ANIMAL CELLS AND SYSTEMS, 2017
VOL. 21, NO. 2, 77–83
http://dx.doi.org/10.1080/19768354.2017.1284156

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

From sequencing data to gene functions: co-functional network approaches

Jung Eun Shim*, Tak Lee* and Insuk Lee

Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, Seoul, Korea

ABSTRACT
Advanced high-throughput sequencing technology accumulated massive amount of genomics and transcriptomics data in the public databases. Due to the high technical accessibility, DNA and RNA sequencing have huge potential for the study of gene functions in most species including animals and crops. A proven analytic platform to convert sequencing data to gene functional information is co-functional network. Because all genes exert their functions through interactions with others, network analysis is a legitimate way to study gene functions. The workflow of network-based functional study is composed of three steps: (i) inferencing co-functional links, (ii) evaluating and integrating the links into genome-scale networks, and (iii) generating functional hypotheses from the networks. Co-functional links can be inferred from DNA sequencing data by using phylogenetic profiling, gene neighborhood, domain profiling, associalogs, and co-expression analysis from RNA sequencing data. The inferred links are then evaluated and integrated into a genome-scale network with aid from gold-standard co-functional links. Functional hypotheses can be generated from the network based on (i) network connectivity, (ii) network propagation, and (iii) subnetwork analysis. The functional analysis pipeline described here requires only sequencing data which can be readily available for most species by next-generation sequencing technology. Therefore, co-functional networks will greatly potentiate the use of the sequencing data for the study of genetics in any cellular organism.

ARTICLE HISTORY
Received 14 January 2017
Accepted 15 January 2017

KEYWORDS
Sequencing data; gene functions; co-functional networks

Introduction

Revolutionary advances in high-throughput sequencing technology enabled genomics and transcriptomics approaches to the study of gene functions in any living organism. As of December 2016, Genomes Online Database (GOLD) (Mukherjee et al. 2017) reported over 80,000 cellular organisms with sequenced genomes including approximately 3000 animal and plant species. One of the key goals of sequencing projects is to identify functions of genes in an organism. Sequencing data alone may suggest gene functions to some extent based on sequence homology between evolutionarily conserved genes. However, abundant sequencing data for diverse species and the wide variety of biological contexts may allow the extraction of higher-order functional information. Since all genes exert their functions through their molecular interactions, exploiting the map of functional associations between genes (i.e. co-functional network) will facilitate identification of gene functions. There are several steps in the network-based workflow to identify gene functions from sequencing data: (i) inferencing co-functional links, (ii) evaluating and integrating the inferred links into genome-scale networks, and (iii) generating functional hypotheses from the networks. Here, we review recent progress in network-based approaches to the study of gene functions in the respect of those three work steps. Since sequencing technology is accessible for most organisms, the workflow described here will be applicable to most organisms as well.

Inference of co-functional links from sequencing data

Based on co-inheritance of ancestral genes

The genes that are conserved or lost in a similar pattern across species can be due to their functional relevance. The pattern of gene inheritance during speciation can be profiled by homology search for reference species with fully sequenced genomes, called *phylogenetic profiling* (PG) (Figure 1, Inference step A) (Pellegrini et al. 1999; Kensche et al. 2008). Thus, a phylogenetic profile for each gene is a vector of presence or absence of homologs across reference species genomes. The profile can be based on binary score, indicating presence or absence of homology, or based on significance scores derived

CONTACT Insuk Lee insuklee@yonsei.ac.kr
*These authors contributed equally.

