Bioinformatics Study of Functional Associations Observed in Multiple Sources of Human Genome Data

Seungwoo Hwang[1,2] and Igor B. Kuznetsov[*,1]

[1]*Gen*NY*sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, University at Albany, SUNY, One Discovery Drive, Rensselaer, NY 12144, USA*

[2]*Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea*

Abstract: High-throughput genome analysis techniques produce the ever increasing number of heterogeneous large-scale datasets. Studies of these mutually complementary sources of data promise insights into a global picture of the living cell. Here, we present a simple bioinformatics methodology for the analysis of multiple heterogeneous sources of 'omic' (genomic, proteomic, etc) data. We apply this methodology to study associations among four types of human 'omic' data: protein-protein interactions, gene expression, transcription factor binding sites, and functional pathways. The results of our study indicate that the proposed approach can be used to identify and rank statistically significant functional associations among genes. We show that combinations of multiple data types provide additional insights into the properties of functional pathways. The proposed methodology can also be used as a quantitative procedure for evaluating the quality of 'omic' datasets.

INTRODUCTION

Recent technological advances in high-throughput data acquisition provide us with various types of large-scale datasets, such as whole genome sequences, gene expression, protein-protein interactions, functional pathways, location of transcription factor binding sites, etc. This gives us a unique opportunity to use bioinformatics to integrate diverse and mutually complementary sources of 'omic' data (*e.g.*, genomic, proteomic, etc.) into a single coherent systems biology framework in order to provide functional inference, reveal essential features of gene and protein interaction networks, and ultimately to model these networks. The results of such integrative studies have several key advantages (reviewed in [1-3]). In particular, using multiple sources of information may allow us to reduce systematic noise inherently present in all types of experimental data. Integrative approaches are also important for the studies of complex diseases, such as cancer [4], since predicting the status of disease cases based on multiple biomarkers represents a starting point towards translating genomics research into clinical medicine. The integrative approach can also be used for predicting properties of one type of data based on other types of 'omic' (genomic, proteomic, etc) data [2, 5-7], for evaluating 'omic' datasets [8], and for functional prediction and inference [9-11]. Such a promise of the integrative approach is based on the general assumption that, within a given genome, there exist inter-relationships between heterogeneous types of genomic data [12]. Since even seemingly different data types describe various functional aspects of the same genome (*e.g.*, the human genome), it seems reasonable to anticipate the existence of non-random associations among them. However, the existence of such associations needs to be verified and their strength needs to be quantified for each particular combination of data types [3]. In this manuscript, we use the term 'association' instead of 'correlation' in order not to confuse it with correlation between expression profiles.

A number of studies have demonstrated the existence of non-random pairwise associations between different types of large-scale 'omic' datasets. 'Non-random association' or just 'association' in this context means that genes that are functionally related with respect to one data type also tend to be related with respect to another data type. For the first time, such an association was demonstrated on the example of the yeast interactome and transcriptome. Since interacting proteins must be present within the cell at the same time, genes that encode them should also be expressed during the same time intervals. Consistent with this reasoning, it was shown that yeast genes with similar expression profiles are more likely to encode interacting proteins than randomly chosen genes [13]. A related study of the yeast genome showed that genes encoding interacting proteins exhibit higher than average co-expression [12]. This study also showed that the yeast protein-protein interaction (PPI) dataset contains a larger proportion of strongly co-expressed proteins, compared to their baseline proportion in the entire yeast proteome. Similarly, yeast proteins from the same protein complex show a stronger co-expression than random proteins [14]. The interactome-transcriptome correlation demonstrated in yeast was also demonstrated for multicellular organism *C. elegans* [10, 15-16]. Another important type of association is that between expression and transcription factors (TF). It was shown for the yeast genome that when the same TFs target the same genes, these genes exhibit stronger co-expression than randomly selected ones [17].

The existence of two associations, PPI-expression and expression-TF locations, implies that there should also exist an association between PPI and TF locations. Consistent with this expectation, it was shown for proteins from the human *N*-methyl *D*-aspartate (NMDA) receptor that regulatory regions of the genes that encode interacting proteins are

*Address correspondence to this author at the Gen*NY*sis Center for Excellence in Cancer Genomics, Department of Epidemiology and Biostatistics, University at Albany, SUNY, One Discovery Drive, Rensselaer, NY 12144, USA; E-mail: IKuznetsov@albany.edu

1874-1363/07

targeted by similar sets of TFs [18-19]. The correlation between PPI and TF data was also employed in order to discover cooperative TF pairs that synergistically influence the expression of proteins that are located close to each other in the yeast protein-protein interaction network [20]. Correlations that involve biological pathways were also studied. Since genes that belong to the same pathway are functionally related, they can be expected to be co-expressed and co-regulated. An association between pathways and expression was shown for both tumor [21-22] and normal cells [22] from the human genome. An association between pathway data and data on transcriptional regulation was also demonstrated for several selected human pathways [18]. In yeast, relationships in a combination of three or more heterogeneous types of genome-wide datasets have also been studied [23-25].

Most integrative studies have been done on the example of the yeast genome. Because of its relative simplicity, yeast is the best experimentally characterized eukaryotic organism for which many experimental large-scale datasets, such as PPI and locations of transcription factor binding sites (TFBS), are readily available. The human genome, on the other hand, is much more complex in nature and significantly harder to study experimentally. For instance, no comprehensive experimental datasets on protein-protein interactions and TFBS locations are yet available for the human genome. Due to the absence of such experimental datasets, information about multiple genome-wide associations that involve PPI and TFBS locations in the human genome is lacking. A possible way to overcome this limitation is to study associations using computationally inferred genome-wide datasets.

In this work, we use a novel computational approach to perform a comprehensive analysis of four types of data that describe the following functional features of the human genome: functional pathways, expression profiles, inferred protein-protein interactions, and inferred locations of transcription factor binding sites. We use inferred protein-protein interactions from OPHID (Online Predicted Human Interactome Database), the largest publicly available PPI database [26], that includes 8,687 human proteins. This PPI dataset is more than two orders of magnitude larger than the dataset of only 76 proteins used in a previously reported study of correlations involving the human interactome [18]. We analyze types of associations that have not been studied previously for the human genome, including associations between expression and TFBS locations, PPI and expression, and pathway information and PPI. We study associations not only in pairwise combinations, but also in combinations of three and four data types.

