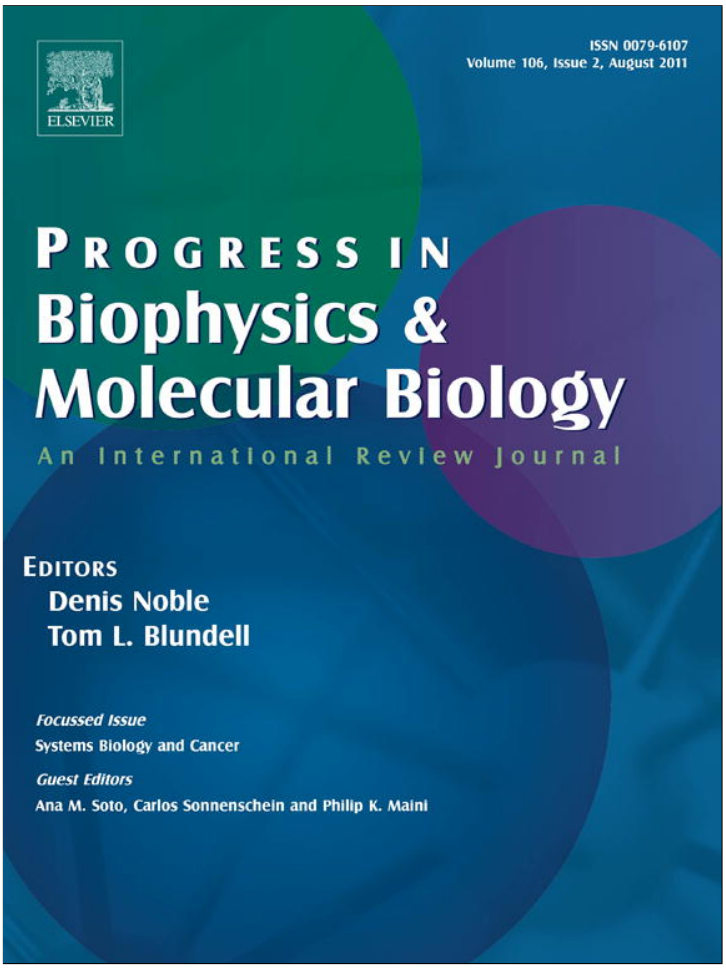

Progress in Biophysics and Molecular Biology 106 (2011) 435–442

Contents lists available at ScienceDirect

Progress in Biophysics and Molecular Biology

journal homepage: www.elsevier.com/locate/pbiomolbio

Review

Probabilistic functional gene societies

Insuk Lee

Department of Biotechnology, College of Life Science and Biotechnology, Yonsei University, 262 Seongsanno, Seodaemun-gu, Seoul 120-749, Republic of Korea

ARTICLE INFO

Article history:
Available online 31 January 2011

Keywords:
Complex phenotypes
Gene networks
Data integration
Guilt-by-association
Genotype–phenotype associations

ABSTRACT

A cellular system may be viewed as a social network of genes. Genes work together to conduct physiological processes in the cells. Thus if we have a view of the functional association among genes, we may also be able to unravel the association between genotypes and phenotypes; the emergent properties of interactive activities of genes. We could have various points of view for a gene network. Perhaps the most common standpoints are protein–protein interaction networks (PPIN) and transcriptional regulatory networks (TRN). Here I introduce another type of view for the gene network; the probabilistic functional gene network (PFGN). A 'functional view' of association between genes enables us to have a holistic model of the gene society. A 'probabilistic view' makes the model of gene associations derived from noisy high-throughput data more robust. In addition, the dynamics of gene association may be presented in a single static network model by the probabilistic view. By combining the two modeling views, the probabilistic functional gene networks have been constructed for various organisms and proved to be highly useful in generating novel biological hypotheses not only for simple unicellular microbes, but also for highly complex multicellular animals and plants.

Contents

1. Introduction

Genes are social. Most genes collaborate with other genes to conduct physiological processes in the cells or organisms. Their strong social nature is analogous to that of human beings. The sociality underlies the multi-functionalities of both humans and genes. For example, a man may have multiple roles as a husband, father, son, and friend in different social contexts by having different relationships with his wife, children, parents, and friends, respectively. Likewise, a single gene may be involved in various cellular processes by having different collaborative gene partners.

E-mail address: insuklee@yonsei.ac.kr.