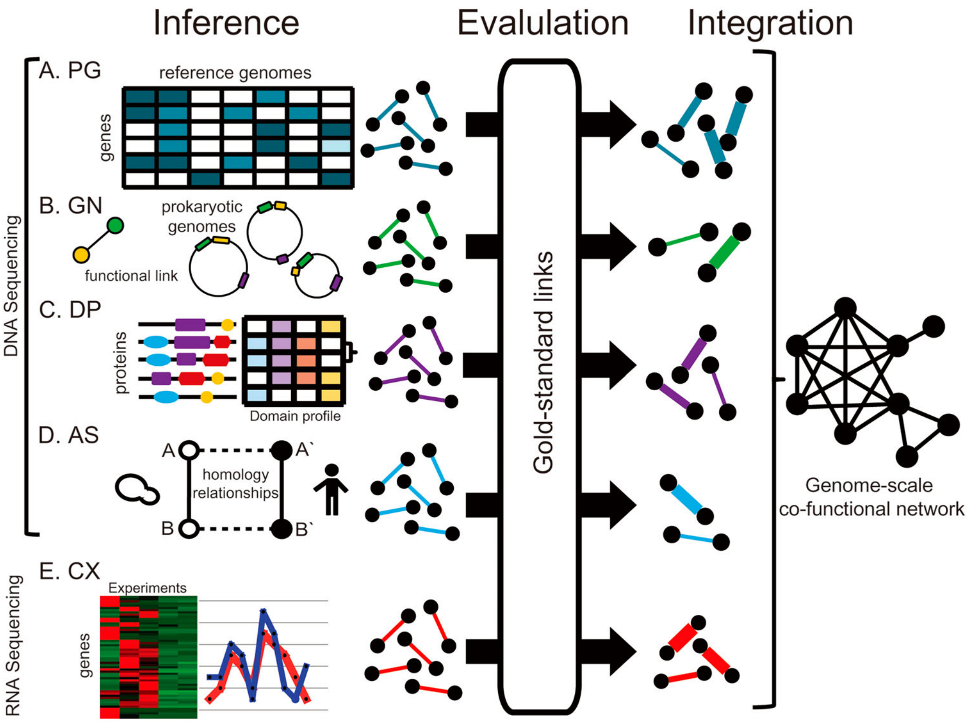

Figure 1. From sequencing data to co-functional networks. Functional links between genes can be inferred by (A) phylogenetic profiling (PG), (B) gene neighborhood (GN), (C) domain profiling (DP), (D) associalogs (AS) using DNA sequencing data, and by (E) co-expression (CX) analysis using RNA sequencing data. All inferred links are evaluated by gold-standard co-functional links derived from pathway annotation databases. The inferred links are scored for likelihood (represented by edge thickness, in which the thicker edge indicates higher likelihood of functional association), and then integrated into a genome-scale co-functional network.

from sequence alignment software, such as Basic Local Alignment Search Tool (McGinnis & Madden 2004). Similarity of phylogenetic profiles between two genes can be measured by various metrics, and mutual information (MI) scores generally give high correlation between phylogenetic profile similarity and degree of functional coupling between genes (Shin & Lee 2017). The PG method has been used to successfully infer functional association between genes in bacterial species, but not in higher eukaryotes such as animals and plants. Recently, we found that the PG method can be more effective when profile similarity is measured within each of three domains of life: Archaea, Bacteria, and Eukaryota (Shin & Lee 2015). We also demonstrated that, with domain-specific PG, the size of the inferred human gene networks increased as additional reference species genomes were used. This study suggests that better understanding of speciation and functional evolution may improve the effectiveness of PG even further in the future.

Based on conserved genomic neighborhood relationship in bacterial genomes

In bacterial genomes, genes for operating the same pathway are frequently encoded as co-transcriptional gene clusters, called *operons*. Functional association between genes, therefore, can be inferred based on neighborhood relationships between genes in bacterial genomes (Dandekar et al. 1998). The principle of the *gene neighborhood* (GN) method (Figure 1, Inference step B) can be applied not only for bacterial genes but also for eukaryotic genes with bacterial orthologs. If two genes of a eukaryotic organism have counterpart orthologous genes that tend to be in proximity to each other in bacterial genomes, they are likely to be involved in similar processes. The degree of GN can be measured by either distance or probability of being neighbors in bacterial genomes. Recently, we showed that the two different measures of GN are complementary so that their integration improves the quality of co-functional networks (Shin et al. 2014).

Based on similar domain compositions

Protein domains are considered as structural, functional, and evolutionary units of proteins. Therefore, functional associations between protein coding genes can be inferred based on domain-level information of each proteins. For example, extrapolation of domain–domain interactions (DDIs) from known protein–protein interactions (PPIs) can be used to identify functional associations between protein coding genes (Sprinzak & Margalit 2001; Deng et al. 2002). Many computational methods have been developed to identify DDIs from PPIs and to infer new PPIs from the DDIs, which are now available from meta-databases (Yellaboina et al. 2011). However, these methods require reference PPIs or known DDIs to identify new functional associations between coding genes. Recently, we proposed *domain profiling* (DP) (Figure 1, Inference step C), a domain-based method to infer functional links that requires only domain annotations for each protein coding genes (Shim & Lee 2016). In this method, the domain composition of each protein coding gene is represented as a domain profile, which is a vector of presence or absence of each domain of a comprehensive domain database, Interpro (Mitchell et al. 2015). Next, functional associations between protein coding genes are measured based on the similarity between domain profiles. There are various metrics to measure profile similarity. We developed a new metric, a weighted version of MI, and found that this metric outperformed other popular metrics including traditional MI (Shim & Lee 2016).

