2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing

Connecting biological themes using a single human network of gene associations

Hai Fang, Ji Zhang and Kan-Kan Wang

State Key Laboratory of Medical Genomics
Ruijin Hospital affiliated to Shanghai Jiao-Tong University School of Medicine, Shanghai, China

hfang@sibs.ac.cn, jizhang@sibs.ac.cn, kankanwang@shsmu.edu.cn,

Abstract: **The accumulations of biological data in various knowledge domains (biological themes) provide rich resources for extracting deeper biological insights into biosystems. However, relations among these biological themes remain uncharacterized. Here, we present a systematic approach to the discovery of these relationships using a single human network of gene associations. We first constructed the network by incorporating the compiled human interactome and the predicted human associatome resulting from the Bayesian supervised integration of functional annotation data, model organism functional linkage data, and human functional linkage data. Using the network, we then performed the binomial distribution-based enrichment assessment to examine the high-level relationships among biological themes. Applications of the approach in biological processes from the Gene Ontology indicated that, although intra-process connections are of highly modular, distinct processes differ considerably in their abilities to interconnect with others. We also showed that such interconnectedness of processes cannot be explained by the modular constraints, but is largely due to the connectivity of their individual members to other processes. Moreover, we extended applications in other biological themes to find connections among regulatory profiles of transcription factors and microRNAs. These results demonstrated the feasibility of the approach, combining network biology with systematic information, to characterize high-level connections of any new biological themes.**

Keywords: **human gene network; gene associations; biological themes; interconnectedness matrix**

INTRODUCTION

Recent advances in technologies have propelled the rapid accumulation of large-scale biological data in various knowledge domains (biological themes). On one hand, the high-throughput omics technologies quantify genome-wide biological information at the multiple levels. On the other hand, technologies in the field of computational biology and bioinformatics speed up curations of biological annotated databases, as highlighted by databases of gene annotations (e.g., Gene Ontology [1;2]), transcription factor–targeted genes (e.g., TRANSFAC [3]), microRNA target genes (e.g., miRBase [4]). With the availability of these quantified biological information and curated databases, the next task is to how to explore higher level relations among these biological themes. Integrated gene networks [5;6] offer promising opportunities for exploring such relationships. In a variety of species, networks constructed through Bayesian integration of multiple sources of data have proved powerful to predict gene function and perturbation phenotypes [7-9].

In this study, we aim to construct a single human network of gene associations with high-coverage and high confidence with the tasks of inferring the relationships among various biological themes. We reason that these high-level relationships are principally encoded in the gene-level topological structure of the network. The strength of relationships can be measured by the enrichment of the observed links against the expected links under the binomial distribution-based model. As demonstrated by characterizing interplays among diversified biological themes, our proposed approach provides an analytical framework for systems-level surveys in the diversified aspects of human biomedicine.

METHODS

A. Outline

The procedures to implement the approach described in this setting are summarized in Figure 1. We first constructed the network (HumanNet) through the Bayesian supervised integration of four predictive data sets, and the further compilation of the existing human interactome data (**Figure 1A**). To quantify the relationships of biological themes, we then calculated binomial transformed Z-scores, measuring the strength of topological links of their individual members mapped to HumanNet. After the calculation of Z-scores of the observed links against the expected links under the binomial distribution-based model, we applied hierarchical clustering of the Z-score matrix, followed by the subsequent statistical significance assessment. The resultant interconnectedness matrix permits a global view of these relationships, wherein some biological themes tend to be highly connected together while others are disconnected into single nodes corresponding to specific themes (**Figure 1B**).

B. The probabilistic supervised Bayesian formalism

Bayesian statistical approach [5;6] provides a supervised learning framework for integrating highly heterogeneous types of data into a single coherent network of gene associations in human. The approach measures the likelihood of associations between gene pairs conditioned on the predictive data sets considered using Bayes rule. Taking into account the dependencies among the predictive data sets and to avoid the overestimation, only the data set with largest data-set likelihood ratio remains for the integration. Thus, the maximum likelihood ratio (LR) is calculated to measure the