METHODS

Sources of Genomic Data

This work deals with multiple heterogeneous sources of genomic data. We therefore need to use consistent unique gene identifiers for each of these sources. We utilize the human genome annotation version 38 from the Ensembl database [27] to assign a unique id to each gene and keep this id for each data type. We obtained the following four types of data for the human genome using publicly available sources:

1. Biological pathways from the KEGG database [28]. In KEGG, each gene from the human genome is assigned to one or more functional pathways. By mapping KEGG identifiers onto Ensembl identifiers, we generated a list of 4,024 genes for which pathway annotation is available.

2. Protein-protein interactions (PPI) from the OPHID database [26]. OPHID catalogs human protein-protein interactions that are either determined experimentally or inferred from known protein-protein interactions in model organisms (*S. cerevisiae*, *C. elegans*, *D. melanogaster*, and *M. musculus*). By mapping OPHID identifiers onto Ensembl identifiers, we generated a list of 8,687 genes whose protein products are annotated in the OPHID database.

3. Gene expression data from the SymAtlas database [29]. SymAtlas reports genome-scale gene expression measurements for 73 normal human tissues and 6 disease state tissues hybridized to Affymetrix HG-U133A array. Two replicates were used for each tissue. In our analysis, we excluded disease state tissues and only used 73 normal tissues. Expression levels for each tissue were averaged over the two replicates. Thus, each gene was represented by an expression profile that consists of 73 data points. By mapping Affymetrix identifiers onto Ensembl identifiers, we generated a list of 12,306 genes whose expression profiles are annotated in the SymAtlas database.

4. The data on transcription factor binding sites (TFBS) were obtained as follows. First, we used the Ensembl human genome assembly version 38 [27] to retrieve regulatory upstream region of each gene. We define regulatory upstream region as a 2KB region upstream of the transcription start site. In Ensembl, a gene can be annotated as producing multiple transcripts, 1.3 transcripts per gene on average [30]. In cases when more than one transcript is annotated for a given gene, we use known transcript with most 5' transcription start site. We choose known transcripts over novel transcripts because the former have more supporting evidence that the latter [30]. We used this procedure to retrieve regulatory upstream regions of all protein-coding genes (a total of 23,326 genes). Second, we used the Match software program [31] to scan the upstream regions for TFBS annotated in the TRANSFAC database [32]. The TRANSFAC database is a library of experimentally identified transcription factor binding sites represented in the form of a position weight matrix (PWM). Match is a tool that searches for putative TFBS in input DNA sequences by using a library of PWMs. Match was run using the library of high-quality vertebrate PWMs and the option to minimize the number of false positives. By parsing Match output, we obtained a list of putative TFBS found in the upstream regions of 23,326 human genes.

Conversion of Genomic Data into a Unified Matrix Format

Each type of genomic data was converted into a unified matrix format. In this format, a symmetric n by n matrix nu-

merically summarizes a particular type of functional relationships observed among n genes. Each of the four types of data described above was converted into a matrix format as follows.

1. KEGG pathways are represented by matrix K (size 4,024 x 4,024). An element k_{ij} in K matrix is equal to 1 if products of genes i and j belong to at least one common KEGG pathway and 0 otherwise.

2. Protein-protein interactions are represented by matrix P (size 8,687 x 8,687). An element p_{ij} in P matrix has a binary value of 1 or 0, indicating the presence or absence of protein-protein interaction between products of genes i and j.

3. Expression profiles are represented by matrix R (size 12,306 x 12,306). An element r_{ij} in R matrix is the Pearson correlation coefficient (PCC) between expression profiles of genes i and j. For the cases when at least one gene in a pair (i,j) is mapped onto multiple Affymetrix probe sets (3,837 out of 12,306 genes), we calculate PCC between all probe set pairs that correspond to (i,j) and choose a PCC with the largest magnitude. Negative correlations in R matrix were set to zero. For analyses that involve computing association scores (see below), we use a binary version of R matrix in which all elements that have values equal to or greater than 0.7 (strong correlation) are set to 1 and all elements that have values below 0.7 are set to 0.

4. The *cis*-similarity between promoter regions of genes is represented by matrix T (size 23,326 x 23,326). An element t_{ij} in T matrix is the number of unique TFBSs observed in the promoter regions of both gene i and j. Unique means that all occurrences of binding sites for the same TF are counted only once for each promoter region. For instance, if the promoter region of gene i contains 4 sites for transcription factor A and 1 site for transcription factor B, whereas the promoter region of gene j contains 2 sites for transcription factor A and 3 sites for transcription factor B, the value of t_{ij} will be equal to 2. The idea of this definition of *cis*-similarity is to attempt to account for the number of common transcription factors that control both gene i and j.

When we study a combination of two or more types of data, we only use genes for which all types of required annotation are available and exclude genes with missing annotation. For example, when we study associations between K and P matrices, we take a set of genes for which both KEGG pathway and protein-protein interaction data are annotated.