Based on co-functional links between orthologous genes

Functions can be evolutionarily conserved not only at the gene-level but also at the network-level. Evolutionarily conserved interactions between proteins whose homologous proteins in other organisms also interact are called *interologs* (Walhout et al. 2000). This network inference method has been particularly useful for species with not much experimental data available. Although interologs can be used for any sequenced species, the resulting networks generally have limited size because many functional associations are not based on physical interactions between proteins. Furthermore, the majority of the known PPIs were identified from only few species: *Saccharomyces cerevisiae* (budding yeast), *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fruit fly), *Arabidopsis thaliana,* and human. Recently, we applied the evolutionarily conserved relationship between genes to functional association, *associalogs* (AS) (Figure 1, Inference step D), which are evolutionarily conserved functional associations between genes whose orthologs are also functionally associated in other organisms (Kim et al. 2013). There are substantially more functional associations that can be identified by AS than those by interologs.

Based on co-expressions across various contexts

RNA sequencing data are also widely used for inferring co-functional links between genes. Under the hypothesis that two genes that operate same cellular processes are likely to show similar expression patterns across experimental conditions, *co-expression* (CX) (Figure 1, Inference step E) can be used to infer functional associations. The degree of CX between genes can be measured by using various metrics such as Pearson correlation coefficient, Spearman's correlation coefficient, and MI (Song et al. 2012). There is no universally best measure for the correlation. Rather, selecting the one that fits the given data would be a better strategy. CX analysis has two major advantages in inferring functional associations between genes. First, the large amount of publicly available transcriptomics data based on both microarray and RNA sequencing allows to identify many functional links. There are central public repository databases such as Gene Expression Omnibus (Edgar et al. 2002), ArrayExpress (Parkinson et al. 2005), and The Sequence Read Archive (Kodama et al. 2012). These databases provide downloadable transcriptomics data which were submitted by individual researchers. The amount of data in these databases is sharply rising recently as RNA sequencing popularity is increasing. RNA sequencing has several merits over microarray: platform-independence, higher sensitivity, and it is more robust against technical noises such as probe cross-hybridization. Therefore, the rapid growth of RNA sequencing data will potentiate CX analysis for the study of gene functions. Second, CX links can provide context-associated functional models. Transcriptomics data are often produced in a particular biological context. Thus, accounting the context information allows the investigation of gene functions that are associated with specific biological conditions such as cell-types, tissue-types, developmental stages, and environmental stresses. For example, if two genes show CX as the disease progresses, they are highly likely to function together in the disease progression. By exploiting context-specific CX networks, researchers are able to gain functional insight into disease pathways (Gargalovic et al. 2006; Miller et al. 2010).

Evaluation and integration of inferred co-functional links

Success of network-based functional study relies on the quality of individual network links. Therefore, we need to evaluate the quality of inferred functional links. The quality of inferred links is measured by using gold-standard functional links (Figure 1, Evaluation step). Thus, the quality of gold-standard data could influence the entire process of studying gene functions from sequencing data via co-functional networks. The gold-standard set of co-functional links can be derived from pathway annotation databases such as Gene Ontology (GO) (Gene Ontology 2015), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2017), and MetaCyc (Caspi et al. 2016). GO provides controlled vocabulary of gene attributes in a loosely hierarchical structure and is composed of three categories: molecular function, cellular component, and biological process. Two genes that share biological process terms are highly likely to be participating in the same biological pathway. Pairing all genes of a single biological process term can generate gold-standard sets of co-functional links. Since GO, KEGG, and MetaCyc databases provide complementary pathway annotations, the combination of these resources would be useful in generating less biased and more comprehensive gold-standard sets of co-functional links. The quality of inferred co-functional links can be assessed by measuring how much the inferred links overlap with the gold-standard links. Co-functional links inferred from multiple sequencing data can then be integrated based on the standardized quality score into a genome-scale co-functional network (Figure 1, Integration step). A popular scheme of evaluation and integration of co-functional links is the log likelihood score based on the Bayesian statistics framework (Lee et al. 2004).