978-0-7695-3739-9/09 $25.00 © 2009 IEEE
DOI 10.1109/IJCBS.2009.88

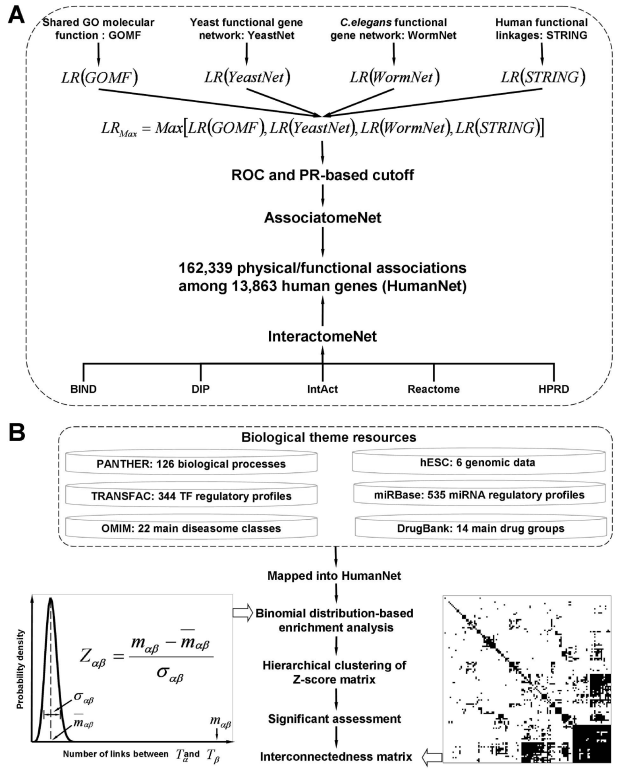

Figure 1. Schematic flow diagram of the proposed methodology to connect the biological themes using a human network of gene associations. (A) The construction of the human gene network (HumanNet). Four predictive data sets (GOMF, 'shared molecular function annotations of GO; YeastNet, 'yeast functional gene network; WormNet, '*C.elegans* functional gene network; STRING, 'human functional linkages) are first rescored with likelihood ratio, and then integrated in a maximum Bayesian model into human associatome network (AssociatomeNet) under the empirical cutoff determined by receiver operator characteristic (ROC) and precision-recall (PR) analyses. Additional associations from human interactome network (InteractomeNet, 'BIND, DIP, IntAct, Reactome and HPRD) are also included to create the final network of high-confidence and high-coverage associations. **(B)** The network-oriented enrichment analysis in terms of high-level relations among biological themes. The biological themes compiled from diversified resources are first mapped into the network, followed by the Binomial distribution-based enrichment analysis of the observed links against the expected links among their members. The resultant Z-score matrix is subjected to hierarchical clustering and statistical significance assessment of each relationship, resulting in interconnectedness matrix as the representation of potential relationships among biological themes.

likelihood of associations between gene pairs, given the presence of predictive data sets (**Eq. 1**).

$$LR(d_1,\cdots,d_n)=\underset{i=1}{\overset{n}{Max}}\frac{\Pr(d_i|A)}{\Pr(d_i|\sim A)} \quad \text{[Eq. 1]}$$

Where $\Pr(d_i|A)$ and $\Pr(d_i|\sim A)$ are respectively the conditional probabilities for gene pairs to be associated (A) or not to be associated ($\sim A$), predicted by the considered data set d_i, respectively.

C. The construction of human gene network

Diverse types of data sets differ considerably in their utility for inferring human gene associations. Such gene associations can be either direct or indirect. The direct associations are highlighted by human physical protein-protein interactions (i.e., BIND [10], DIP [11], IntAct [12], Reactome [13] and HPRD [14]), representing gold-standard associations. These interactions are compiled into human interactome network (termed as InteractomeNet). The indirect associations are functional, rather than physical, which can be predicted and transferred from heterogeneous functional genomic data across organisms. Four predictive data sets are integrated in a maximum Bayesian model into human associatome network (termed as AssociatomeNet), including shared molecular function annotations of GO (GOMF) [2], yeast functional gene network (YeastNet) [8], C.elegans functional gene network (WormNet) [9] and human functional linkages (STRING) [15]. Using the gold-standard references of associations and non-associations, likelihood ratios can be calculated for each predictive data set to determine how the predictive data set impacts the chance that gene pairs are associated.

The gold-standard positive (GSP) and negative (GSN) gene associations are defined as references to evaluate and integrate predictive data sets. GSP gene associations are based on HPRD, containing 34,989 distinct interactions among 9,456 distinct protein-encoding genes. As for GSN gene associations, lists of gene pairs assigned to separate subcellular compartments (i.e., plasma membrane *vs.* nucleus according to Gene Ontology) are first created, and then restricted to those annotated in HPRD, resulting in 1,372,503 unique gene pairs of non-associations in total. Thus, estimations of the conditional probabilities $\Pr(d_i|A)$ and $\Pr(d_i|\sim A)$ in **Eq. 1** can be obtained by calculating the fraction of gene pairs in the data set d_i that are found in the GSP or GSN, respectively (**Eq. 2**).

$$LR(d_1,\cdots,d_n)=\underset{i=1}{\overset{n}{Max}}\frac{\Pr(d_i|GSP)}{\Pr(d_i|GSN)} \quad \text{[Eq. 2]}$$

Gene pairs that share the small, specific molecular functions are more likely to be associated than those sharing large, general functions. Molecular function annotations are downloaded from the Gene Ontology (GO), deriving 28,169 assignments of 15,850 genes to one or more of 1,557 molecular functions. As a measure of functional similarity, the smallest shared molecular function (SSMF) for each pair

of annotated genes is first identified. Then, gene pairs with increasing number of SSMF (i.e., 5, 10, 50, 100, 500, and 1,000) are binned, followed by the calculation of likelihood ratios for each bin by testing against GSP and GSN (**Eq. 3**).

$$LR(GOMF) = \frac{\Pr(GOMF|GSP)}{\Pr(GOMF|GSN)} \quad \text{[Eq. 3]}$$

The yeast functional gene network (YeastNet) covers 102,803 linkages among 5,483 yeast genes. Each linkage is associated with log-likelihood score (LLS) which measures the probability of a true functional linkage between two yeast genes. Using yeast-human homology data retrieved via INPARANOID [16], YeastNet is transferred into humanized functional network, involving 14,281 functional associations among 1,375 human orthologs. These associations are then binned by LLS to assess the likelihood ratios, based on the intersection with GSP and GSN (**Eq. 4**).

$$LR(YeastNet) = \frac{\Pr(YeastNet|GSP)}{\Pr(YeastNet|GSN)} \quad \text{[Eq. 4]}$$

The C.elegans functional gene network (WormNet) comprises 384,700 linkages among 16,113 worm genes. LLS of each linkage indicate the probability of a true functional linkage between two C.elegans genes. Complemented by INPARANOID human-worm homology data [16], humanized functional network is created from WormNet, containing 36,201 functional associations among 2,661 human orthologs. Following the calculation of overlaps between LLS-binned associations with GSP and GSN, the likelihood ratios for WormNet data set can be assessed (**Eq. 5**).

