GMR
Genetics and Molecular Research
ISSN 16765680

Mining topological structures of protein-protein interaction networks for human brain-specific genes

W.J. Cui, X.J. Gong, H. Yu and X.C. Zhang

School of Computer Science and Technology, Tianjin University, Tianjin, China

Corresponding author: H. Yu
E-mail: yuhua@tju.edu.cn

Genet. Mol. Res. 14 (4): 12437-12445 (2015)
Received February 4, 2015
Accepted July 23, 2015
Published October 16, 2015
DOI http://dx.doi.org/10.4238/2015.October.16.10

ABSTRACT. Compared to other placental mammals, humans have unique thinking and cognitive abilities because of their developed cerebral cortex composed of billions of neurons and synaptic connections. As the primary effectors of the mechanisms of life, proteins and their interactions form the basis of cellular and molecular functions in the living body. In this paper, we developed a pipeline for mining topological structures, identifying functional modules, and analyzing their functions from publically available datasets. A human brain-specific protein-protein interaction network with 1482 nodes and 3105 edges was built using a MapReduce based shortest path algorithm. Within this, 7 functional cliques were identified using a network clustering method, 98 hub proteins were obtained by the calculation of betweenness and connectivity, and 5 closest relationship to clique connector proteins were recognized by the combination scores of topological distance and gene ontology similarity. Furthermore, we discovered functional modules interacting with TP53 protein, which involves several fragmented research study conclusions and might be an important clue for further *in vivo* or *in silico* experiments to confirm these associations.

Key words: Brain-specific genes; Functional cliques; GO similarity; Betweenness; Hub protein

INTRODUCTION

During the process of gene expression in the human brain, from transcription to translation, any critical change will lead to a change in the physiology and behavior of a neuron. Numerous research efforts have attempted to disclose the working mechanisms of the brain both from the viewpoint of physiological structure using imaging technologies (Amunts et al., 2014; Di Martino et al., 2014) and that of gene expression using genome-wide association studies (Hernandez et al., 2012; Medland et al., 2014).

System biology and biological interaction networks have led to new paradigms in these analyses. The discovery of the key changes in genes or proteins that regulate the responses of the brain to external stimuli or lead to neuropsychiatric disorders can accelerate the pace of gene and drug treatments. Protein interaction is crucial for building the networks in the brain that underlie these phenomena (Amunts et al., 2014). Neural networks are formed by the interconnection of specific neurons in the brain. The molecular mechanisms involved in creating these connections, however, are poorly understood (Zilles and Amunts, 2013).

In this paper, we developed a pipeline for mining topological structures, identifying functional modules, and analyzing their functions. A human brain-specific protein-protein interaction network (PPIN) with 1482 nodes and 3105 edges was built using a MapReduce-based shortest path algorithm. Within this, 7 functional modules were identified using the molecular complex detection (Mcode) algorithm (Bader and Hogue, 2003), 98 hub proteins were obtained by the calculation of betweenness and connectivity, and 5 closest relationship to clique connector (CRCC) proteins were recognized by the combination of scores of topological distance and gene ontology (GO) similarity. Furthermore, we discovered functional modules interacting with the TP53 protein, which involves several fragmented research study conclusions, and might be an import clue for further *in vivo* or *in silico* experiments to confirm or build upon these associations.

MATERIAL AND METHODS

PPI network construction for human brain-specific genes

Gene expression datasets such as those from the Genomics Institute of the Novartis Research Foundation, serial analysis of gene expression, and expressed sequence tags are very widely used as data sources for classifying housekeeping (HK) and tissue-specific (TS) genes. For a given gene, applying predefined thresholds of its expression levels usually allows its identification as HK or TS. However, because of the noise contained in expression datasets and human involvement in defining the thresholds, the reliability of the identifications is often not high. In this paper, we attempted to rectify this problem by obtaining brain-specific genes from two sources: scientific publications from PubMed and known tissue-specific databases such as TisGeD (Xiao et al., 2010) and TiGER (Liu et al., 2008). From this, we finally obtained a total of 56 brain tissue-specific genes.