Testing the Statistical Significance of the Associations Among Multiple Data Types

The main idea of presenting a particular type of genomic data as a symmetric matrix that describes a certain type of functional relationship between gene pairs is to reveal statistically significant functional associations among multiple matrices by using multiplication of equivalent matrix elements. In general, when elements from k matrices of dimension n, $M_1 \ldots M_k$, that represent k types of genomic data for n genes are multiplied, and a final matrix is obtained, $F[i,j]=M_1[i,j]*\ldots*M_k[i,j]$ (note that this is an element-wise multiplication, not a conventional matrix product). In this final matrix F, gene pairs that exhibit strong associations across all k types of data will correspond to elements with large absolute value. The overall strength of functional associations within a group of n genes represented by k matrices can be quantified by computing the sum of all elements in the final matrix, $S(n,k)$, as follows:

$$S(n,k) = \sum_{i<j}^{n} M_1[i,j] \cdot M_2[i,j] \cdot \ldots \cdot M_k[i,j] \qquad \text{Eq. 1}$$

If $S(n,k)$ is significantly higher than that expected by chance, it will indicate that genes in the multiplied matrices exhibit a strong non-random association across k types of genomic data. We estimate the statistical significance of $S(n,k)$ by comparing it to the distribution of random scores. A random score is obtained by randomly permuting elements in each matrix $M_1, \ldots, M_k$ and then using these permuted matrices to obtain a score according to Eq. 1. For each matrix combination we generate 10,000 random scores. The p-value of the observed score, $P(R(n,k) \geq S(n,k))$, is computed as follows:

$$P(R(n,k) \geq S(n,k)) = \frac{N(R(n,k) \geq S(n,k))}{10{,}000} \qquad \text{Eq. 2}$$

where $R(n,k)$ is random score and $N(R(n,k) \geq S(n,k))$ is the number of random scores that are equal to or larger than $S(n,k)$. We applied the Shapiro-Wilk normality test and found random association scores to be normally distributed (data not shown). Histograms of the distributions of random scores can be found in Supplementary information. Since most p-values obtained from random simulations are zero, we use the z-score to rank the associations:

$$z - score(S(n,k)) = \frac{S(n,k) - < R(n,k) >}{\sigma_{R(n,k)}} \qquad \text{Eq. 3}$$

where $<R(n,k)>$ is the average and $\sigma_{R(n,k)}$ is the standard deviation of the random score.

RESULTS

Representation of Genomic Data in a Matrix Format

For each gene in the human genome, we collected four types of raw functional data (see Methods for details): KEGG pathways, protein-protein interactions, expression profiles, and putative transcription factor binding sites (see Table **1**). Each raw data type was converted into a unified matrix format, abbreviated as follows: K (KEGG pathway matrix), P (PPI matrix), R (co-expression matrix), and T (*cis*-similarity matrix). In this format, each symmetric n by n matrix numerically summarizes a particular type of functional relationships observed among n genes. An element k_{ij} in K matrix has a binary value of 1 if products of genes i and j belong to at least one common KEGG pathway and 0 otherwise. An element p_{ij} in P matrix has a binary value of 1 or 0, indicating the presence or absence of protein-protein interaction between products of genes i and j. An element r_{ij} in R matrix is the Pearson correlation coefficient (PCC) between expression profiles of genes i and j. An element t_{ij} in T matrix is a degree of *cis*-similarity between genes i and j. We define *cis*-similarity as the number of unique TFBSs shared by the promoter regions of genes i and j. Unique means that

Table 1. Four Types of Data Used in this Study

Data Type	Description	Source	Genes
K	Functional pathways	KEGG [28]	4,024
P	Protein-protein interactions	OPHID [26]	8,687
R	Expression profiles	SymAtlas [29]	12,306
T	Putative TFBS found in the promoter regions	Ensembl [27], TRANSFAC [32]	23,326

all occurrences of TFBSs for the same transcription factor are counted only once for each promoter region. Thus, our definition of *cis*-similarity is the number of common TFs involved in transcriptional control of both gene *i* and *j*.

Associations Observed Among Multiple Types Of Genomic Data

In this section, we perform a qualitative study of genome-wide associations observed among the four types of genomic data. The idea of this study is to examine whether the properties of genes with respect to one type of functional data are correlated with other types of functional data. For example, we can classify pairs of genes into interacting and non-interacting categories and examine the average correlation coefficient between their expression profiles in order to see whether expression profiles of genes whose products interact tend to have a higher correlation coefficient than the profiles of non-interacting ones. Here, we study global genome-wide associations for the following combinations of data types: K-P, R-T, K-R, P-T, K-P-R, and K-P-T and demonstrate the existence of potentially significant relationships observed among these data types. A rigorous statistical analysis of the significance of associations for all possible combinations of data types that confirms the qualitative trends discussed here is presented in the following sections of the manuscript.

First, we analyze the associations between functional pathways and protein-protein interactions (K-P association). The comparison of pathway information for interacting and non-interacting proteins shows that 40.82% of interacting protein pairs share at least one functional pathway (meaning that both proteins in the pair belong to the same pathway), whereas only 5.55% of non-interacting protein pairs share pathway annotation (Fig. **1A**). This means that interacting protein pairs are seven times more likely to participate in the same pathway than non-interacting protein pairs. Analysis of the reverse relation shows that if two proteins participate in the same pathway, they are eleven times more likely to interact than proteins from different pathways (Fig. **1B**).

Second, we analyze the associations between co-expression and *cis*-similarity of promoter regions (R-T association). This analysis shows that, on average, correlation between expression profiles of genes that share common TFBS is higher (PCC=0.375) than that between expression profiles of genes that do not share any common TFBS (PCC=0.369) (Fig. **1C**). Analysis of the reverse relation shows that an increase in the level of co-expression of gene pairs is associated with an increase in the number of common TFBS found in their promoter regions (Fig. **1D**). These results confirm to an empirical expectation that co-expressed genes should have similar cis-profiles and *vice versa*. However, the trends shown in Figs. (**1C**) and (**1D**) are very subtle and their statistical significance is not obvious. One possible reason of weak trends is that the computational identification of putative TFBS *via* sequence motif-based methods is inherently prone to noise because of a very high percentage of false positive predictions [33]. We will address the issue of statistical significance in the next section of the manuscript.

Third, we analyze the following three types of associations: K-R, P-R, and K-P-R. We divided all gene pairs into six categories according to whether their products are interacting and/or participating in same functional pathways and compared the average correlations between expression profiles for these six categories (Fig. **1E**). From right to left in Fig. (**1E**), the largest average PCC between expression profiles is found for gene pairs that both interact and participate in same pathways (PCC = 0.4331), whereas the smallest average PCC is found for gene pairs that neither interact nor participate in same pathways (PCC = 0.3216). We also observe that the average PCC is higher for gene pairs that participate in same pathways (PCC = 0.4028) than for interacting pairs (PCC = 0.3773). These observations suggest that, with respect to concerted expression, genes from the same pathway act as a more cohesive biological module than genes producing physically interacting proteins. Experimental evidence shows that interacting proteins from the same complex are not necessarily produced by co-regulated genes. For example, cyclin-dependent kinase and cyclin together form a protein complex. While the former is produced from a constantly transcribed gene, the latter is produced in a regulated manner [13].