$$LR(WormNet) = \frac{\Pr(WormNet|GSP)}{\Pr(WormNet|GSN)} \quad \text{[Eq. 5]}$$

STRING contains a comprehensive body of known and predicted protein-protein associations, containing 545,521 associations among 13,297 human genes. Each binary association is annotated with a combined confidence score resulting from three different types of evidence: genomic context associations (conserved gene neighborhood, gene fusion, and phylogenetic co-occurrence), high-throughput experimental data (physical protein interactions and gene coexpression), and the mining of databases and literatures. Combined scores are grouped into 6 bins of increasing confidence (i.e., 250, 400, 550, 700, 850, 1000), which are subsequently tested against GSP and GSN to assign the likelihood ratios for STRING data set (**Eq. 6**).

$$LR(STRING) = \frac{\Pr(STRING|GSP)}{\Pr(STRING|GSN)} \quad \text{[Eq. 6]}$$

Since the possible correlations and redundant information among the predictive data sets (i.e., GOMF, YeastNet, WormNet, STRING), especially those multiple resources-curated functional gene networks, the likelihood rations from these predictors are integrated in the maximum Bayesian approach to obtain the most conservative estimations LR_{Max}. To determine the cutoff that achieves high accuracy of predictions while maintaining high coverage, analyses of receiver operator characteristic (ROC) and precision-recall (PR) are employed, respectively. Within the true associations (i.e., GSP) and the true non-associations (i.e., GSN) as a function of the specific threshold LR_{Max}^{cutoff}, ROC analysis measures the true-positive prediction rate (sensitivity/recall) *vs.* the false-positive prediction rate (1-specificity), while PR analysis calculates the rate of predicted associations that are truly positive (precision) *vs.* recall. An LR_{Max}^{cutoff} of 200, corresponding to a minimum 95% specificity, 80% sensitivity and 95% precision, is determined to define the high accuracy and coverage of associations. This ROC and PR-based cutoff results in 136,237 high-confidence associations among 12,181 genes (i.e., AssociatomeNet).

D. Enrichment analysis in terms of high-level relationships among biological themes

Given any two biological themes T_α (and the number n_α of its members) and T_β (and the number n_β of its members), the number of observed links between their members in the HumanNet is labeled as $m_{\alpha\beta}$ (excluding self-associations if genes annotated to both themes). The number of possible links when their members are randomly distributed can be approximatively modeled by Binomial distribution, resulting in the expected number of links $\overline{m}_{\alpha\beta}$ (**Eq. 7**) and its variance $\sigma^2_{\alpha\beta}$ (**Eq. 8**).

$$\overline{m}_{\alpha\beta} = n_\alpha n_\beta \frac{2M}{N(N-1)} \quad \text{[Eq. 7]}$$

$$\sigma^2_{\alpha\beta} = n_\alpha n_\beta \frac{2M}{N(N-1)}\left(1 - \frac{2M}{N(N-1)}\right) \quad \text{[Eq. 8]}$$

Where *M* and *N* represent the total number of links and the total number of genes in the HumanNet, respectively.

Thus, statistically significant deviations of observed links $m_{\alpha\beta}$ from its expected links $\overline{m}_{\alpha\beta}$ can be assessed by Binomial transformed Z-score $Z_{\alpha\beta}$ (**Eq. 9**). The larger the Z-score between two themes within the network, the more likely it is that their members topologically interact.

$$Z_{\alpha\beta} = \frac{m_{\alpha\beta} - \overline{m}_{\alpha\beta}}{\sigma_{\alpha\beta}} \quad \text{[Eq. 9]}$$

In particular, for one given biological theme T_α, the likelihood of the number $m_{\alpha\alpha}$ of interacting with each other in the network can be similarly assessed by $Z_{\alpha\alpha}$ (**Eq. 10-12**).

$$\overline{m}_{\alpha\alpha} = \frac{n_\alpha(n_\alpha - 1)}{2} \frac{2M}{N(N-1)} \quad \text{[Eq. 10]}$$

$$\sigma^2_{\alpha\alpha} = \frac{n_\alpha(n_\alpha - 1)}{2} \frac{2M}{N(N-1)}\left(1 - \frac{2M}{N(N-1)}\right) \quad \text{[Eq. 11]}$$

$$Z_{\alpha\alpha} = \frac{m_{\alpha\alpha} - \overline{m}_{\alpha\alpha}}{\sigma_{\alpha\alpha}} \quad \text{[Eq. 12]}$$

Where $\overline{m}_{\alpha\alpha}$ and $\sigma^2_{\alpha\alpha}$ respectively represent the expected number of links (**Eq. 10**) and the corresponding variance (**Eq. 11**) when the members of the biological theme T_α are randomly distributed in the network.

Taken together, Z-score matrix $Z_{\alpha\beta}$ encodes the network-oriented relationships among biological themes. After hierarchical clustering of the Z-score matrix, the resultant map provides a global view of such relationships. When simultaneously comparing multiple hypothesis tests, statistical significance of each relationship can be estimated by the method of false discovery rate (FDR) [17]. Briefly, a randomized network was first generated by randomly shuffling the biological theme-gene associations while keeping the adjacency matrix of the network unchanged. From this network, an expected Z-score matrix $Z_{\alpha\beta}{}^b$ was then similarly calculated to obtain a random distribution for observed Z-score matrix. By repeating randomization-calculation steps B (i.e., 1,000) times, the resulting random distributions of Z-score matrix $Z_{\alpha\beta}{}^b, b = 1, \cdots, B$ were used to estimate the FDR value by determining the number of relationships called significant (i.e., those with Z-scores no less than a specific value Z_t) and dividing by the median number of relationships falsely called significant (i.e., the median number of Z-scores among each of B Z-scores matrix, whose $Z_{\alpha\beta}{}^b$ satisfy: $Z_{\alpha\beta}{}^b \geq Z_t, b = 1, \cdots, B$).