0079-6107/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.pbiomolbio.2011.01.003

Therefore, the sociality of genes is an organisms' fundamental mechanism to create a larger number of traits than the number of genes.

Social network modeling is a successful approach to study human society. For instance, a social network enables the identification of hub individuals influencing a large number of people in the society. Hub individuals are major targets in marketing strategies. Recently, the similar network approach has become more pragmatic in biology, because technologies for mapping connections between genes have dramatically improved during the last decade. Thus now we can identify hub genes in gene networks and observe a strong correlation between network centrality and functional essentiality of genes not only in microorganisms (Jeong et al., 2001) but also in animals (Lee et al., 2008). Social affinity of genes may also be exploited to discover novel gene functions by guilt-by-association (GBA) (Lee et al., 2006). Furthermore, it has become more evident that many phenotypes are implemented in the organismal system as communities of genes (Lee et al., 2008), and functional interactions between the communities provide ways of modulating phenotypes (Lee et al., 2010b). Therefore, the network modeling of a gene society will provide new opportunities in genetics research and rational therapeutics.

The same gene society may be modeled by various network views. First, the protein–protein interaction network (PPIN) that maps direct or indirect physical contacts between proteins is the most popular modeling view of gene societies. This popularity is rooted in recently advanced detection methods of protein–protein interaction (PPI) (Lalonde et al., 2008). The experimental detection of PPIs, however, is still quite limited in animals and plants compared to yeast, in which the consolidated set of PPIs covers the majority of the proteome (Kim et al., 2010). In fact, there are many connections between genes mediated by non-physical associations. Another popular view of gene society is the transcriptional regulatory network (TRN) that maps directional relationships from transcriptional factors (TF) to their target genes. Experimental approaches such as chromatin immuno-precipitation followed by a promoter sequence analysis based on a DNA chip (ChIP–chip analysis) or next generation sequencing (ChIP-seq analysis) have suggested the majority of the known TF-target gene relations (Kim et al., 2009).

In this article, I introduce a relatively newer modeling view of the gene network; a probabilistic functional gene network (PFGN) that maps functional associations between genes (Fraser and Marcotte, 2004; Lee et al., 2004). What would be merits of a probabilistic functional view for modeling a gene network? How do we construct the PFGN? How do we utilize the PFGN for new biological discoveries? The entire workflow of network-driven biology from the analysis of genomics data to the generation of new biological hypotheses may be largely divided into two data integration steps: (1) integrating diverse omics data into a probabilistic functional link and (2) integrating multiple functional links into new biological hypotheses. Currently PFGNs are used to generate three types of biological hypotheses: (1) novel gene functions, (2) genetic dissection of complex phenotypes, and (3) novel genetic (epistatic) interactions. I will describe the basic concepts, construction methods, and recent progresses in the development of PFGNs for various organisms and their contributions to modern genetics research.

2. Key ideas about the probabilistic functional view of a gene network.

2.1. Why a functional view of the gene network?

Genes may connect to each other by various types of associations in the cell. For example, a set of genes for the same protein complex are associated via physical contacts between proteins, while another set of genes may work together for a biological process without protein–protein interactions. Each gene pair is associated by specific types of connections – protein–protein interaction, regulator–target relation, co-expression, co-transcription, and so forth. In other words, none of the specific types of association alone can explain the entire interactive activities among genes.

Then, what are the merits of a functional view over other modeling views of the gene network? Functional association is fuzzy and is a more generalized notion of gene–gene relations that may include more specific definitions of the association between genes, such as the physical contact between gene products and regulator–target relations. Thus the specific types of relations between genes can be represented by a more inclusive type of relation, the functional association. The consolidation of various types of associations by using the more inclusive functional association results in a more extended coverage of genome by the gene network. The extended coverage of the genome turns out to be a critical factor determining the general predictive power of a gene network (Lee et al., 2008; McGary et al., 2007). Extending the coverage of gene networks is more challenging for higher eukaryotes such as animals and plants, because of their larger search space of pair-wise relationship modeling; the number of possible gene pairs increases combinatorially as the number of genes does linearly. Therefore, the benefits from the functional view of the gene network may be much bigger for higher animal and plant species such as human and crops (Kim et al., 2010).