Generating functional hypotheses from co-functional networks

No gene is standalone. The functional interdependence of genes is the conceptual foundation for network-based generation of functional hypothesis. Network-based analysis can generate functional hypotheses by using the networks alone or by integrating networks with other external data such as functional genomics and genetics data. Numerous methods of network-based generation of functional hypotheses have been proposed, and most of them belong to one of the following categories.

Based on network connectivity

There are several different ways to use network connectivity for generating functional hypothesis (Figure 2(A)). In general, genes that are connected to many other genes are functionally important. The hypothesis generation using network connectivity exploits genes with high network centrality (i.e. hub genes) to identify key genes called essential genes for cellular viability (Jeong et al. 2001). Network connectivity information can also be integrated with functional genomics data to identify key modulators of the relevant process. For example, the MAster Regulator INference algorithm (MARINa) (Lefebvre et al. 2010) and Context-Associated Hubs (CAH) method (Cho et al. 2014) identify genes that are associated with a particular disease by significantly enriched connections to the differentially expressed genes (DEGs) in the disease-relevant context. In addition, group-wise connectivity measure can be used to identify functional modules for disease processes. For example, Disease Association Protein–Protein Link Evaluator (DAPPLE) (Trost et al. 2016) evaluates the significance of the observed network connectivity within a group of candidate genes from genome-wide association study (GWAS) to identify functional modules for the given disease.

Based on network propagation

Gene networks generally contain some genes already known for functions or phenotypes. Since genes are connected by functional associations in the co-functional networks, we can infer the functions of uncharacterized genes by propagating functional information of known genes through networks. The network propagation algorithms are divided into two conceptual categories (Figure 2(B)): (i) *direct neighborhood*, in which node information can propagate only to direct neighbors, and (ii) *network diffusion*, in which node information diffuses throughout the entire network (Shim et al. 2015). A popular form of direct neighborhood method is the naïve Bayes algorithm, in which the score of propagated function is based on the sum of all edge weights to the connected neighbors for the given function. Typical network diffusion methods include *random walk with restart*, *iterative ranking*, and *Gaussian Smoothing* (Wang & Marcotte 2010). Information from GWAS and mutation studies can be propagated through co-functional gene networks to prioritize candidate genes for diseases (Shim & Lee 2015). For examples, propagated GWAS significance scores from candidate genes through the network were used to identify additional candidate disease genes with low GWAS significance (Lee et al. 2011), and propagated mutation occurrence scores were used to identify cancer genes with low mutation occurrences among patients (Cho et al. 2016).

Based on subnetwork analysis

Identification of highly connected network communities that are enriched for genes of a phenotype will facilitate to study functional organization of phenotypes such as