RESULTS AND DISCUSSION

A. *Constructing a high-confidence gene network in human*

With the purpose of exploring interrelationships among biological themes in the context of the human network, we sought to construct a network containing gene associations of high accuracy while maintaining the high coverage to include as many human genes as possible. To such end, we not only considered direct associations of genes, as highlighted by human physical protein-protein interactions (i.e., InteractomeNet, the union of BIND, DIP, IntAct, Reactome and HPRD), but also applied a Bayesian supervised approach to integrate the indirect associations of genes (AssociatomeNet) which were predicted and transferred from heterogeneous data sets across organisms (**Figure 2**).

Integration using a maximum Bayesian model [5;6] allows data sets with dissimilar types (i.e., numerical and categorical) and redundant information (e.g., those multiple resources-curated functional gene associations in model organisms) to be combined into a common, most conservative estimator of maximum likelihood ratio (LR_{max}). We fist assembled four sources of predictive data sets, including shared molecular function annotations of GO (GOMF) [2], yeast functional gene network (YeastNet) [8], C.elegans functional gene network (WormNet) [9] and human functional linkages (STRING) [15]. Each predictive data set was then subjected to the estimation of how the predictive data set impacts the chance that gene pairs are associated. As shown in **Figure 2A**, these four resources of

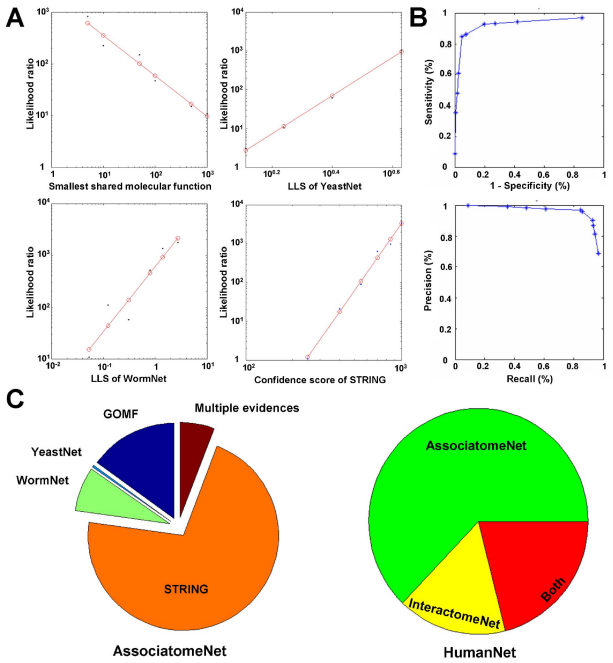

Figure 2. Construction and content of the human gene network (HumanNet). (A) Diverse predictive datasets contributing to the Bayesian supervised model of human gene associations (AssociatomeNet). Top left, 'The likelihood ratio of gene associations decreases with the increasing number of smallest shared molecular function. Top right, 'The likelihood ratio of associations transferred from functional network in yeast (YeastNet) positively correlates with increasing LLS. Bottom left, 'The humanized functional network of *C.elegans* shows positive correlations between LLS of WormNet and the likelihood ratios of human gene associations. Bottom right, 'Human functional linkage data (STRING) with high confidence score tends to link genes evaluated by gold-standard positive and negative gene associations. A blue point indicates the calculated likelihood ratio of the specific bin, while the red line indicates the fitted values using the log-linear regression model. Note that the axes are on a log scale. **(B)** Evaluation and determination of AssociatomeNet with high-confidence and high-coverage associations. ROC curve measures the true-positive prediction rate (sensitivity/recall) versus the false-positive prediction rate (1-specificity) as a function of the maximum likelihood ratio (Top panel), while PR curve the rate of predicted associations that are truly positive (precision) versus recall (Bottom panel). **(C)** The content of the constructed network. The pie charts illustrate the relative contributions of each dataset to the final network HumanNet.

data sets all exhibit strong predictive power to infer human gene associations. To predict a larger number of gene associations while avoiding the overestimation, we combined the data set-specific LR in the maximum Bayesian model, which only retains the maximum LR for a gene pair. Thus far, we yielded a complete network of a large number of gene associations with low confidence. To obtain the accurate, high-coverage gene network, we performed receiver operator characteristic (ROC) and precision-recall (PR) analyses (**Figure 2B**) according to overlaps of predicted associations with the true associations and the true non-associations under a series of LR_{Max}. Applying an ROC- and PR-based cutoff (i.e., LR_{Max}^{cutoff} of more than 200) to the complete network, we finally built a higher confidence network, corresponding to minimum 95% specificity, 80% sensitivity and 95% precision. This predictive network of gene associations (AssociatomeNet), comprising 136,237 high-confidence associations among 12,181 genes, was contributed differentially by the predictive data sets (the left pie chart of **Figure 2C**). About 71% of gene associations in AssociatomeNet are explained by STRING alone, 15% by GOMF alone, less than 1% by YeastNet alone, 8% by WormNet alone, and 6% by multiple datasets.

Unionizing the compiled human interactome network (i.e., InteractomeNet) and the predicted human associatome network (AssociatomeNet), we constructed a single high-confidence gene network in human (termed as HumanNet), covering 162,339 physical/functional associations among 13,863 human genes (~ 55% of the human protein-coding genes). As shown in the right pie chart of **Figure 2C**, AssociatomeNet alone, InteractomeNet alone and both of them account for 63%, 16% and 21% of HumanNet, respectively. This comprehensive high-confidence gene network of human offers an unprecedented basis to explore the high-level relationships among biological themes by examining gene-level topological associations of their members in the constructed human gene network.