2.2. Why a probabilistic view of the gene network?

Then why is a probabilistic view more appropriate than a deterministic view for modeling a gene network? First, the probabilistic view of complex systems provides more robust models. In general, genome-scale data used for modeling cellular systems are not perfect in either accuracy or completeness. Thus a systemic model constructed with incomplete and erroneous data would have different view-qualities for different parts of the system; some parts clear and the others blurry. The partially blurry view of the whole system needs to be adjusted based on clarity of each part to enhance the robustness of inference for the whole system. For example, conclusions about a part of the system that is viewed unclearly should be taken by a low probability score. Second, the dynamicity of interactions between genes can be partially implemented by the probabilistic view. Some gene-pair associations are highly stable (e.g., genes for stable protein complexes such as ribosomes and proteasomes), while others are context-specific (e.g., genes for stress response). Most of the genes may have different collaboration partners for different biological contexts. This is actually the way genes play multiple roles in the cells. We can assign higher probabilities for more stable gene interactions, because a more consistent interaction is more probable to be observed for a particular biological context. In the same sense, the gene interactions for rarer contexts tend to be scored with lower probability. If we construct a single static gene network modeling all possible cellular contexts, the probabilistic view partially complements the lack of context-specific information of the static model.

2.3. Two levels of data integration

Both the probabilistic and functional views enhance robustness as well as completeness of the gene network model, because they facilitate data integration in the bottom-up approach of system modeling. The whole gene network is composed of various types of functional relations among genes that are largely complementary to each other. Therefore, well-designed data integration of the

various experimental data and types of gene association improves the power of inference about the gene network.

The whole process of learning biology from various omics data, such as transcriptome, proteome, interactome profiles, to a network-based hypothesis generation can be divided into two levels of data integration: (1) integrating various omics data into each probabilistic functional link, and (2) integrating multiple functional links into new functional hypotheses. The quality of a gene network model relies heavily on the accuracy of each link between the genes. Upon integrating multiple omics data, we may observe a general increase in the accuracy of many functional links by multiple supporting evidences (Fig. 1A). The positive correlation between the number of evidence and likelihood of a functional link suggests that the reliability of each functional association between genes improves with multiple supporting data. The next level of data integration is subject to the functional links that may be already the products of data integration. To generate hypotheses about gene functions and others, we can use multiple links connected to the query gene. Thus each new hypothesis is based on the integration of information from multiple functional links to the query gene. The comparison between the predictability of a gene network by using only one nearest neighbor (i.e., one that is most confidently connected to the network neighbor) and using all neighbors of the query gene shows a clear difference (Fig. 1B). For the rest of the article, I will describe the whole workflow from diverse types of omics data to the new biological hypotheses with two sections for each level of data integration.

3. Integrating diverse omics data into a probabilistic functional link

3.1. Inference of functional links between genes

Nodes and edges are the basic building blocks of networks. Accordingly, genes and functional associations between them are basic building blocks of functional gene networks. There are various types of omics data that are able to imply functional associations between genes. I will describe four major classes of such data types here.

First, protein–protein interactions via physical contacts are strong predictors of functional associations between protein coding genes. Physical contacts mediate many functional communications between gene products. Protein complexes conducting whole biological processes comprise multiple proteins physically associated to each other. Signaling proteins convey cellular signals via a series of physical contacts between them. Yeast two hybrid is perhaps the most widely used experimental technique to detect protein–protein interactions. This method has been employed for genome-wide protein–protein interaction mapping in various species including yeast (Ito et al., 2001; Uetz et al., 2000; Yu et al., 2008), worm (Li et al., 2004; Simonis et al., 2009), fly (Giot et al., 2003), and human (Rual et al., 2005; Stelzl et al., 2005). Affinity purification followed by mass spectrometry analysis (APMS) is another popular large-scale protein–protein interaction mapping method that has been used for yeast (Gavin et al., 2006; Krogan et al., 2006) and human (Ewing et al., 2007; Hutchins et al., 2010) so far.

Second, co-expression is a prevalent pattern supporting functional associations between genes. Genes for the same biological processes tend to co-express across various biological contexts. Huge amount of microarray data are currently deposited into public databases. Gene Expression Omnibus (Barrett et al., 2009) only contained >470,000 array samples on September, 2010. With their versatility by various meta-analysis and high sensitivity for gene expression signals, mapping gene functional associations using gene expression microarray data is recognized more importantly than others now. For example, co-expression links cover >30%, >70%, and >30% of the total functional links of yeast, worm, and Arabidopsis gene networks, respectively (Lee et al., 2010a, 2010b, 2007).