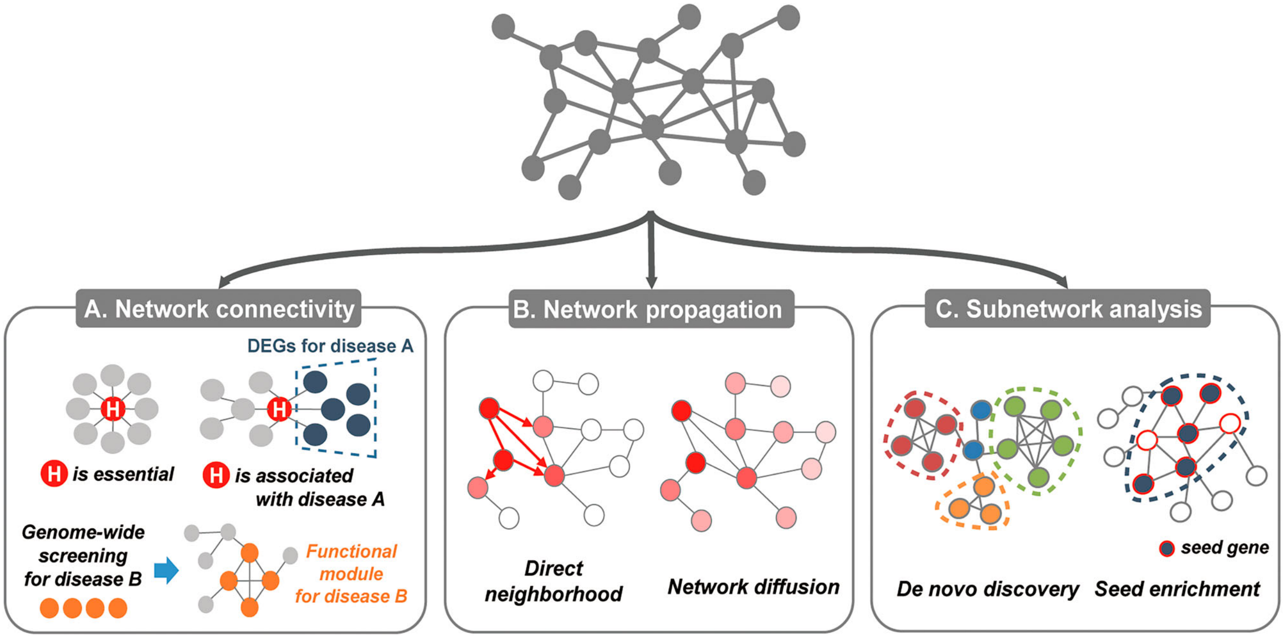

Figure 2. From co-functional networks to gene functions. Functional hypotheses can be generated by three different network approaches. (A) Methods based on network connectivity identify hub genes as essential genes, disease-associated genes by network connections to the DEGs in disease conditions, and disease-associated modules based on network connectivity within a group of candidate disease genes from genome-wide unbiased screening. (B) Functional information of known genes can be propagated to the direct neighbors or throughout the entire network by network diffusion. (C) Modules for processes and phenotypes can be identified by de novo discovery based on clustered network communities or by subnetwork enriched for seed genes, which are already known for the processes or phenotypes.

diseases and to identify additional phenotype-associated genes. Highly connected subnetwork structures may reflect functional modules, in which functionally coupled genes are highly interconnected. There are two approaches to finding subnetworks (Figure 2(C)): (i) *de novo discovery* and (ii) *seed enrichment*. The *de novo* discovery of subnetworks is generally based on network topology and finds a cohesive community in the network through network clustering algorithms such as greedy search for non-overlapping communities (Clauset et al. 2004; Newman 2004) and clique-based search for overlapping communities (Palla et al. 2005). In the case of seed enrichment approaches, subnetworks are identified by enrichment for seed genes which are known for a phenotype, often derived from external genome-wide unbiased genetic analysis such as GWAS and whole exome sequencing for disease samples. For example, dmGWAS (Jia et al. 2011) provides a dense module searching algorithm to identify candidate subnetworks or genes for diseases. HyperModules (Leung et al. 2014) software identifies significantly mutated subnetworks among patients using local network search heuristics to detect closely connected network regions. A software jActiveModule (Ideker et al. 2002) was developed to identify subnetworks enriched for DEGs.

Conclusions

We are facing an upcoming onslaught of sequencing data for diverse species due to revolutionary next-generation sequencing technology. Effective conversion of this major type of high-throughput data into functional information will be critical in enhancing our understanding of gene-to-phenotype associations in the era of genomics. The strategy described above can be easily generalized to any cellular organism. Methods of network inference and network-based hypothesis generation will continue to improve by incorporating additional data mining and graph analysis technologies. Systematic approaches to study gene functions have been possible in several laboratory model organisms only for many years. Integration of high-throughput sequencing technology and network science will open new avenues for genetics in all living organisms.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by grants from the National Research Foundation of Korea [2012M3A9B4028641, 2012M3A9C 7050151, and 2015R1A2A1A15055859] to I. L.

References

Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, Latendresse M, Mueller LA, et al. 2016. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucl Acids Res. 44:D471–D480.

Cho A, Shim JE, Kim E, Supek F, Lehner B, Lee I. 2016. MUFFINN: cancer gene discovery via network analysis of somatic mutation data. Genome Biol. 17:129.

Cho A, Shin J, Hwang S, Kim C, Shim H, Kim H, Kim H, Lee I. 2014. Wormnet v3: a network-assisted hypothesis-generating server for *Caenorhabditis elegans*. Nucl Acids Res. 42:W76–W82.

Clauset A, Newman MEJ, Moore C. 2004. Finding community structure in very large networks. Phys Rev E Stat Nonlin Soft Matter Phys. 70:066111.

Dandekar T, Snel B, Huynen M, Bork P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. 23:324–328.

Deng MH, Mehta S, Sun FZ, Chen T. 2002. Inferring domain–domain interactions from protein–protein interactions. Genome Res. 12:1540–1548.

Edgar R, Domrachev M, Lash AE. 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucl Acids Res. 30:207–210.

Gargalovic PS, Imura M, Zhang B, Gharavi NM, Clark MJ, Pagnon J, Yang WP, He A, Truong A, Patel S, et al. 2006. Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. Proc Natl Acad Sci U S A. 103:12741–12746.