Fourth, we analyze K-T, P-T, and K-P-T associations by computing the average number of common TFBS for the same six categories of gene pairs described above. The results of this analysis, shown in Fig. (**1F**), reveal a trend very similar to the one shown in Fig. (**1E**): the largest number of common TFBS is observed for gene pairs that both interact and participate in the same pathways, whereas the smallest number of common TFBS is observed for gene pairs that neither interact nor participate in the same pathway. These two related trends indirectly indicate that the level of co-expression (measured by PCC) and the *cis*-similarity (measured by the number of common TFBS) are correlated with each other, which is in agreement with the direct relationship between them shown in Figs. (**1C**) and (**1D**). The small differences in the number of common TFBS observed in Fig. (**1F**) can be attributed to the fact that the computational procedure for the identification of putative TFBS produces a very large number of false positives [33].

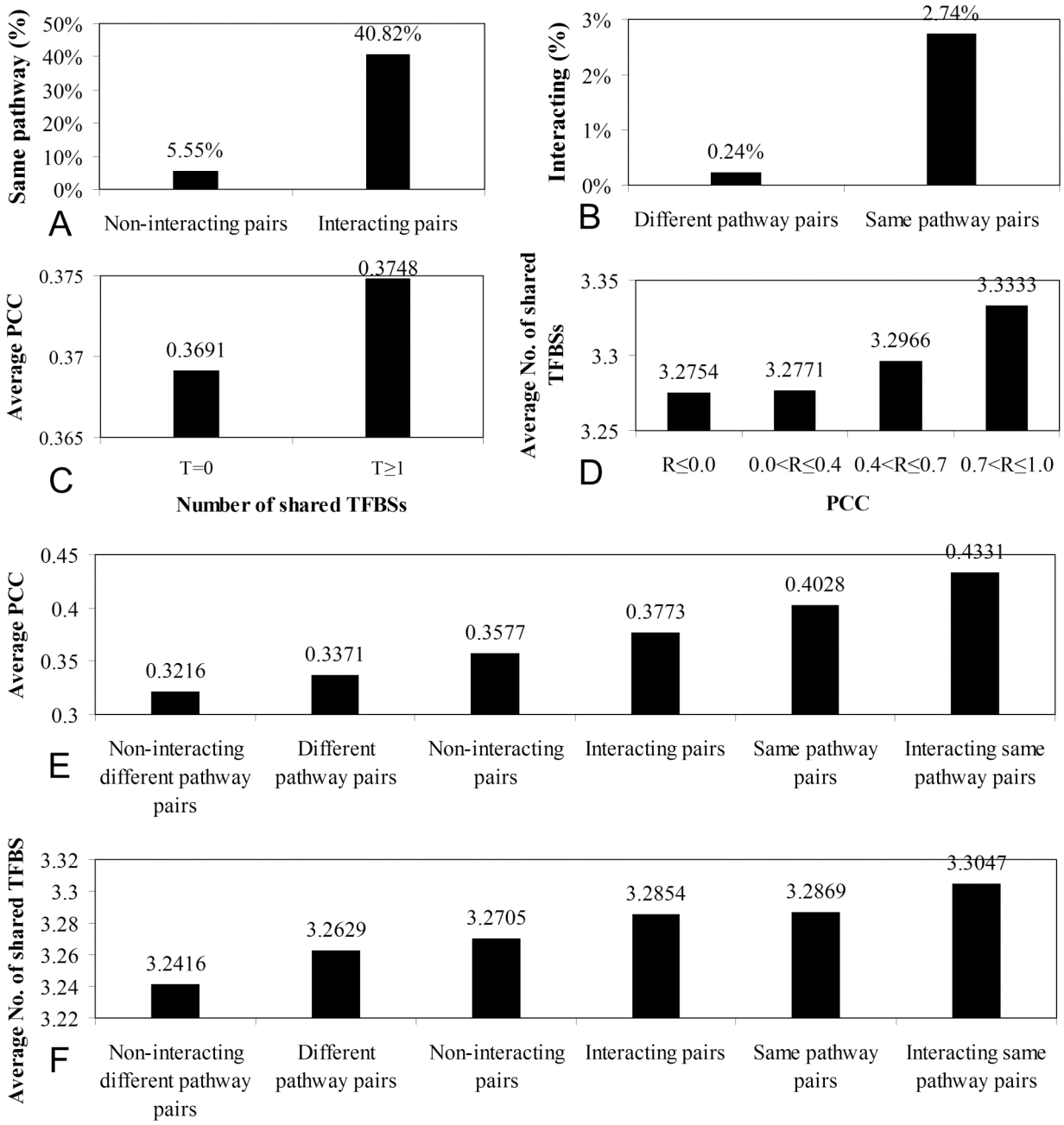

Fig. (1). Associations exist among the four types of functional data. (**A**) Interacting protein pairs are more likely to participate in the same pathway than non-interacting protein pairs. (**B**) Protein pairs from the same pathway are more likely to interact than protein pairs from different pathways. (**C**) A pair of genes that share common TFBS in the promoter regions shows a higher correlation between expression profiles than a pair without any shared TFBS. (**D**) An increase in correlation between expression profiles is associated with an increase in the number of shared TFBS. (**E**) Pairs of proteins from the same pathway and/or pairs of interacting proteins are more likely to show correlated expression. (**F**) Pairs of protein from the same pathway and/or pairs of interacting proteins are more likely to share common TFBS in their promoter regions.