Tissue-specific genes generally exhibit high expression in the corresponding tissue of living organisms. The external environment, which is primarily reflected in the transcription and expression of non-specific-tissue genes, exerts a considerable influence in the process of biological regulation. Proteins encoded by tissue-specific genes are instead greatly influenced by their associated proteins. To examine these for the brain, we developed a MapReduce-based Dijkstra shortest path algorithm (cloudSPA) to extract associated proteins from the whole human PPI data repository (http://string-db.org/), and then constructed a highly-associated PPIN.

A PPIN can be represented as an undirected weighted graph. Nodes represent proteins, edges denote interactions between proteins, and weights indicate the tightness of such interactions. To construct the brain-specific PPIN, firstly, we constructed a PPIN $G_0 \leq N_0, E_0, W_0 >$, in which N_0 represented the tissue specific protein set, E_0 was the interaction set of proteins in N_0, and W_0 was the associated strength set of interactions in E_0.

Secondly, on the basis of the G_0, the full human PPI was partitioned into several PPINs. Then, we queried the strong correlation paths between tissue-specific proteins in these networks using cloudSPA. All strong correlation paths were retrieved and added into G_0 to construct a PPIN $G_1 \leq N_1, E_1, W_1 >$, which still contained many redundancy connections. Thirdly, we discarded those non-shortest paths (non-strong-correlation paths) between tissue-specific proteins in G_1, and finally constructed a PPIN $G_2 \leq N_2, E_2, W_2 >$ (tissue specific PPIN), which was a PPIN with strong correlations and few redundancies of data.

Identifying date hub proteins

Hub proteins and functional modules are usually regarded as two key players in a PPIN. Joy et al. (2005) pointed out that the hub proteins might act as important links between these modules. Interactions between these topological structures play an important role in maintaining the steady state of life systems.

A date hub represents a protein with high connectivity and low co-expression. The terms degree and betweenness are used to describe a date hub. For every node v in a PPIN, degree centrality (DC) is defined as in Equation 1:

$$\mathrm{DC(v)} = \frac{(\mathrm{deg(v)})}{(\mathrm{N-1})} \qquad \text{(Equation 1)}$$

where *deg(v)* = $\{e| e\in E\ \Lambda v\in V\}$ is the number of links for node v.

Betweenness assesses the number of shortest paths passing through a given node. The concept derives from the analysis of social networks. A node with high betweenness centrality is more likely to be located on the shortest paths between multiple node pairs in the network and therefore more information should be passed through it. This means that betweenness centrality will easily lead to congestion and be more likely to become the bottleneck of a network. For every node v, betweenness centrality is defined as in Equation 2:

$$\mathrm{BC(v)} = \Sigma_{s,t\in V}\frac{\sigma(\mathrm{s,t|v})}{\sigma(\mathrm{s,t})} \qquad \text{(Equation 2)}$$

where *σ(s, t)* is the number of shortest paths from s to t; and *σ(s, t│ v)* is the number of shortest paths passing through v from s to t.

Essential genes are those indispensable for the survival of an organism, and their functions are therefore considered a foundation of life. Date hub proteins are identified by the proportion of proteins encoded by essential genes.

Proteins in a PPIN are divided into four groups: high betweenness, high connectivity; low betweenness, high connectivity; high betweenness, low connectivity, and low betweenness; low connectivity.

Here, we used a threshold α to describe whether a protein was high connectivity and low co-expression (high betweenness). The best α is chosen when the proportion of proteins encoded by essential genes is the largest, as shown in Figure 1.

