity dates and plant height in early maturing soybean (*Glycine max*) germplasm," *BMC Genomics*, vol. 16, no. 1, p. 217, 2015/03/20 2015, doi: 10.1186/s12864-015-1441-4.
- [11] X. Zhang, F. Zou, and W. Wang, "Efficient algorithms for genome-wide association study," *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 4, p. Article 19, 2009, doi: 10.1145/1631162.1631167.
- [12] S. Kumari et al., "Evaluation of Gene Association Methods for Coexpression Network Construction and Biological Knowledge Discovery," *PLoS One*, vol. 7, no. 11, p. e50411, 2012, doi: 10.1371/journal.pone.0050411.
- [13] F. Dikbaş, "A novel two-dimensional correlation coefficient for assessing associations in time series data," *International Journal of Climatology*, vol. 37, no. 11, pp. 4065-4076, 2017, doi: <https://doi.org/10.1002/joc.4998>.
- [14] F. Dikbaş, "A New Two-Dimensional Rank Correlation Coefficient," *Water Resources Management*, vol. 32, no. 5, pp. 1539-1553, 2018/03/01 2018, doi: 10.1007/s11269-017-1886-0.
- [15] S.-J. Chou et al., "Analysis of spatial-temporal gene expression patterns reveals dynamics and regionalization in developing mouse brain," *Sci. Rep.*, vol. 6, no. 1, p. 19274, 2016/01/20 2016, doi: 10.1038/srep19274.
- [16] E. Martinez, K. Yoshihara, H. Kim, G. M. Mills, V. Trevino, and R. G. W. Verhaak, "Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects," *Oncogene*, Original Article vol. 34, no. 21, pp. 2732-2740, 05/21/print 2015, doi: 10.1038/onc.2014.216.
- [17] J. A. Bubier et al., "Integration of heterogeneous functional genomics data in gerontology research to find genes and pathway underlying aging across species," *PLoS One*, vol. 14, no. 4, p. e0214523, 2019, doi: 10.1371/journal.pone.0214523.
- [18] D. I. Scheffer, J. Shen, D. P. Corey, and Z. Y. Chen, "Gene expression by mouse inner ear hair cells during development," *Journal of Neuroscience*, vol. 35, no. 16, pp. 6366-6380, 2015, doi: 10.1523/JNEUROSCI.5126-14.2015.
- [19] J. Delfini et al., "Population structure, genetic diversity and genomic selection signatures among a Brazilian common bean germplasm," *Sci. Rep.*, vol. 11, no. 1, p. 2964, 2021/02/03 2021, doi: 10.1038/s41598-021-82437-4.
- [20] A. R. Marderstein, E. R. Davenport, S. Kulm, C. V. Van Hout, O. Elemento, and A. G. Clark, "Leveraging phenotypic variability to identify genetic interactions in human phenotypes," *The American Journal of Human Genetics*, vol. 108, no. 1, pp. 49-67, 2021/01/07/ 2021, doi: <https://doi.org/10.1016/j.ajhg.2020.11.016>.
- [21] M. Perros, "A sustainable model for antibiotics," *Science*, vol. 347, no. 6226, pp. 1062-1064, 2015, doi: 10.1126/science.aaa3048.
- [22] F. Dikbaş, "Compositional Correlation for Detecting Real Associations Among Time Series," in *Academic Researches in Mathematic and Sciences*, Z. Yildirim Ed., 1 ed. Ankara: Gece Kitaplığı, 2018, pp. 27-46.
- [23] S. Heubach and T. Mansour, "Compositions of n with parts in a set," *Congressus Numerantium*, vol. 168, p. 127, 2004.
- [24] G. E. Andrews, *The Theory of Partitions (Encyclopedia of Mathematics and its Applications)*. Cambridge: Cambridge University Press, 1984.
- [25] G. E. Andrews and K. Eriksson, *Integer Partitions*. Cambridge: Cambridge University Press, 2004.
- [26] G. H. Hardy and E. M. Wright, *An introduction to the theory of numbers*. Oxford university press, 1979.
- [27] J. J. Watkins, *Number theory: a historical approach*. Princeton University Press, 2013.
- [28] A. P. Stakhov, "The golden section in the measurement theory," *Computers and Mathematics with Applications*, vol. 17, no. 4-6, pp. 613-638, 1989, doi: 10.1016/0898-1221(89)90252-6.
- [29] L. Lindroos, "Integer Compositions, Gray Code, and the Fibonacci Sequence," 2012.
- [30] P. T. Spellman et al., "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization," *Molecular Biology of the Cell*, vol. 9, no. 12, pp. 3273-3297, 1998. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0031742022&partnerID=40&md5=212944b877cb8836ca1f33a585f0b8c9>.
- [31] D. N. Reshef et al., "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, pp. 1518-1524, 2011, doi: 10.1126/science.1205438.
- [32] V. Subbarayan et al., "Inverse relationship between 15-lipoxygenase-2 and PPAR- $\gamma$  gene expression in normal epithelia compared with tumor epithelia," *Neoplasia*, vol. 7, no. 3, pp. 280-293, 2005, doi: 10.1593/neo.04457.
- [33] Y. Zhang, S. Jia, H. Huang, J. Qiu, and C. Zhou, "A novel algorithm for the precise calculation of the maximal information coefficient," *Sci. Rep.*, Article vol. 4, 2014, Art no. 6662, doi: 10.1038/srep06662.
- [34] M. Sardi et al., "Genome-wide association across *Saccharomyces cerevisiae* strains reveals substantial

- variation in underlying gene requirements for toxin tolerance," *PLoS Genet.*, vol. 14, no. 2, p. e1007217, 2018, doi: 10.1371/journal.pgen.1007217.
- [35] C. G. Liu, Y. H. Lin, and F. W. Bai, "Global gene expression analysis of *Saccharomyces cerevisiae* grown under redox potential-controlled very-high-gravity conditions," (in eng), *Biotechnol J*, vol. 8, no. 11, pp. 1332-40, Nov 2013, doi: 10.1002/biot.201300127.
- [36] C. F. Connelly and J. M. Akey, "On the prospects of whole-genome association mapping in *Saccharomyces cerevisiae*," (in eng), *Genetics*, vol. 191, no. 4, pp. 1345-1353, 2012, doi: 10.1534/genetics.112.141168.
- [37] S. Bergmann, J. Ihmels, and N. Barkai, "Similarities and Differences in Genome-Wide Expression Data of Six Organisms," *PLoS Biol.*, vol. 2, no. 1, p. e9, 2003, doi: 10.1371/journal.pbio.0020009.
- [38] D. Wang, A. Arapostathis, C. O. Wilke, and M. K. Markey, "Principal-Oscillation-Pattern Analysis of Gene Expression," *PLoS One*, vol. 7, no. 1, p. e28805, 2012, doi: 10.1371/journal.pone.0028805.
- [39] U. de Lichtenberg, L. J. Jensen, A. Fausbøll, T. S. Jensen, P. Bork, and S. Brunak, "Comparison of computational methods for the identification of cell cycle-regulated genes," (in eng), *Bioinformatics*, vol. 21, no. 7, pp. 1164-71, Apr 1 2005, doi: 10.1093/bioinformatics/bti093.
- [40] J. Kelleher, *Encoding Partitions as Ascending Compositions*. NUI, 2005 at Department of Computer Science, UCC., 2005.