B. Discovering the network-oriented connections among biological processes

Genes involved in the same biological process are expected to be topologically linked. Furthermore, two or more biological processes may be topologically related to achieve the same goals of cellular functions. We therefore took advantages of Panther classification [1], which contains 126 abbreviated categories of biological processes, to examine the human network for Binomial distribution-based enrichment of associations in which (i) both partners were involved in the same biological process (intra-process connections), (ii) and one partner involved in a biological process while the other one involved in another process (inter-process connections). As expected, we observed a strong enrichment for intra-process connections, as shown in diagonal of Z-score matrix (**Figure 3A**), indicative of a highly modular structure. When examining the inter-process connections, we observed that biological processes distinguish themselves by their abilities to interconnect others. As shown in upper (or lower) diagonal of Z-score matrix (**Figure 3A**), most processes are linked to a few other processes, whereas a few processes represent the hubs

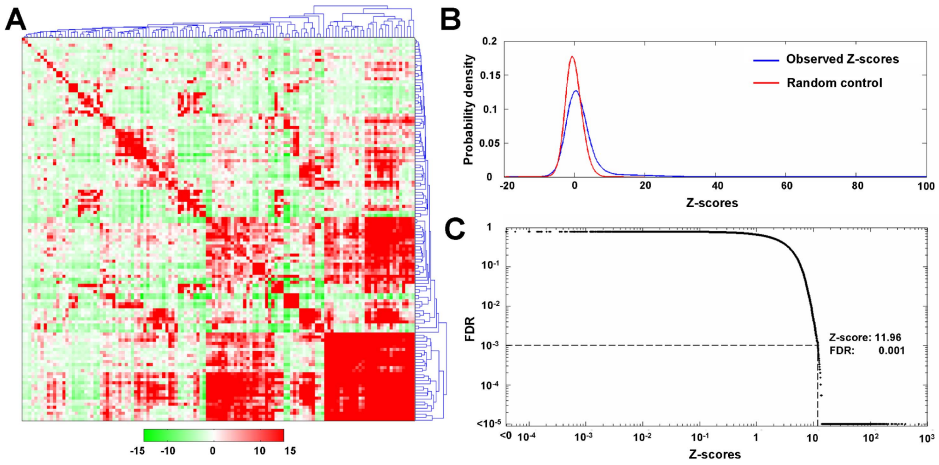

Figure 3. Binomial distribution-based enrichment analysis of the high-level relationships among Panther biological processes in the context of the network. (A) The Binomial transformed Z-score matrix corresponding to the network-oriented relationships among biological processes, together with the hierarchical tree indicating larger classes based on their interconnectedness. Categories of biological processes were derived from PANTHER database, which contains 126 abbreviated terms. The coded color of the matrix denotes the Z-score calculated as the normalized deviations of observed links in the network between/within biological processes from its expected links under Binomial distribution model. The larger the Z-score between/within biological processes, the more likely it is that their members topologically interact in the network. **(B)** Distribution of the observed Z-scores for all the possible relationships among biological processes (in blue) versus the distribution of the expected Z-scores generated from random network as control (in red). **(C)** The FDR values versus their respective Z-scores. The estimation of statistical significance of each relationship is inferred by comparing the observed distribution of Z-scores with their expected distribution from random network to control the FDR, calculated as the number of relationships called significant divided by the median number of relationships falsely called significant.

that are connected to a large number of distinct processes. To measure how the observed Z-score matrix of biological processes deviated from random control, we randomly shuffled the biological process-gene associations while keeping the adjacency matrix of the network unchanged. Compared with the random control, we found that distribution of the observed Z-scores was shifted toward higher values ($P=1.9 \times 10^{-17}$, Kolmogorov-Smirnov test), indicating that the observed universal enrichment of intra-process connections and preferential enrichment of inter-process connections are inherent features of biological processes (**Figure 3B**). We further utilized a method of false discovery rate (FDR) [17] to estimate the statistical significance of enrichment among these connections. **Figure 3C** shows a plot of the FDR values versus their Z-scores, wherein the indicated point by the dash lines show that 0.1% of those relationships with Z-scores more than 11.96 called significant turn out truly false (i.e., FDR=0.001). Here and thereafter, we used Z-score of 11.96 as the threshold to determine the statistical significance of relationships among biological themes.

Applying the FDR-based assessment of significant connections, we obtained the interconnectedness matrix of biological processes, which, together with the information revealed from the hierarchical clustering of Z-score matrix (**Figure 3A**), allows the reconfiguration of biological processes into eight larger classes (I-VIII) (**Figure 4A**). Although the reconfiguration of biological processes was generated independently of any priori knowledge about the process categories, the resulting classes tend to be functionally similar. The most prominent example is a tightly interconnected immunity-related cluster within Class VII (or called immunity clique). Except for complement-mediated immunity in Class I, all other immunity-related processes are exclusively grouped in Class VII. Likewise, most developmental processes remain together in Class V (development clique). As for Class VIII, these processes are tightly interconnected to form a clique, and are functionally associated to core cellular processes, such as cell cycle, apoptosis, signaling cascades, oncogenesis and stress response. Of note, the interconnectivity of developmental clique (and immunity clique) with the clique of core cellular processes is consistent with the essential roles of development system and immunity system in maintaining proper physiological functions of individual organism. In contrast, metabolism-related processes do not form a single class. They are distributed among multiple classes (i.e., Classes I, IV and VI), which share low interconnectedness. We hypothesized that high interconnectedness of core cellular events in Class VIII and immunity-related processes in Class VII arises from the presence of genes which overwhelmingly prefer to link many other biological processes, while low interconnectedness of metabolism-related processes is largely due to the absence of such process-linking genes. To quantify such differences, we first characterized the role of each gene based on its participation coefficient (pC) [18], which measures its tendency of connecting genes in all biological processes. According to