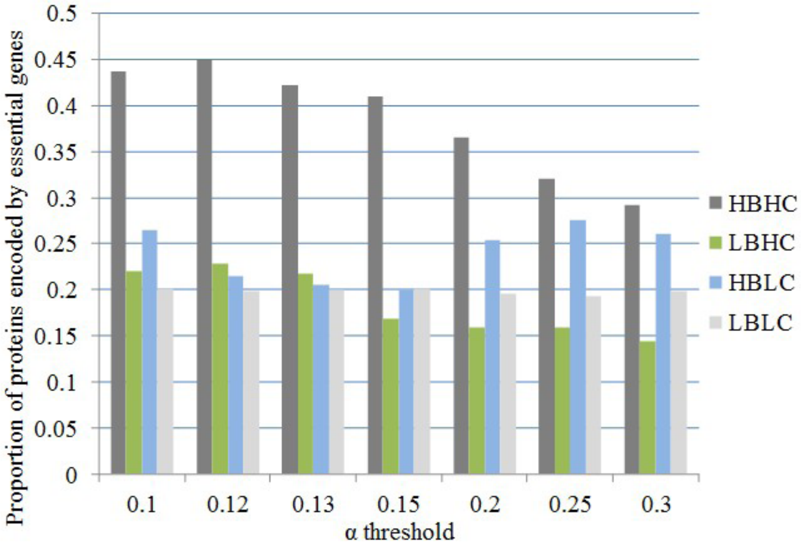


Figure 1. Identification of date hub proteins. The y-axis represents the proportion of proteins encoded by essential genes and the x-axis reflects the threshold, when α = 0.12, that the proportion of proteins encoded by essential genes is the largest. HBHC, high betweenness, high connectivity; LBHC, low betweenness, high connectivity; HBLC, high betweenness, low connectivity; LBLC, low betweenness, low connectivity.

Mining functional modules

Functional modules consist of proteins that participate in a particular cellular process while binding each other at different times and places (Schwikowski et al., 2000; Spirin and Mirny, 2003). Proteins inside a module have strong functional relevance. Cellular functions, such as signal transmission, are carried out by “modules” made up of many species of interacting molecules (Hartwell et al., 1999). Identifying functional modules is similar to detecting communities (clusters) in a network. To date, there are many clustering algorithms developed to find such groups; for instance, Mcode (Bader and Hogue, 2003), highly connected subgraph (Hartuv and Shamir, 1999), and restricted neighborhood search clustering (King et al., 2004). We performed the clustering analysis using above methods, and carried out functional enrichment analysis on the results. Only the results that were generated using the Mcode algorithm demonstrated rich GO sets. Thus, we adopted the Mcode algorithm to identify the functional modules in the PPIN.

Identification of CRCC proteins

Date hubs are considered to connect different modules. To obtain a better understanding of the biological function of date hub proteins as key connectors in the PPIN, we combined the topological distance and the GO similarity to measure the relationships as shown in Equation 3:

$$\mathrm{ProS}(\mathrm{V},\mathrm{C}) = \frac{\mathrm{TDS}(\mathrm{V},\mathrm{C}) * (1 - \mathrm{GSS}(\mathrm{V},\mathrm{C}))}{\mathrm{AvgTDS}(\mathrm{C}) * (1 - \mathrm{AvgGSS}(\mathrm{C}))} \qquad \text{(Equation 3)}$$

where *TDS(V, C)* is the score of topological distance, indicating the shortest distance from the date hub to its corresponding modules. *GSS(V, C)* is the score of GO similarity. *AvgTDS(C)* and *AvgGSS(C)* are the average topological and average GO similarity scores, respectively.

RESULTS

PPIN of human brain-specific genes

The final PPIN of human brain-specific genes is shown in Figure 2. It consists of 1482 nodes and 3105 edges. The features of its topological properties are shown in Table 1. From the table, we can observe that the PPIN follows the law of a small-world network (Telesford et al., 2011). The values of path length, centralization, and density indicate that the whole network is quite sparse.