Third, we can infer functional associations among genes by their similar genomic context. There are three major genome-context patterns we use currently: phylogenetic profiling pattern (Huynen et al., 2000; Pellegrini et al., 1999; Wolf et al., 2001), gene

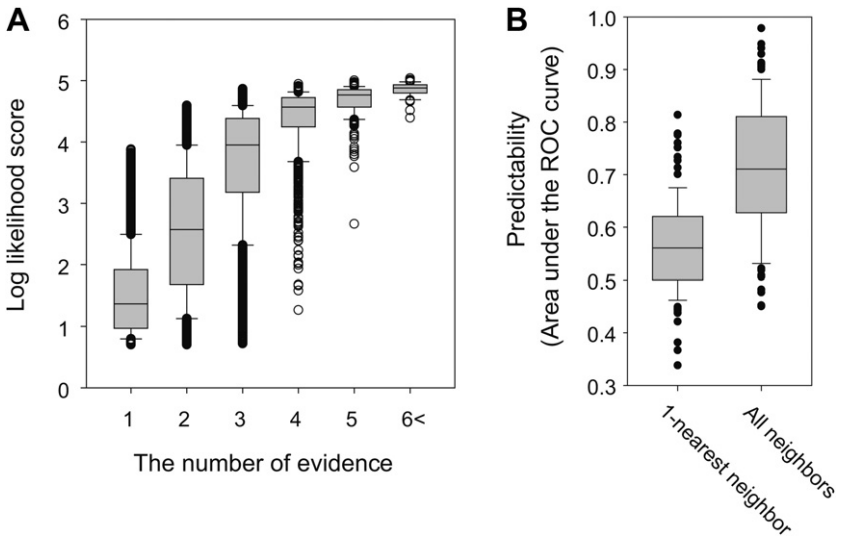

Fig. 1. Effects of two levels of data integration. (A) The effect of integrating multiple genomics data into each functional association. The latest version of the functional gene network for yeast (Lee et al., 2007) has functional associations supported by various number of evidences, from a single evidence to eight at most. The number of supporting evidence for each functional association positively correlates with its log likelihood score, indicating an improvement in inference power for each functional link by support from more genomics data. The lines in the middle of the box represents the median and the two edges represent the top 25% and 75% of the log likelihood scores of the set of links supported by same number of evidences, respectively. (B) Effect of integrating multiple functional links into new biological hypotheses. In the network, each gene may have multiple neighboring genes. Functional inference by guilt-by-association may use either one nearest neighboring gene or all neighboring genes for each query gene. A network-based classifier for known phenotypic genes can be assessed by the receiver operating characteristics (ROC) curve (Lehner and Lee, 2008). The predictability for known genes for each of the 100 knock-out phenotypes (McGary et al., 2007) has been assessed by the area under the ROC curve (AUC) with two options of guilt-by-association. The box-and-whiskers plot clearly shows an improved predictability by using all network neighbors including ones connected by weak associations.

neighboring pattern (Bowers et al., 2004; Dandekar et al., 1998; Overbeek et al., 1999), and gene fusion pattern (or Rosetta stone protein) (Enright et al., 1999; Marcotte et al., 1999). During speciation, descendants inherit a portion of the ancestors' biological processes. Because many cellular functions are carried at pathway-level, proteins for the same pathways or processes are co-inherited. This co-inheritance pattern can be detected from phylogenetic profiles across many completely sequenced genomes. With the advent of era of next generation sequencing (NGS), we expect a large number of fully sequenced genomes at hands (Genome Online Database (GOLD, http://www.genomesonline.org) reports >1300 fully sequenced genomes and ~7000 genomes committed by September 2010). The newly sequenced genomes will expand the profile size, thus potentially provide more information. Consequently, the usefulness of this method would extend to complex eukaryotes such as animals and plants in near future. The next genome-context approach, the gene neighboring method, uses bacterial operon structures. In prokaryotes, genes involved in the same metabolic pathway tend to be transcribed as a single mRNA molecule encoding all the members of the proteins on it, and thus the co-transcriptional unit is called operon. Although eukaryotes have lost the operon structure from their genomes during evolution, we can deduce the context of eukaryotic proteins from their ancestral prokaryotic genomes by orthology. If two eukaryotic genes frequently have their orthologs in the proximal chromosomal loci of prokaryotes, their ancestral genes would likely be located in the same operon, thereby implicating functional association. The last genome-context approach is the Rosetta Stone (gene fusion) method. Two proteins encoded in separate loci on a genome could have their orthologs in another genome fused into a single gene. The gene fusion event in a different genome-context tends to supports the functional association between two genes. We can efficiently detect the gene fusion by ortholog mapping based on BLASTP (Altschul et al., 1990) among genomes. However, the gene fusion also suggests many promiscuous gene pairs that need to be filtered out (Marcotte et al., 1999).