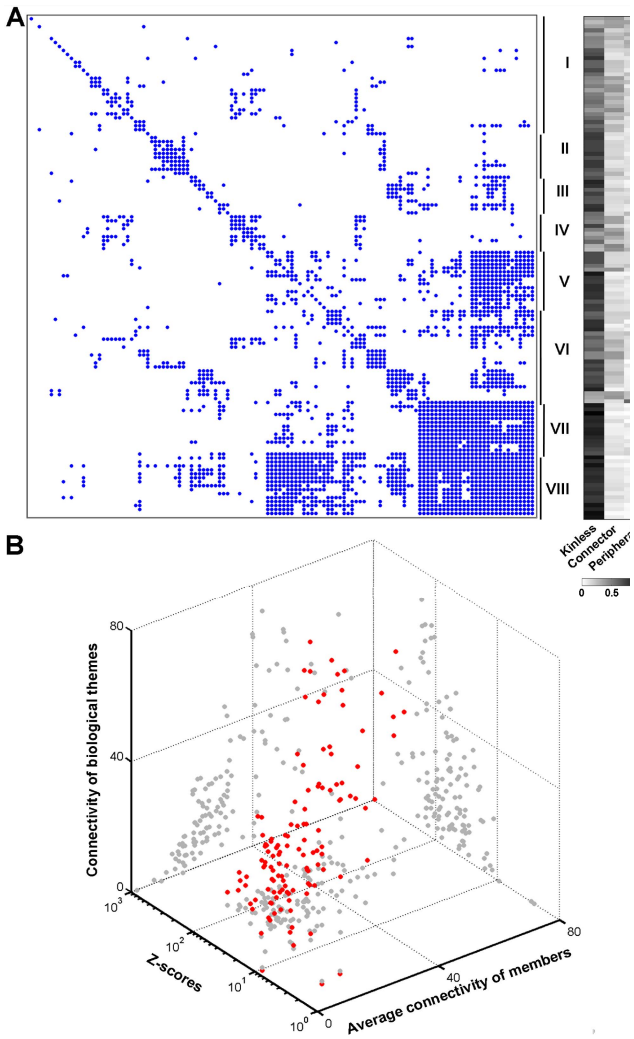

Figure 4. The network-driven interconnectedness among Panther biological processes and its contributing factors. (A) The interconnectedness matrix resulting from the Z-score matrix under the threshold of FDR<0.001 (corresponding to Z-score>11.96, see Figure 3) and the indicated larger classes (I-VIII) based on the interconnectedness. Also shown in the most right display is the gray-coded proportions of genes, for each biological process, which are classified as kinless (i.e., those genes with links homogeneously distributed among all biological processes), connector (i.e., those genes with many links to other biological processes) and peripheral (i.e., those genes with most links within their own biological process). **(B)** The contributing factors for the observed interconnectedness. Each point (in red) corresponds to a biological process, plotted on the 3-dimensional space spanned by the connectivity of biological processes (interconnectedness), within-process Z-score and the average connectivity of its individual members. Grayed points represent the orthogonal projection onto the corresponding 2-dimensional plane.

participation coefficient, we classified those genes into three different roles as follows: kinless genes with links homogeneously distributed among all biological processes

(pC＞0.857); connectors with many links to other biological processes (0.695＜pC≤0.857); peripheral genes with most links within their own biological process (pC≤0.695). For each biological process, we then calculated the proportion of genes classified into these three roles. As shown in the most right display of **Figure 4A**, those biological processes with high interconnectedness (e.g., Classes VII and VIII) are enriched with genes classified as kinless and are absent of connectors and peripheral genes. On the contrast, in those biological processes with low interconnectedness (e.g., Classes I and IV), the percentage of genes classified as connector or peripheral approximately equals that of genes classified as kinless. As a further examination of the contributing factors for the observed interconnectedness, we considered three process-centered quantities: the connectivity of biological processes (interconnectedness), within-process Z-score and the average connectivity of its members in the network. As shown in **Figure 4B**, the absence of correlation between interconnectedness and the within-process Z-scores (Left vertical plane), and the appearance of correlation between interconnectedness and the connectivity of their members (Right vertical plane) indicate that the observed interconnectedness of a biological process is contributed by the connectivity of their individual members to other processes rather than the modular constraints of their individual members.