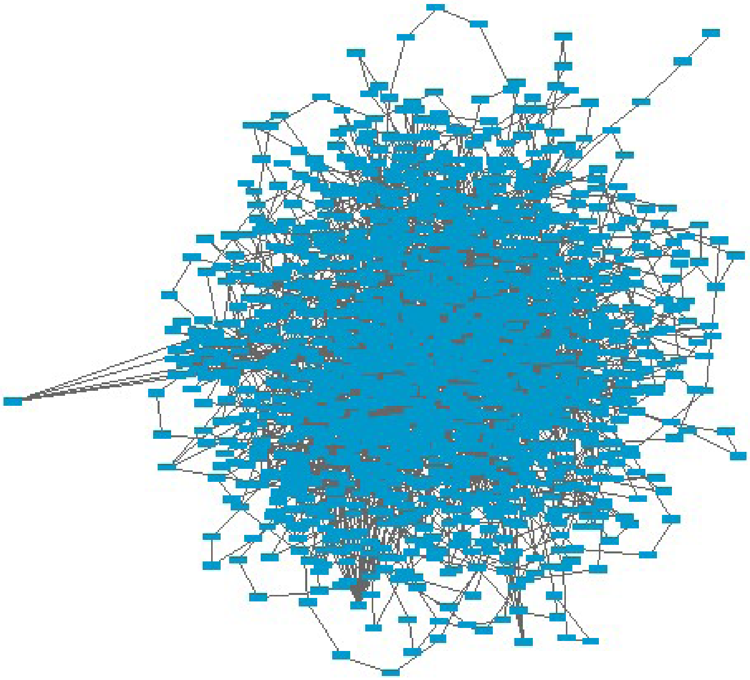


Figure 2. PPIN of human brain-specific genes.

Table 1. Topological properties of the PPIN.

Simple parameters	Value	Simple parameters	Value
Clustering coefficient	0.108	Number of nodes	1482
Connected components	1	Network density	0.003
Network diameter	11	Network heterogeneity	1.654
Network radius	6	Number of edges	3105
Network centralization	0.131	Number of self-loops	0
Characteristic path length	4.390	Avg. number of neighbors	4.192

PPIN = protein-protein interaction network.

Functional modules

A total of 7 functional modules were identified (Table 2). We performed functional enrichment analysis on those clusters, and all of them matched rich GO terms. For example, the functions of cluster 1 primarily included metabolic processes such as the Wnt signaling pathway, TGF-β signaling pathway, and ubiquitin mediated proteolysis. Functions of cluster 2 primarily consisted of responses to external stimulation including the B cell receptor, Toll-like receptor, and T cell receptor signaling pathways. Cluster 3 was found to mainly regulate biological processes; its functions included the gonadotropin-releasing hormone (GnRH) signaling pathway and the calcium signaling pathway. The function of cluster 4 was the regulation of cell protein metabolism. Cluster 5 played a significant role in immune response and immune regulation. The functions of cluster 6 primarily included RNA splicing, nuclear mRNA splicing via spliceosomes, and RNA splicing via transesterification reactions. The functions of cluster 7 mainly included certain fundamental cellular functions such as forebrain development, lymphocyte activation, adherent junctions, focal adhesion, synaptic transmission, and positive regulation of cell communication.

Table 2. Identification of seven functional modules.