Fourth, the functional association between genes in an organism may be transferred to another by their ortholog pairs which are dubbed as associalogs (Lee et al., 2008). Each organism has study bias towards different biological processes. Thus, borrowing functional information from other organisms may compensate for less studied parts of cellular systems. Noticeably we can predict many plant specific traits by associalogs from animals or yeast (Lee et al., 2010a). This observation indicates that not only the genes but also the 'gene associations' are evolutionary reused for diverse biological processes and traits in different species.

In addition to the above methods, gene functional associations can be inferred also from domain co-occurrence patterns (Lee et al., 2010a), co-citation patterns (Jenssen et al., 2001; Stapley and Benoit, 2000), on tertiary structures of proteins (Aloy and Russell, 2003), and so forth. The introduction of a comprehensive list of linkage discovery methods is beyond the scope of this review article. As we observe continuous approaches to many novel types of omics data, we expect to see many more new methods in discovering the edge building blocks, the functional connections between genes, in near future.

3.2. Construction of a probabilistic functional gene network

In cells or organisms, gene networks may be composed of heterogeneous types of relationships. This heterogeneous nature of relations among genes is in fact one of the major reasons why we would benefit from using diverse omics data for the network modeling. The potential of heterogeneity, however, comes with the technical complication of integration. Different data types may have intrinsically different predictive power. Moreover, individual data sets based on the same technical platform may also result in various qualities. Therefore, data standardization needs to precede the integration. For efficient integration, a functional association that represents an inclusive notion of gene interactions is advantageous as we discussed earlier. The Bayesian statistics framework of data standardization with a view of functional association has been developed (Lee et al., 2004) and proved to be highly robust for diverse data sets and organisms. The scoring scheme is called log likelihood score (LLS),

$$LLS = \ln\left(\frac{P(A|E)/P(\sim A|E)}{P(A)/P(\sim A)}\right)$$

where $P(A|E)$ and $P(\sim A|E)$ are the frequencies of gene associations (A) observed with the given experimental evidence (E) between annotated genes operating in the *same* biological processes and in *different* processes, respectively, while $P(A)$ and $P(\sim A)$ represent the prior expectations (*i.e.*, the total probability of associations between all annotated genes operating in the *same* biological processes and operating in *different* processes, respectively). A *LLS* score that is greater than zero indicates that the data set supports the associations between two genes in the same biological processes, with higher scores indicating a more confident support of the association. Fig. 2 illustrates an example of the log likelihood score calculation from yeast mRNA expression profiles across various cell-cycle conditions.

None of data sets is perfect, so that integration of multiple data sets with various sensitivities and specificities for different areas of cellular systems would improve not only the accuracy but also the coverage of the gene network (Fig. 3). By standardizing the heterogeneous data into a unified log likelihood score, we can achieve data integration with a mathematical equation taking account of the dependence among different data sets. The weighted sum (WS), a modified naïve Bayesian integration method to integrate multiple *LLS* scores for a given gene pair was calculated as:

$$WS = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \cdot i}, \text{ for all } L \geq T$$

where L_0 represents the primary (i.e., highest) *LLS* score for a given gene pair, D is a free parameter determining the decay rate of the weight for secondary evidences, and i is the rank order index of multiple *LLS* scores associated with a given gene pair. Ranking starts from the second highest *LLS* with a descending magnitude for all n remaining *LLS* scores. To exclude noisy linkage information, we consider only the *LLS* scores above the empirically chosen threshold T during integration. The free parameter D ranges from 1 to $+\infty$, and is optimized to maximize the overall performance of the integrated model, assessed by a precision-recall curve for the recovery of test set gene pairs. As the optimal value of D approaches $+\infty$, *WS* approaches the L_0, and the lower scoring *LLS* scores do not provide any additional information, implying that all data sets are completely interdependent. Each individual gene–gene association may have a different data type with its primary *LLS*. We independently test the performance of a *naïve* Bayesian integration of the *LLS* scores (which is simply the sum of the *LLS* scores for each given gene pair), then select the integration model that maximizes the area under the plot of *LLS* versus coverage of genes incorporated in the network.