Statistical Significance of the Associations Among Multiple Types of Genomic Data

The qualitative analyses shown in the previous section indicate the existence of potentially significant associations among various types of 'omic' data. In this section, we use a rigorous quantitative approach to evaluate the statistical significance of the observed associations on the genome-wide scale. Given two or more matrices that represent particular types of 'omic' data for a group of genes, we measure the strength of association among these data types by means of an association score. This association score is defined as the sum of products between corresponding elements of the matrices under consideration (Eq. 1, see Methods for details). Statistical significance of the observed association score is estimated by comparing it to the distribution of random association scores obtained from randomly permuted matrices.

Table 2. P-Values and Z-Scores Estimated from the Random Permutation Experiment

	Type of Association	Genes	P-Value	Z-Score
Pairwise combinations	K-P	2,424	0	160.97
	K-R	3,271	0	85.35
	R-T	11,775	0	75.63
	P-R	6,784	0	7.70
	K-T	3,939	0	7.59
	P-T	8,513	0.0498	1.66
Triple combinations	K-P-T	2,424	0	142.26
	K-P-R	2,154	0	82.22
	K-R-T	3,271	0	73.90
	P-R-T	6,784	0	7.47
Quadruple combination	K-P-R-T	2,154	0	75.65

Since we have four matrices that correspond to the four types of data, there are eleven possible combinations of two, three and four matrices. P-values and z-scores for each combination are shown in Table **2**. Histograms of all random distributions can be found in the Supplementary information, with three selected distributions shown in Figs. (**2-4**).

The results shown in Table **2** indicate that all eleven combinations of data types demonstrate significant associations as evidenced by low *p*-values. Below, we briefly discuss biological implications of each association. The results for pairwise combinations indicate:

K-P (z-score=160.97, p=0) - the existence of a highly significant association between protein-protein interactions and protein function. Since in our methodology associations are not directional, K-P association is equivalent to P-K association, thus implying that interacting proteins tend to participate in the same functional pathway, and *vice versa*, proteins from the same functional pathway tend to interact.

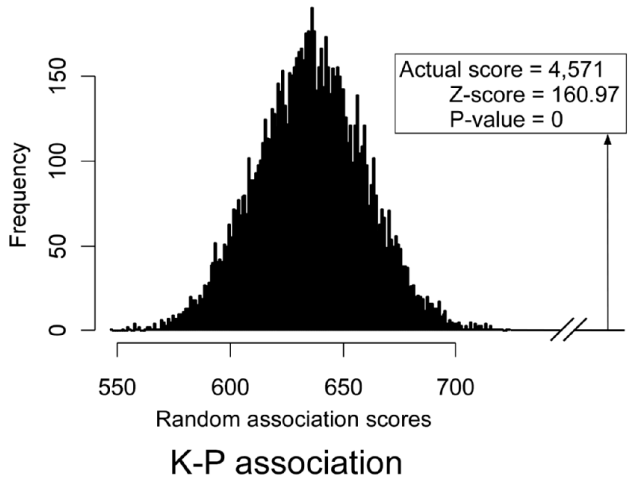

Fig. (2). The distribution of random K-P scores.

K-R (z-score=85.35, p=0) - the existence of a highly significant association between co-expression and gene function. It shows that genes participating in the same pathway tend to be co-expressed, and *vice versa*, co-expressed genes tend to participate in the same pathway.

R-T (z-score=75.63, p=0) - co-expressed genes tend to share similar *cis*-profiles, and *vice versa*, genes with similar *cis*-profiles tend to be co-expressed.

P-R (z-score=7.70, p=0) - genes that encode interacting proteins tend to be co-expressed, and *vice versa*, co-expressed genes tend to encode interacting proteins.

K-T (z-score=7.59, p=0) - genes from the same pathway tend to have similar *cis*-profiles, and *vice versa*, genes with similar *cis*-profiles tend to participate in the same pathway.

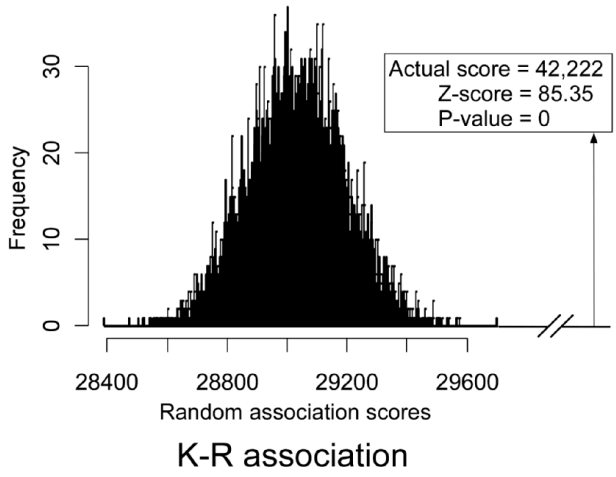

Fig. (3). The distribution of random K-R scores.

P-T (z-score=1.66, p=0.0498) - the existence of a marginally significant association between protein-protein interactions and the similarity of *cis*-profiles of the genes that encode interacting proteins.

The results for all combinations of three data types, described below, also demonstrate highly statistically significant genome-wide associations:

K-P-T (z-score=142.26, p=0) - the existence of a highly significant association that links gene function, K, interac-

tions between gene products, P, and *cis*-similarity of the promoter regions, T. This association implies that genes from the same pathway both tend to code for interacting proteins and share a similar set of TFs in their promoter regions.

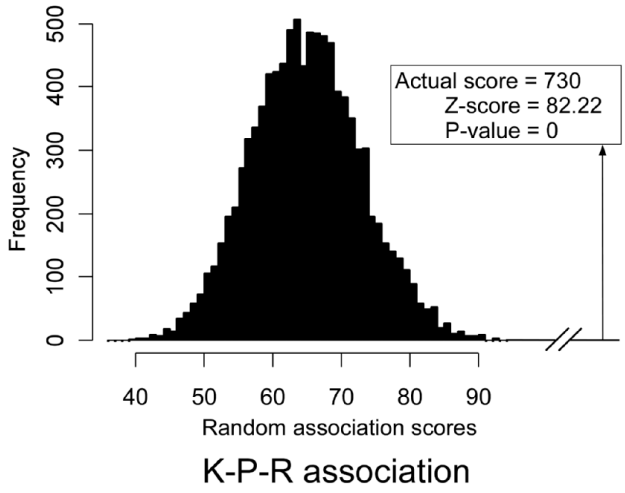

Fig. (4). The distribution of random K-P-R scores.