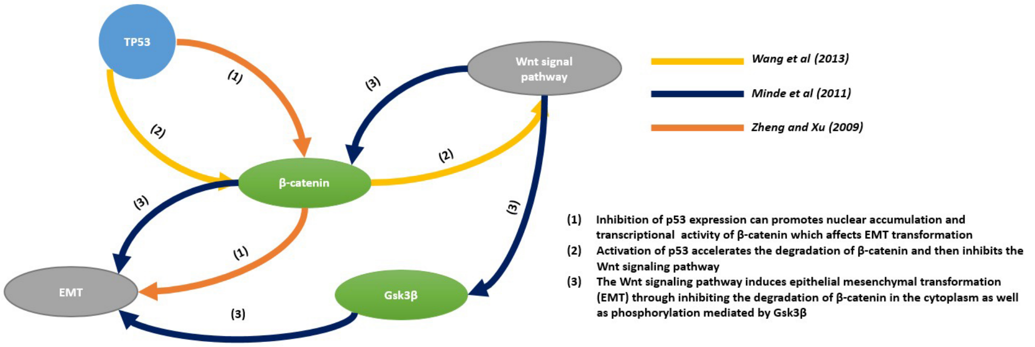
	Protein nodes	Genes
Cluster 1	ENSP00000251547, ENSP00000216225, ENSP00000347834, ENSP00000344866, ENSP00000224862, ENSP00000380256, ENSP00000281708, ENSP00000326804, ENSP00000265094, ENSP00000276326	*FBXO44*, *RBX1*, *FBXL3*, *FBXL5*, *FBXL15*, *CCNF*, *FBXW7*, *CUL1*, *FBXW11*, *FBXO25*
Cluster 2	ENSP00000274335, ENSP00000335657, ENSP00000306245, ENSP00000303452, ENSP00000263967	*PIK3R1*, *CCK*, *FOS*, *TRH*, *PIK3CA*
Cluster 3	ENSP00000286355, ENSP00000311405, ENSP00000334051, ENSP00000401548	*ADCY8*, *ADCY6*, *GNAL*, *TCF19*
Cluster 4	ENSP00000351885, ENSP00000280892, ENSP00000398350	*PPP2R4*, *EIF4E*, *C11orf68*
Cluster 5	ENSP00000258743, ENSP00000402956, ENSP00000264832	*IL6*, *HLA-B*, *ICAM1*
Cluster 6	ENSP00000361162, ENSP00000318861, ENSP00000333001	*TOE1*, *SF3B2*, *RBM8A*
Cluster 7	ENSP00000344456, ENSP00000344818, ENSP00000357656	*CTNNB1*, *UBC*, *FYN*

CRCC proteins

We also identified 5 CRCC proteins, as shown in Table 3. Among these, we discovered an interesting connection with TP53 proteins, which connect several fragmented research study conclusions (Figure 3).

Table 3. Relationships between the CRCC protein and its two closest partner groups.

CRCC (protein ID)	Gene	1st partner (ProS score)	2nd partner (ProS score)
ENSP00000269305	*TP53*	Cluster 1 (0.3227)	Cluster 7 (0.144205)
ENSP00000341551	*SMAD4*	Cluster 2 (0.471616)	Cluster 7 (0.122718)
ENSP00000251453	*RPS16*	Cluster 2 (0.948999)	Cluster 3 (0.840729)
ENSP00000256442	*CCNB1*	Cluster 6 (0.408316)	Cluster 7 (0.134237)
ENSP00000332643	*NDN*	Cluster 3 (0.550684)	Cluster 7 (0.247597)



(1) Inhibition of p53 expression can promotes nuclear accumulation and transcriptional activity of β-catenin which affects EMT transformation
(2) Activation of p53 accelerates the degradation of β-catenin and then inhibits the Wnt signaling pathway
(3) The Wnt signaling pathway induces epithelial mesenchymal transformation (EMT) through inhibiting the degradation of β-catenin in the cytoplasm as well as phosphorylation mediated by Gsk3β

Figure 3. Regulatory effects of TP53 between cluster 1 and cluster 7. Different color arrows represent different conclusions (1-3) proposed in different papers as cited. Gray represents the function of the clusters. Blue indicates the hub protein.

The protein encoded by TP53 (p53) acts as a tumor suppressor. The nuclear accumulation and transcriptional activity of β-catenin has been shown to be modulated by p53, and the inhibition of p53 expression can promote nuclear accumulation and the transcriptional activity of β-catenin which affects the epithelial-mesenchymal transition (EMT) (Wang et al., 2013). Activation of p53 accelerates the degradation of β-catenin and then inhibits the Wnt signaling pathway (Zheng and Xu, 2009). In turn, the Wnt signaling pathway induces EMT through inhibiting the degradation of β-catenin in the cytoplasm as well as phosphorylation mediated by glycogen synthase kinase (Gsk3β) (Giles et al., 2003). As mentioned previously, functions of cluster 1 include the Wnt signaling pathway and functions of cluster 7 include some fundamental cellular functions such as adherent junctions that are related to EMT. Obviously, the expression of TP53 affects the content of β-catenin and further regulates these two functional modules.