3.3. Probabilistic functional gene networks for various organisms

We have constructed probabilistic functional gene networks (PFGN) for various organisms – baker's yeast *Saccharomyces*

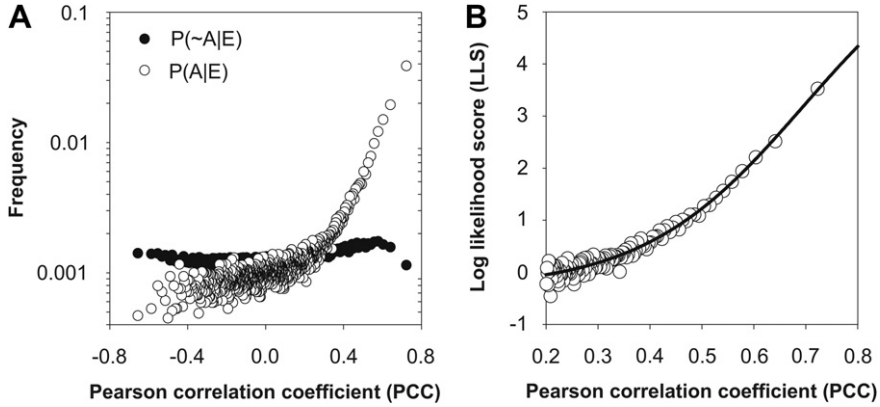

Fig. 2. Calculation of log likelihood score of functional association derived from gene expression data. (A) Genes operating same biological processes tend to co-express through the course of gene expression experiments. Thus, the Pearson correlation coefficient (PCC) of two genes' expression vectors shows a positive correlation with their frequency of sharing pathway annotation. For example, yeast mRNA co-expression patterns across various cell-cycle conditions (Spellman et al., 1998) were compared to their frequencies participating in the same KEGG pathways (Kanehisa et al., 2002), and plotted for all annotated gene pairs as a function of PCC. The frequencies of gene pairs sharing KEGG pathways for each PCC value ($P(A|E)$) were calculated for bins of 20,000 gene pairs, and showed strong correlations with the PCC. In contrast, the frequency of gene pairs not sharing the pathway annotation ($P(\sim A|E)$) showed no significant correlation with the PCC. (B) The ratio of these two frequencies that were normalized by the prior expectation ($P(A)$ and $P(\sim A)$) and transformed by logarithm provided the log likelihood scores (LLS) belonging to the same pathway for each PCC. Using a regression model for the relationship between PCC and LLS, we can assign LLS to all gene pairs with PCC (including gene pairs that are not annotated). This figure has been adapted from Lee et al. (2004).

cerevisiae (Lee et al., 2004, 2007), a simple animal *Caenorhabditis elegans* (Lee et al., 2008, 2010b), and the reference plant *Arabidopsis* (Lee et al., 2010a) during the past several years. Unlike protein–protein interaction network (PPIN), PFGN covers the majority of genes for even animals and plants with large genome sizes (Table 1). Consequently, we can generate novel biological hypotheses for more genes with PFGN than with PPIN. One reason for the high genome coverage of PFGN is the transferred network-linkage information among different organisms by orthology. Each organism has its own niche in modeling biological processes. For example, yeast have largely contributed in understanding basic biological processes such as transcription and translation, the worm is a major model system in studying programmed cell death and aging, and Arabidopsis is an excellent system to study stress responses. Because yeast, worm, and Arabidopsis all share many of such physiologies, a new discovery about a particular biological process from one of organisms may contribute to the network of other organisms in which study of the same biological process is not accessible yet. In other words, PFGNs for various organisms have been constructed by another level of data integration, integrating network-linkage data from multi-organisms. One interesting observation during PFGN construction using information transfer among multi-organisms was that the functional gene association of animals contributes to modeling plant specific biological processes such as trichome development (Lee et al., 2010a). Construction of PFGNs of human and crops will be feasible with similar modeling framework in near future.