C. Connections among regulatory profiles

Transcription factors (TFs) and microRNAs (miRNAs) are trans-acting regulotors of gene expression, both exerting their activities by binding to cis-regulatory elements (e.g., promoters regions and 3' untranslated regions) of their target genes in a combinatorial manner [19;20]. To search potential regulatory interplays within/between TFs and miRNAs, we performed enrichment analyses of network-context associations among predicted targets of TFs and miRNAs. We began by defining the TF regulatory profiles from the TRANSFAC database [3] and the miRNA regulatory profiles from miRBase database [4]. Then, we evaluated the enrichment of the links among targets of TFs and miRNAs mapped to the human networks under the binomial distribution-based model. After applying the hierarchical clustering of Z-score matrix and the FDR-based assessment of significant interplays, we obtained the interconnectedness matrix, displaying the significant connections among regulatory profiles. Compared to the prevalence of intra-processes connections (**Figure 4A**), the connections of within-regulatory profiles are rare, as shown in diagonal of interconnectedness matrix (**Figure 5A**). It is consistent with the fact that the topological structure of human gene network most reflects the functionality rather than the regulation. Moreover, the connections of between-regulatory profiles are biased in terms of TF and miRNA regulatory levels. Connections between TF-TF pairs or TF-miRNA pairs are more abundant compared to those between miRNA-miRNA pairs (**Figure 5A**). It suggests that TFs have the tendency to function in a combinatorial fashion with other TFs or miRNAs, whereas miRNAs are lack of such cooperation with other miRNAs. Since genes in human gene network have a power-law degree distribution with the degree exponent of 1.731, we wondered whether the same kind distribution holds as for the high-level regulatory profiles. As shown in **Figure 5B**, the degree distribution of regulatory profiles approximates a power law with the degree exponent of 0.728. It is well known that the smaller the value of degree exponent, the more important the role of the highly connected nodes (i.e., hubs) is in the network [21]. Therefore, the high-level network of regulatory profiles is more skewed to scale-free than the gene-level network. Neighborhood connectivity distribution is also a relevant property of a scale-free network. The neighborhood connectivity of a node measures the average of the neighborhood connectivity of the node. The Neighborhood connectivity distribution of regulatory profiles, following the power-law distribution with the exponent of -0.368, exhibits a decreasing function of the degree (**Figure 5C**). It indicates the prevalence of edges between low connected and highly connected nodes in the network of regulatory profiles. The network visualization of connections among regulatory profiles manifests these topological parameters (**Figure 5D**). Notably, the tightly interconnected TFs are cell cycle-relevant regulators. The clique of cell cycle-relevant regulators includes E2F1, E2F4, DP1, DP2, SP1, NRF1, NRF2, NFY and ATF. Although co-regulations of cell cycle have been long shown [22], our findings that targets of cell cycle-regulators are also topologically linked substantially expand the generality of functional synergism between these cell cycle-relevant regulators. Among the most connected miRNAs are hsa-miR-561, hsa-miR-495, hsa-miR-552, and hsa-miR-208b. Interestingly, these miRNAs are present both in TF-miRNA pairs and miRNA-miRNA pairs. Although their underlying regulatory programs remain unclear, preferential cooperation especially with the clique of cell cycle-relevant TFs further points to the importance of the intricate regulatory design in the cell cycle biosystem.

CONCLUSIONS

We have carried out a systems analysis to uncover the high-level relationships in various biological themes, ranging from biological processes to regulatory programs. The success depends on two factors. First, we have constructed a single human gene network, covering ~ 55% of the human protein-coding genes with the high-confident physical/functional associations. This network represents the most comprehensive resource with regard to interactions between human genes. The topological structure of the network inherently encodes the biology-operating information. Second, we have developed a model for decoding the underlying information. Principally, converting topological relationships of their members in the gene-level network into interconnectedness matrix allows the identifications of general properties of biological theme-level network. As demonstrated in this pilot study, our analytical framework combining with systematic information about biological themes provides a platform for testing known knowledge and, more importantly, generating new sound hypotheses. Connections among biological