Along the same lines, the *CCNB1* gene encodes a cell cycle protein. Cluster 6 plays a role in transcription and cluster 7 primarily comprises some basic cellular functions such as cell cycle progression, gene expression regulation and transcriptional regulation. Thus, CCNB1 is clearly closely related to these two modules.

The protein encoded by *SMAD4* can serve as a tumor suppressor. SMAD directly participates in signal transduction of the TGF-β superfamily and is an important mediator of TGF-β superfamily members for their participation in life activities (Li et al., 2011). TAK1 was originally identified as a mitogen-activated protein kinase kinase kinase (MAP3K) activated by TGF-β (Yamaguchi et al., 1995) and was characterized as a central player in multiple immune and inflammatory signaling pathways including cytokine receptors, Toll-like receptor, T-cell receptor, and B-cell receptor-mediated signaling (Sato et al., 2005; Schuman et al., 2009; Shinohara and Kurosaki, 2009; Chen, 2012; Sakurai, 2012) (cluster 2). On the other hand, research by Scheel et al. (2011) shows that TGF-β is able to induce EMT, which is closely related to the function of adherent junctions and focal adhesion (cluster 7).

NDN is an imprinted gene and one of the many actions of Nhlh2 might be through necdin, a downstream target of Nhlh2 that was shown to augment *GnRH* gene transcription by interrupting Msx repression (Miller et al., 2009) (cluster 3). There are no obvious research results that illustrate the relationship between *NDN* and cluster 7. However, cluster 7 mainly comprises some basic cellular functions and many functions of *NDN* are inseparable from synaptic transmission, the positive regulation of cell communication, and cell mobility (cluster 7).

The *RPS16* gene encodes a ribosomal protein that is the main component of the ribosome, which is essential in protein biosynthesis. Although we were unable to identify evidence to verify the relationship between RPS16 and its corresponding clusters, we might predict that it plays a role in the response to external stimuli and regulation of cell biological processes.

DISCUSSION

Hub proteins act as important links between functional modules. We can both verify some relevant published conclusions and predict some as-yet unknown biological processes. Meanwhile, some relevant conclusions can be integrated to describe a complete biological process that is significant for understanding PPIN robustness and integrity. With the rapid development of science and technology, increasing numbers of biological experiments can be completed efficiently. If conclusions in these papers cannot be efficiently utilized and incorporated, that would be a great loss. Therefore, the question of how to integrate them remains a challenging problem. Our discovery might provide a clue to achieving this goal.

On the other hand, we note that the results of the enrichment analysis focus primarily on basic functions and rarely contain brain tissue-specific functions. One reason for this might be that the majority of brain function modules play a role in basic biological processes. Another reason could be the small set of data currently known to be related because we are not very certain regarding the mechanisms of cognition and thinking, and we intend to examine this issue further in future studies.

Conflicts of interest

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

Research partly supported by the Natural Science Funding of China under grant #61170177, the National Basic Research Program of China under grant #2013CB32930X, and Innovation Funding of Tianjin University.

REFERENCES

Amunts K, Hawrylycz MJ, Van Essen DC, van Horn JD, et al. (2014). Interoperable atlases of the human brain. *Neuroimage* 99: 525-532.

Bader GD and Hogue CW (2003). An automated method for finding molecular complexes in large protein interaction networks.

BMC Bioinformatics 4: 2.

Chen ZJ (2012). Ubiquitination in signaling to and activation of IKK. *Immunol. Rev.* 246: 95-106.

Di Martino A, Yan CG, Li Q, Denio E, et al. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* 19: 659-667.

Giles RH, van Es JH and Clevers H (2003). Caught up in a Wnt storm: Wnt signaling in cancer. *Biochim. Biophys. Acta* 1653: 1-24.