K-P-R (z-score=82.22, p=0) - the existence of a highly significant association that links gene function, K, interactions between gene products, P, and co-expression, R. Biologically, this association is similar to K-P-T association and implies that genes from the same pathway both tend to code for interacting proteins and to be co-expressed.

K-R-T (z-score=73.90, p=0) - the existence of a highly significant association that links gene function, K, co-expression, R, and *cis*-similarity of the promoter regions, T. This association implies that genes from the same pathway tend to be both co-expressed and have a similar set of TFs in their promoter regions.

P-R-T (z-score=7.47, p=0) - the existence of a significant association that links interactions between gene products, P, co-expression, R, and *cis*-similarity of the promoter regions, T. This association implies that genes whose products interact tend to be co-expressed and have a similar set of TFs in their promoter regions. However, it should be pointed out that the strength of P-R-T association is much weaker than that of other triple associations as indicated by a considerably lower z-score.

Finally, the results for the combination of all four data types, K-P-R-T, indicate that this quadruple association is also highly significant (z-score=75.65, p=0). This association indicates that genes from the same pathway simultaneously tend to encode interacting proteins, be co-expressed, and have a similar set of TFs in their promoter regions.

Pathway-Level Analysis of the Associations Among Data Types

The analysis reported in the previous section summarizes global genome-wide relations among data types by considering all genes in the human genome simultaneously. A similar analysis can be performed by considering a group of genes that belong to a particular functional category. A good example of a functional category is a functional pathway, which can be considered as a biological module that carries out a specific genomic function. Depending on the function of the pathway, one may expect certain pathway-specific associations to be more pronounced than the others. In this section, we use the classification of functional pathways from the KEGG database [28]. The main difference from the global analysis reported in the previous section is that pathway-level analysis is done for a group of functionally related genes that belong to a particular KEGG pathway. This kind of analysis enables us to categorize the associations between the pathway data and other types of genomic data and to determine which types of associations are most profound in a particular functional category. Since the TFBS data seem to be noisy and therefore least reliable, we use only the PPI data and the gene expression data for the pathway-level analysis. This leaves us with three combinations to analyze for pathway-level associations: K-P, K-R and K-P-R. These three combinations provide the following biological information for a given pathway: K-P describes its relative enrichment in interacting proteins, K-R describes its relative enrichment in co-expressed genes, and K-P-R describes its relative enrichment in genes that are both co-expressed and code for interacting proteins. When we use Eqs. 2,3 to analyze a group of *m* genes that belong to a particular pathway, each random *m* by *m* matrix for a given type of data is obtained by randomly sampling, without replacement, *m* genes from a list of all human genes annotated with this particular data. Pathways containing less than five annotated genes were excluded from this analysis.

Out of 174 pathways annotated in the KEGG database for the human genome, we identified 98 pathways that are significantly (p<0.05) enriched in interacting proteins, 34 pathways that are significantly enriched in co-expressed genes, and 75 pathways that are significantly enriched in genes that are both co-expressed and code for interacting proteins. Lists for all pathways and all combinations, ranked by z-score, are given in Supplementary information. Top 10 scoring pathways for each combination are shown in Tables **3**, **4** and **5**. It should be noted that when pathways are analyzed with respect to concomitant enrichment in co-expressed genes whose protein products also interact (the triple K-P-R association), several additional pathways emerge as significant (Table **5** and Supplementary Table **3**). For example, two

Table 3. Top 10 Pathways Enriched in Interacting Proteins

KEGG ID	Z-Score	Pathway Name
hsa03050	237.19	Proteasome
hsa03020	134.22	RNA polymerase
hsa03010	86.94	Ribosome
hsa00193	61.66	ATP synthesis
hsa00240	48.63	Pyrimidine metabolism
hsa04110	44.15	Cell cycle
hsa03022	40.98	Basal transcription factors
hsa04130	39.30	SNARE interactions in vesicular transport
hsa00020	39.19	Citrate cycle (Tca cycle)
hsa04350	34.76	Tgf-beta signaling pathway

pathways ('Glutamate metabolism' and 'Glutathione metabolism') are identified as showing significant concomitant enrichment, even though they do not show enrichment in interacting proteins or co-expressed genes. Some pathways are concomitantly enriched even though they show enrichment in either interacting proteins or co-expressed genes, but not both. For example, 'Cholera' pathway shows very strong concomitant enrichment (top 5th in Table **5**), but it does not show enrichment in co-expressed genes. Similarly, 'Olfactory transduction' pathway shows a significant concomitant enrichment without being enriched in interacting proteins. These observations indicate that combining multiple types of genomic data reveals additional functional features of individual pathways that cannot be revealed by studying simple pairwise associations.

Table 4. Top 10 Pathways Enriched in Co-Expressed Genes

KEGG ID	Z-Score	Pathway Name
hsa04080	9.73	Neuroactive ligand-receptor interaction
hsa04630	5.11	Jak-Stat signaling pathway
hsa04620	4.94	Toll-like receptor signaling pathway
hsa00190	4.91	Oxidative phosphorylation
hsa04020	4.69	Calcium signaling pathway
hsa00602	4.60	Glycosphingolipid biosynthesis-neo-lactoseries
hsa04010	3.86	MAPK signaling pathway
hsa04060	3.81	Cytokine-cytokine receptor interaction
hsa04730	3.57	Long-term depression
hsa04664	3.51	Fc epsilon RI signaling pathway