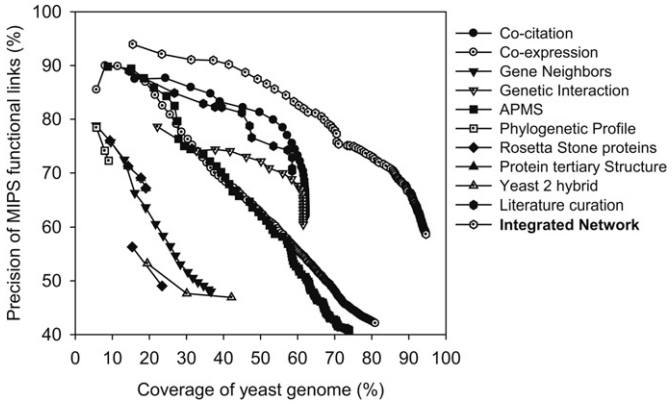

Fig. 3. Improvement of both accuracy and coverage of gene functional network by integration of diverse omics data. The latest version of the functional gene network of yeast (Lee et al., 2007) was trained by Gene Ontology annotations (Christie et al., 2009). Thus reference gene pairs by sharing the same MIPS (The Munich Information center for Protein Sequence) functional annotation (Mewes et al., 2002) could be an independent test set for assessing the quality of networks by individual data sets and the integrated network. The quality of each network was measured by the percentage of gene pairs sharing the same MIPS functional annotations (accuracy) and the percentage of genome (coverage) for the set of gene pairs at every thousand gene pairs in cumulative manner. The resulting curve suggests a clear qualitative improvement in both accuracy and coverage of the integrated network compared to any other networks based on a single data set.

4. Integrating multiple functional links into new biological hypotheses

Data integration with a functional view as described above provides highly accurate and comprehensive gene networks. What then can we do with the gene networks for a novel biological discovery? Indeed, the "how-to-use" is more important than "how-to-make" in the field of network biology. Here, we propose the guilt-by-association (GBA) approach ushering to the discovery of novel gene functions, genetic dissection of complex phenotypes, and mapping of epistatic interactions between genes. The generation of biological hypotheses from guilt-by-association using multiple network links is another level of the data integration process that improves the inference power.

4.1. Discovery of gene functions

Network-based inference has been applied to various complex systems. Guilt-by-association (GBA) is a major approach of network-

Table 1
Functional gene networks for various organisms.

Organism	Genome coverage (# gene)	# links	URL of network database
Yeast	95% (5483)	102,803	www.functionalnet.org/yeastnet
C. elegans	75% (15,139)	999,367	www.functionalnet.org/wormnet
Arabidopsis	73% (19,647)	1,062,222	www.functionalnet.org/aranet

based inference about functions of network nodes, such as individual persons of a social network or individual genes of a gene network. In human society, people who work for the same task functionally connect to each other, forming a task-oriented community. Example task-oriented communities in human society are advisory boards, managing committees, academic institutions, and so on. Similarly, genes for the same biological process tend to connect to each other in functional gene networks and thus form functional modules. Therefore, if we have a set of genes known to be involved in a particular biological function and are interconnected in the gene network, uncharacterized genes that are connected to the functionally known genes are likely to be novel candidate genes for the same function (Fig. 4A). The feasibility of GBA in identification of gene functions has been experimentally validated by discovery of novel yeast ribosomal biogenesis genes (Li et al., 2009).

4.2. Genetic dissection of complex phenotypes

One of the fundamental goals in genetics is mapping genotype–phenotype associations. Gene-to-trait links are critical information platforms to step towards manipulating and engineering traits, for example, chemical inhibition of disease genes to manipulate disease symptoms, or genetic engineering of stress response genes in crops to enhance their stress resistance. Most of phenotypes are complex, that is, they cannot be explained by activity of a single gene, but rather by complex interactions among many genes. Thus, complex phenotypes are emergent properties of complex interactive activities among many genes. Genetic alliance among genes for a given complex phenotype can be dissected by identifying all component genes and mapping interactions among them.

Dissecting complex phenotypes into individual component genes may be facilitated by GBA for the discovery of gene functions. The GBA approach may work to study complex phenotypes provided with one assumption that genes modulating a complex phenotype are members of the same pathway or closely related pathways. Then we would see well-interconnected genes for the same phenotype in the functional gene network. If the known phenotypic genes satisfy this condition (Fig. 4A), new candidate genes for the same phenotype can be inferred from network neighbors.