Hartuv E and Shamir R (2000). A clustering algorithm based on graph connectivity. *Inform. Process. Lett.* 76: 175-181.

Hartwell LH, Hopfield JJ, Leibler S and Murray AW (1999). From molecular to modular cell biology. *Nature* 402: C47-52.

Hernandez DG, Nalls MA, Moore M, Chong S, et al. (2012). Integration of GWAS SNPs and tissue specific expression profiling reveal discrete eQTLs for human traits in blood and brain. *Neurobiol. Dis.* 47: 20-28.

Joy MP, Brock A, Ingber DE and Huang S (2005). High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* 2005: 96-103.

King AD, Przulj N and Jurisica I (2004). Protein complex prediction via cost-based clustering. *Bioinformatics* 20: 3013-3020.

Li F, Lan Y, Wang Y, Wang J, et al. (2011). Endothelial Smad4 maintains cerebrovascular integrity by activating N-cadherin through cooperation with Notch. *Dev. Cell* 20: 291-302.

Liu X, Yu X, Zack DJ, Zhu H, et al. (2008). TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics* 9: 271.

Medland SE, Jahanshad N, Neale BM and Thompson PM (2014). Whole-genome analyses of whole-brain data: working within an expanded search space. *Nat. Neurosci.* 17: 791-800.

Miller NL, Wevrick R and Mellon PL (2009). Necdin, a Prader-Willi syndrome candidate gene, regulates gonadotropin-releasing hormone neurons during development. *Hum. Mol. Genet.* 18: 248-260.

Minde DP, Anvarian Z, Rüdiger SG, et al. (2011). Messing up disorder: how do missense mutations in the tumor suppressor protein APC lead to cancer. *Mol. Cancer* 10: 10.1186.

Sakurai H (2012). Targeting of TAK1 in inflammatory disorders and cancer. *Trends Pharmacol. Sci.* 33: 522-530.

Sato S, Sanjo H, Takeda K, Ninomiya-Tsuji J, et al. (2005). Essential function for the dinase TAK1 in innate and adaptive immune responses. *Nat. Immunol.* 6: 1087-1095.

Scheel C, Eaton EN, Li SH, Chaffer CL, et al. (2011). Paracrine and autocrine signals induce and maintain mesenchymal and stem cell states in the breast. *Cell* 145: 926-940.

Schuman J, Chen YH, Podd A, Yu M, et al. (2009). A critical role of TAK1 in B-cell receptor-mediated nuclear factor kappaB activation. *Blood* 113: 4566-4574.

Schwikowski B, Uetz P and Fields S (2000). A network of protein-protein interactions in yeast. *Nat. Biotechnol.* 18: 1257-1261.

Shinohara H and Kurosaki T (2009). Comprehending the complex connection between PKCbeta, TAK1, and IKK in BCR signaling. *Immunol. Rev.* 232: 300-318.

Spirin V and Mirny LA (2003). Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. U. S. A.* 100: 12123-12128.

Telesford QK, Joyce KE, Hayasaka S, Burdette JH, et al. (2011). The ubiquity of small-world networks. *Brain Connect.* 1: 367-375.

Wang Z, Jiang Y, Guan D, Li J, et al. (2013). Critical roles of p53 in epithelial-mesenchymal transition and metastasis of hepatocellular carcinoma cells. *PLoS One* 8: e72846.

Xiao SJ, Zhang C, Zou Q and Ji ZL (2010). TiSGeD: a database for tissue-specific genes. *Bioinformatics* 26: 1273-1275.

Yamaguchi K, Shirakabe K, Shibuya H, Irie K, et al. (1995). Identification of a member of the MAPKKK family as a potential mediator of TGF-beta signal transduction. *Science* 270: 2008-2011.

Zheng Q and Xu LH (2009). Influential factors of B-catenin in the classic Wnt/b-catenin signaling pathway. *Chin. J. Cell Biol.* 31: 183-190.

Zilles K and Amunts K (2013). Individual variability is not noise. *Trends Cogn. Sci.* 17: 153-155.