Table 5. Top 10 Pathways Enriched in Both Interacting and Co-Expressed Genes

KEGG ID	Z-Score	Pathway Name
hsa00193	78.58	ATP synthesis
hsa03050	76.25	Proteasome
hsa03020	44.87	RNA polymerase
hsa00190	36.83	Oxidative phosphorylation
hsa05110	30.64	Cholera
hsa04350	29.41	TGF-beta signaling pathway
hsa04660	27.01	T cell receptor signaling pathway
hsa03010	25.52	Ribosome
hsa05120	25.18	Epithelial cell signaling in helicobacter pylori infection
hsa04610	23.94	Complement and coagulation cascades

DISCUSSION

In general, the results of the quantitative analysis of the genome-wide pairwise associations are consistent with the qualitative study performed on the same datasets (see Fig. **1**) and discussed in the first section of Results, thus confirming the utility of the proposed approach. For instance, Fig. (**1E**) shows that the average correlation between expression profiles is larger for gene pairs from the same pathway, K, (PCC=0.4028) than for gene pairs that encode interacting proteins (PCC=0.3773). This observation is consistent with a larger z-score observed for K-R association (z-score=85.35) compared to that for P-R association (z-score=7.70). Similarly, Fig. (**1F**) shows that the average number of shared TFBS is larger for gene pairs from the same pathway (3.2869) than for gene pairs whose products interact (3.2705). This observation is also consistent with a larger z-score observed for K-T association (z-score=7.59) compared to that for P-T association (z-score=1.66). If we assume that K and P matrices contain similar amounts of noise, then the observation that the z-scores for K-R and K-T associations are larger than those for P-R and P-T suggests that transcriptional co-regulation is more important for genes from the same pathway than for genes that encode interacting proteins. It should also be noted that, to the best of our knowledge, out of the six pairwise combinations of data types utilized in this work, three (P-R, R-T, and K-P) have never been studied for the human genome.

The strongest pairwise associations, indicated by very high z-scores, are observed for combinations involving pathway data, K-P and K-R. This observation is consistent with empirical expectations and confirms that genes from the same functional pathway tend to be co-expressed and code for interacting proteins. The only marginally significant genome-wide association is observed between PPI data and *cis*-similarity of promoter regions (P-T combination, p=0.0498, Table **2**). The relatively low z-scores for two associations involving the T matrix (P-T and K-T) are not straightforward to interpret. On one hand, a large amount of noise present in matrix T may dampen real biological associations. On the other hand, the R-T association is quite significant (z-score=75.63) despite the noise present in the T matrix. The observation that P-R association, which is related to P-T, also has a relatively low z-score of 7.7 provides an additional argument in favor of the assumption that the weakness of the genome-wide P-T association may reflect a real biological phenomenon.

The application of our methodology to study associations in groups of genes from individual functional pathways shows that pathways enriched in interacting proteins (K-P association, Table **3** and Supplementary Table **1**) are mostly the ones for genetic information processing. These pathways tend to contain large protein complexes, such as the ribosome and DNA/RNA polymerases. Pathways enriched in co-expressed genes (K-R association, Table **4** and Supplementary Table **2**) are mostly the pathways for environmental information processing. These pathways can be thought of as biological modules whose genes need to be expressed in a concerted manner in response to external stimuli. Metabolic pathways seem to be under-represented in the list of pathways enriched in co-expressed genes. There are 112 metabolic pathways, comprising 64% of all 174 annotated pathways. However, out of the total of 34 pathways significantly enriched in co-expressed genes, only six (18%) are metabolic pathways. This observation is consistent with previously reported results that metabolic pathways do not show similar *cis*-profiles [18].

Another possible application of the proposed methodology is to benchmark the quality of various large-scale datasets. In this work, we used PPI from the OPHID database [26], where about 60% of all annotated interactions were inferred computationally, rather than obtained experimentally. Obviously, the quality of this inference needs to be validated. Since proteins that participate in the same functional pathway often form multi-protein complexes and can be expected to interact, the strength of K-P association can be used as an indicator of the non-randomness of PPI annotations. The fact that, according to our results, K-P association ranks highest among all pairwise combinations studied, suggests that the assignment of PPI in groups of functionally related proteins is highly non-random, thus confirming the quality of the OPHID annotation. Therefore, the present work can also be considered as an independent validation of OPHID, in addition to the validation provided by the authors of this database. A similar approach can be used to benchmark other types of data. For instance, given several methods for finding TFBS in promoter regions, the R-T association experiment can be used as a quantitative evaluation procedure to benchmark which of these methods gives the best correlation with expression data.

CONCLUSIONS

We presented a methodology for quantifying the significance of associations observed among multiple heterogeneous types of 'omic' data. We used this methodology to analyze and rank associations among four types of data in humans: functional pathways, protein-protein interactions, expression profiles, and transcription factor binding sites. Using the largest datasets currently available for the human genome, we showed that associations in all combinations of data types are significantly higher than random expectation. We find that pathways involved in genetic information processing are enriched in interacting proteins, and that signaling pathways are enriched in co-expressed genes. We show that combinations of multiple data types provide additional insights into the properties of functional pathways. This work can also be considered as an independent validation of the inferred protein-protein interactions annotated in the OPHID database.

ABBREVIATIONS

PPI	=	protein-protein interaction
TF	=	Transcription factor
TFBS	=	Transcription factor binding site
PCC	=	Pearson correlation coefficient
K	=	KEGG pathway data matrix
P	=	PPI data matrix
R	=	Co-expression data matrix
T	=	*Cis*-similarity data matrix

REFERENCES

[1] M. Gerstein, N. Lan, R. Jansen, "Proteomics. Integrating interactomes," *Science*, vol. 295, pp. 284-287, January 2002.

[2] D. Greenbaum, N. M. Luscombe, R. Jansen, J. Qian, M. Gerstein, "Interrelating different types of genomic data, from proteome to secretome: 'oming in on function," *Genome Res.,* Vol. 11, pp. 1463-1468, September 2001.

[3] M. Vida, "A biological atlas of functional maps," *Cell,* Vol. 104, pp. 333-339, February 2001.

[4] S. Wachi, K. Yoneda, R. Wu, "Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues," *Bioinformatics*, Vol. 21, pp. 4205-4208, February 2005.