Gene Ontology C. 2015. Gene ontology consortium: going forward. Nucl Acids Res. 43:D1049–D1056.

Ideker T, Ozier O, Schwikowski B, Siegel AF. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 18 Suppl. 1:S233–S240.

Jeong H, Mason SP, Barabasi AL, Oltvai ZN. 2001. Lethality and centrality in protein networks. Nature. 411:41–42.

Jia PL, Zheng SY, Long JR, Zheng W, Zhao ZM. 2011. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. Bioinformatics. 27:95–102.

Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucl Acids Res. 45:D353–D361.

Kensche PR, van Noort V, Dutilh BE, Huynen MA. 2008. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface. 5:151–170.

Kim E, Kim H, Lee I. 2013. Jiffynet: a web-based instant protein network modeler for newly sequenced species. Nucl Acids Res. 41:W192–W197.

Kodama Y, Shumway M, Leinonen R, International Nucleotide Sequence Database C. 2012. The sequence read archive: explosive growth of sequencing data. Nucl Acids Res. 40: D54-D56.

Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. 2011. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. 21:1109–1121.

Lee I, Date SV, Adai AT, Marcotte EM. 2004. A probabilistic functional network of yeast genes. Science. 306:1555–1558.

Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, Wang K, Sumazin P, Kustagi M, Bisikirska BC, et al. 2010. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol Syst Biol. 6:377.

Leung A, Bader GD, Reimand J. 2014. Hypermodules: identifying clinically and phenotypically significant network modules with disease mutations for biomarker discovery. Bioinformatics. 30:2230–2232.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucl Acids Res. 32: W20–W25.

Miller JA, Horvath S, Geschwind DH. 2010. Divergence of human and mouse brain transcriptome highlights Alzheimer disease pathways. Proc Natl Acad Sci U S A. 107:12698–12703.

Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, et al. 2015. The InterPro protein families database: the classification resource after 15 years. Nucl Acids Res. 43:D213–D221.

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, et al. 2017. Genomes OnLine database (GOLD) v.6: data updates and feature enhancements. Nucl Acids Res. 45:D446–D456.

Newman ME. 2004. Fast algorithm for detecting community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys. 69:066133.

Palla G, Derenyi I, Farkas I, Vicsek T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature. 435:814–818.

Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, Farne A, Lara GG, Holloway E, Kapushesky M, et al. 2005. ArrayExpress – a public repository for microarray gene expression data at the EBI. Nucl Acids Res. 33:D553–D555.

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 96:4285–4288.

Shim JE, Hwang S, Lee I. 2015. Pathway-dependent effectiveness of network algorithms for gene prioritization. PLoS One. 10:e0130589.

Shim JE, Lee I. 2015. Network-assisted approaches for human disease research. Anim Cells Syst. 19:231–235.

Shim JE, Lee I. 2016. Weighted mutual information analysis substantially improves domain-based functional network models. Bioinformatics. 32:2824–2830.

Shin J, Lee I. 2015. Co-inheritance analysis within the domains of life substantially improves network inference by phylogenetic profiling. PLoS One. 10:e0139006.

Shin J, Lee I. 2017. Construction of functional gene networks using phylogenetic profiles. Methods Mol Biol. 1526:87–98.

Shin J, Lee T, Kim H, Lee I. 2014. Complementarity between distance- and probability-based methods of gene neighbourhood identification for pathway reconstruction. Mol Biosyst. 10:24–29.

Song L, Langfelder P, Horvath S. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. BMC Bioinform. 13:328.

Sprinzak E, Margalit H. 2001. Correlated sequence-signatures as markers of protein–protein interaction. J Mol Biol. 311:681–692.

Trost B, Maleki F, Kusalik A, Napper S. 2016. DAPPLE 2: a tool for the homology-based prediction of post-translational modification sites. J Proteome Res. 15:2760–2767.

Walhout AJ, Sordella R, Lu X, Hartley JL, Temple GF, Brasch MA, Thierry-Mieg N, Vidal M. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science. 287:116–122.

Wang PI, Marcotte EM. 2010. It's the machine that matters: predicting gene function and phenotype from protein networks. J Proteomics. 73:2277–2289.

Yellaboina S, Tasneem A, Zaykin DV, Raghavachari B, Jothi R. 2011. DOMINE: a comprehensive collection of known and predicted domain–domain interactions. Nucl Acids Res. 39: D730–D735.