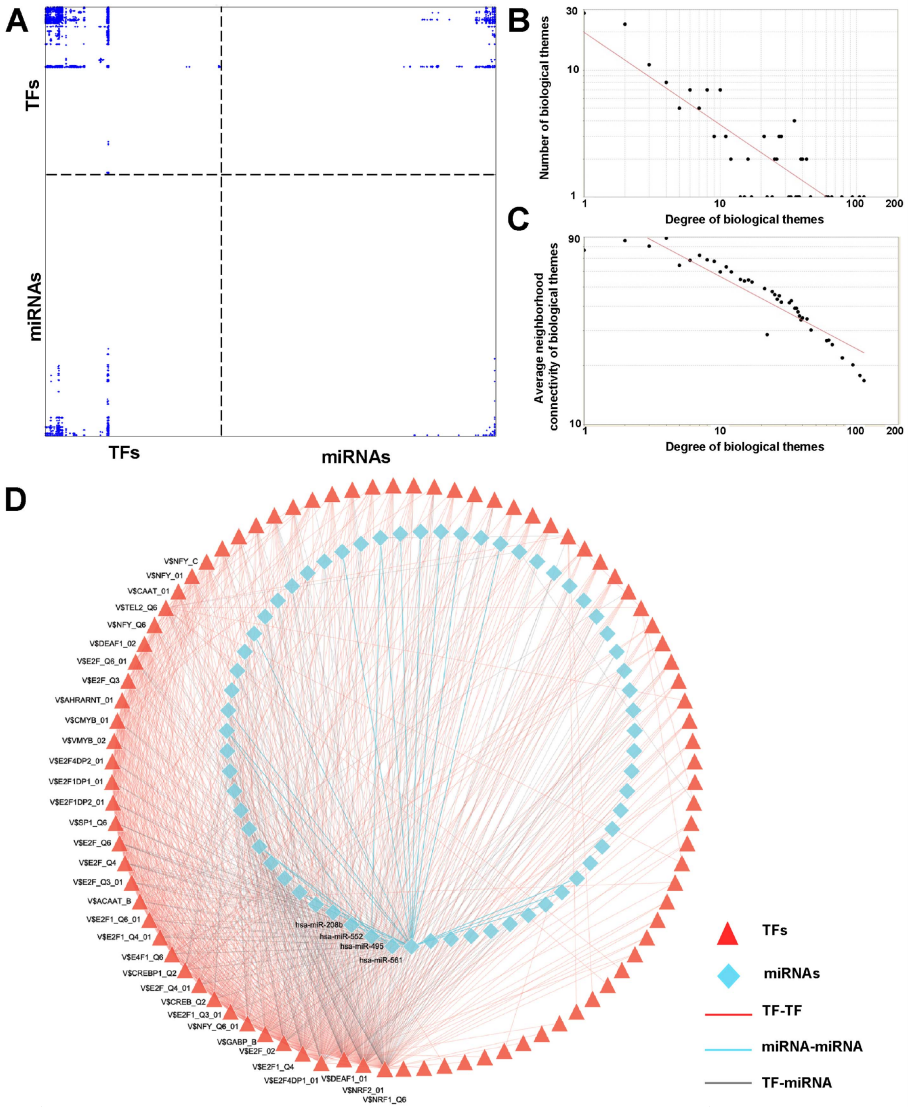

Figure 5. Connections among transcription factor (TF) and microRNA (miRNA) regulatory profiles. (A) The interconnectedness matrix, where each entry (in blue) represents the significant relationship (Z-score>11.96, corresponding to FDR<0.001) between/within regulatory profiles. TF regulatory profiles were identified by a position weight matrix (PWM)-based MATCH program applied to the putative promoter regions of human genes. Regulatory profiles of miRNAs were compiled from miRBase. **(B)** The degree distribution of regulatory profiles. The degree of a node (i.e., regulatory profiles) is the number of edges linked to the node. **(C)** Neighborhood connectivity distribution of regulatory profiles. The neighborhood connectivity of a node measures the average of the neighborhood connectivity of the node. **(D)** Network representations of TF-TF, miRNA-miRNA and TF-miRNA interconnections. The nodes sharing the same attributes (red-filled triangle for TFs, and cyan-filled diamond for miRNAs) are laid out in a separate circle, where the position of nodes in each circle is ordered according to the degree. Only nodes associated with the degree of more than 20 are shown with the identifiers of TFs/miRNAs.

processes show that the capability of a process interconnecting others differs greatly. For example, developmental processes (and immunity-related processes) highly interconnect together to form a clique, which then preferentially interconnects a clique of core cellular processes. Such hierarchical organizations are also observed in regulatory profiles, in which most interconnected miRNAs tend to interconnect a clique of cell cycle-relevant TFs.

ACKNOWLEDGMENT

This work was supported in part by the Knowledge Innovation Program of Chinese Academy of Sciences (J.Z.), Ministry of Science and Technology of China Grants (2006CB910405, 2006CB910700, 2006AA02Z302, 2006AA02Z332, 2007AA02Z335 and 2009CB825607), National Natural Science Foundation Grants (30730033, 30670436, 30600260 and 30570777)

REFERENCES

[1] Mi H, Guo N, Kejariwal A, Thomas PD, 'PANTHER version 6, 'protein sequence and function evolution data with expanded representation of biological pathways' Nucleic Acids Res. 2007, 35:D247-D252.

[2] The Gene Ontology project in 2008. Nucleic Acids Res. 2008, 36:D440-D444.

[3] Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E, 'TRANSFAC, 'transcriptional regulation, from patterns to profiles,' Nucleic Acids Res. 2003, 31:374-378.

[4] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ, 'miRBase, 'tools for microRNA genomics,' Nucleic Acids Res. 2008, 36:D154-D158.

[5] Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M, 'A Bayesian networks approach for predicting protein-protein interactions from genomic data,' Science 2003, 302:449-453.

[6] Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, Kalyana-Sundaram S, Ghosh D, Pandey A, Chinnaiyan AM, 'Probabilistic model of the human protein-protein interaction network,' Nat.Biotechnol. 2005, 23:951-959.

[7] Lee I, Date SV, Adai AT, Marcotte EM, 'A probabilistic functional network of yeast genes,' Science 2004, 306:1555-1558.

[8] Lee I, Li Z, Marcotte EM, 'An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae,' PLoS.ONE. 2007, 2:e988.

[9] Lee I, Lehner B, Crombie C, Wong W, Fraser AG, Marcotte EM, 'A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans,' Nat.Genet. 2008, 40:181-188.

[10] Bader GD, Betel D, Hogue CW, 'BIND, 'the Biomolecular Interaction Network Database,' Nucleic Acids Res. 2003, 31:248-250.

[11] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D, 'The Database of Interacting Proteins--2004 update,' Nucleic Acids Res. 2004, 32:D449-D451.

[12] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H, 'IntAct--open source resource for molecular interaction data,' Nucleic Acids Res. 2007, 35:D561-D565.

[13] Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L, 'Reactome, 'a knowledge base of biologic pathways and processes,' Genome Biol. 2007, 8:R39.

[14] Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, Menon S, Hanumanthu G, Gupta M, Upendran S, Gupta S, Mahesh M, Jacob B, Mathew P, Chatterjee P, Arun KS, Sharma S, Chandrika KN, Deshpande N, Palvankar K, Raghavnath R, Krishnakanth R, Karathia H, Rekha B, Nayak R, Vishnupriya G, Kumar HG, Nagini M, Kumar GS, Jose R, Deepthi P, Mohan SS, Gandhi TK, Harsha HC, Deshpande KS, Sarker M, Prasad TS, Pandey A, 'Human protein reference database--2006 update,' Nucleic Acids Res. 2006, 34:D411-D414.

[15] von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P, 'STRING 7--recent developments in the integration and prediction of protein interactions,' Nucleic Acids Res. 2007, 35:D358-D362.

[16] Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL, 'InParanoid 6, 'eukaryotic ortholog clusters with inparalogs,' Nucleic Acids Res. 2008, 36:D263-D266.

[17] Benjamini Y., Hochberg Y., 'Controlling the false discovery rate, 'a practical and powerful approach to multiple testing,' Journal of the Royal Statistical Society 1995,289-300.

[18] Guimera R, Mossa S, Turtschi A, Amaral LA, 'The worldwide air transportation network, 'Anomalous centrality, community structure, and cities' global roles,' Proc.Natl.Acad.Sci.U.S.A 2005, 102:7794-7799.

[19] Hobert O, 'Common logic of transcription factor and microRNA action,' Trends Biochem.Sci. 2004, 29:462-468.

[20] Chen K, Rajewsky N, 'The evolution of gene regulation by transcription factors and microRNAs,' Nat.Rev.Genet. 2007, 8:93-103.

[21] Barabasi AL, Oltvai ZN, 'Network biology, 'understanding the cell's functional organization,' Nat.Rev.Genet. 2004, 5:101-113.

[22] Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y, 'Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res. 2003, 13:773-780.