[5] A. Drawid, M. Gerstein, "A Byesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome," *J. Mol. Biol.,* Vol. 301, pp. 1059-1075, December 2001.

[6] J. Qian, J. Lin, N. M. Luscombe, H. Yu, M. Gerstein, "Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data," *Bioinformatics*, Vol. 19, pp. 1917-1926, August 2003.

[7] L. V. Zhang, S. L. Wong, O. D. King, F. P. Roth, "Predicting co-complexed protein pairs using genomic and proteomic data integration," *BMC Bioinformatics*, Vol. 5, pp. 38, April 2004.

[8] J. S. Bader, A. Chaudhuri, J. M. Rothberg, J. Chant, "Gaining confidence in high-throughput protein interaction networks," *Nat. Biotechnol.* Vol. 22, pp. 78-85, January 2004.

[9] C. S. Goh, T. A. Gianoulis, Y. Liu, *et al.* "Integration of curated databases to identify genotype-phenotype associations," *BMC Genomics,* Vol. 7, pp. 257, October 2006.

[10] K. C. Gunsalus, H. Ge, A. J. Schetter, *et al.* "Predictive models of molecular machines involved in Caenorhabditis elegans early embryogenesis," *Nature,* Vol. 436, pp. 861-865, August 2005.

[11] I. Lee, S. V. Date, A. T. Adai, E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science,* Vol. 306, pp. 1555-1558, November 2004.

[12] A. Grigoriev, "A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae," *Nucleic Acids Res.,* Vol. 29, pp. 3513-3519, September 2001.

[13] H. Ge, Z. Liu, G. M. Church, M. Vidal, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae," *Nature Genet*., Vol. 29, pp. 482-486, December 2001.

[14] R. Jansen, D. Greenbaum, M. Gerstein, "Relating whole-genome expression data with protein-protein interactions," *Genome Res.*Vol. 12, pp. 37-46, January 2002.

[15] A. J. Walhout, J. Reboul, O. Shtanko, *et al.* "Integrating interactome, phenome, and transcriptome mapping data for the C. elegans germline," *Curr. Biol.,* Vol. 12, pp. 1952-1958, November 2002.

[16] S. Li, H. Ge, T. Hao, *et al.* "A map of the interactome network of the metazoan C. elegans," *Science,* Vol. 303, pp. 540-543, January 2004.

[17] H. Yu, N. M. Luscombe, J. Qian, M. Gerstein, "Genomic analysis of gene expression relationships in transcriptional regulatory networks," *Trends Genet.,* Vol. 19, pp. 422-427, August 2003.

[18] S. Hannenhalli, S. Levy, "Transcriptional regulation of protein complexes and biological pathways," *Mamm. Genome*, Vol. 14, pp. 611-619, September 2003.

[19] O. Alter, G. H. Golub, "Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription," *Proc. Natl. Acad. Sci. USA*, Vol. 101, pp. 16577-16582, November 2004.

[20] N. Nagamine, Y. Kawada, Y. Sakakibara, "Identifying cooperative transcriptional regulations using protein-protein interactions," *Nucleic Acids Res.,* Vol. 33, pp. 4828-4837, August 2005.

[21] H. H. Yang, Y. Hu, K. H. Buetow, M. P. Lee, "A computational approach to measuring coherence of gene expression in pathways," *Genomics,* Vol. 84, pp. 211-217, July 2004.

[22] R. Huang, A. Wallqvist, D. G. Covell, "Comprehensive analysis of pathway or functionally related gene expression in the National Cancer Institute's anticancer screen," *Genomics,* Vol. 87, pp. 315-328, March 2006.

[23] A. Tanay, R. Sharan, M. Kupiec, R. Shamir, "Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data," *Proc. Natl. Acad. Sci. USA*, Vol. 101, 2981-2986, March 2004.

[24] D. Hwang, A. G. Rust, L. Hood, *et al.* "A data integration methodology for systems biology," *Proc. Natl. Acad. Sci. USA*, Vol. 102, pp. 17296-17301, November 2005.

[25] P. Carmona-Saez, M. Chagoyen, A. Rodriguez, O. Trelles, J. M. Carazo, A. Pascual-Montano, "Integrated analysis of gene expression by association rules discovery," *BMC Bioinformatics,* Vol. 7, pp. 54, February 2006.

[26] K. R. Brown, I. Jurisica, "Online predicted human interaction database," *Bioinformatics,* Vol. 21, pp. 2076-2082, May 2005.

[27] T. Hubbard, D. Andrews, M. Caccamo, *et al.* "Ensembl 2005," *Nucleic Acids Res.,* Vol. 33, pp. D447-D453, January 2005.

[28] M. Kanehisa, S. Goto S, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Res.,* Vol. 28, pp. 27-30, January 2000.

[29] A. I. Su, H. Lapp, J. Zhang *et al.* "A gene atlas of the mouse and human protein-encoding transcriptomes," *Proc. Natl. Acad. Sci. USA,* Vol. 101, pp. 6062-6267, April 2005.

[30] V. Curwen, E. Eyras, T. D. Andrews, *et al.* "The Ensembl automatic gene annotation system," *Genome Res.,* Vol. 14, pp. 942-950, May 2004.

[31] A. E. Kel, E. Gossling, I. Reuter, E. Cheremushkin, O. V. Kel-Margoulis, E. Wingender, "MATCH: a tool for searching transcription factor binding sites in DNA sequences," *Nucleic Acids Res*., Vol. 31, pp. 3576-3579, July 2003.

[32] V. Matys, E. Fricke, R. Geffers, *et al*. "TRANSFAC: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res.,* Vol. 31, pp. 374-378, January 2003.

[33] M. Robinson, Y. Sun, R. T. Boekhorst, *et al.* "Improving computational predictions of cis-regulatory binding sites," *Pac. Symp. Biocomput*., Vol. 11, pp. 391-402, November 2006.

Received: May 15, 2007 Revised: May 22, 2007 Accepted: May 26, 2007