Then, are all network neighbors of known phenotypic genes confident candidates? A given gene network may not be highly predictable for all phenotypes, mainly by two reasons. First, none of networks is perfect. This is a general problem of network-based biological predictions. Second, the genetic nature of a given phenotype itself could be not-so-modular. In other words, genes participating in processes that are distantly apart from each other in the network contribute to the same phenotype. In this case, genes for the same phenotype may be sparsely distributed in the network; consequently, they are not well-interconnected in the network. Therefore, the GBA approach to genetic dissection of complex phenotypes requires a pre-evaluation of modularity of known phenotypic genes in the network (Lehner and Lee, 2008). The feasibility of the GBA approach in the discovery of phenotypic genes has been experimentally validated by elongated yeast cell morphology (McGary et al., 2007), suppressors of retinoblastoma (Rb) related tumorigenesis in *C. elegans* (Lee et al., 2008), and Arabidopsis seed pigmentation (Lee et al., 2010a). These results strongly suggest that the same network-based prediction works for gene-to-trait mapping not only for a simple unicellular organism but also for more complex multicellular animals and plants that have multiple types of tissues and cells.

4.3. Prediction of epistatic interactions

Unraveling the genetic organizations of complex phenotypes also requires a map of non-additive functional interactions among genes for each phenotype. An inheritable portion of a phenotype is explained by the collective effects derived from all member genes. However, a combination of multiple genes does not always show a simple additive effect. Some pairs of double mutations cause a much severer or alleviated phenotypic effect than a simple addition of two individual mutational effects (Dixon et al., 2009). Recently, this non-additive combination of genetic variation on phenotype – namely epistasis – is thought to be one of the major reasons for the missing inheritance of complex traits during the genome-wide association study (GWAS) that usually focuses on the phenotypic effect of a single polymorphic position (Manolio et al., 2009). Therefore mapping the epistatic interaction between genes would be a key path towards understanding the genetic organization of complex traits. For example, epistatic interactions between hub cancer genes such as p53 and other cancer related genes would provide important clues in understanding the mechanism of tumorigenesis (Uren et al., 2008).

There are two different types of epistatic interactions: within-pathway interaction and between-pathway interaction (Boone et al., 2007). The classification is based on their locations in the pathway map such as PPIN and PFGN. Genes for the same pathways or functional modules are highly interconnected and usually result in local community structures in PPIN and PFGN. A within-pathway epistatic interaction is observed within a local network community.

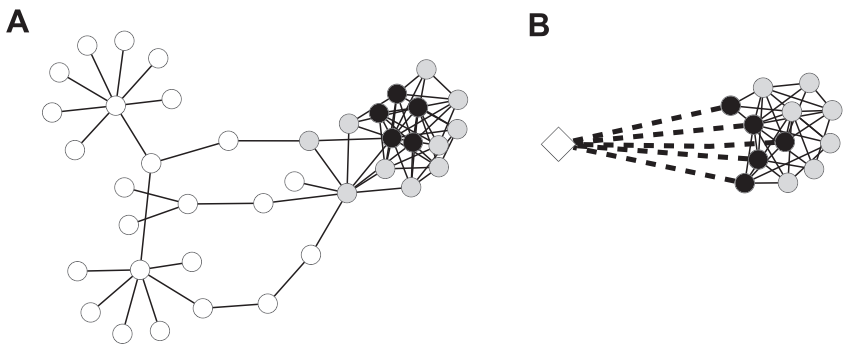

Fig. 4. Guilt-by-association approach to discovery of new gene function, genetic dissection of complex phenotypes, and discovery of epistatic interactions. (A) Black nodes represent genes that are known to be of the same function or involved in the same phenotype. Because the known genes are well-interconnected in the example, their network neighbors (grey nodes) may also be member genes for the same function or phenotype. (B) If we have pre-identified genetic modifier (black nodes) for a particular phenotypic gene (the diamond shaped node) and they are well-interconnected in the network, they may also participate in the same pathway or process. Then, other genes for the same pathway or process (grey nodes connected to the black nodes) may also be novel genetic modifiers for the same phenotypic gene.

On the contrary, a between-pathway epistatic interaction is observed between two local network communities. During the last decade, geneticists have accumulated a large amount of epistatic interaction data through systematic high-throughput experiments, especially in two fungal species, *S. cerevisiae* (Costanzo et al., 2010) and *Schizosaccharomyces pombe* (Roguev et al., 2007), and in an animal species, *C. elegans* (Lehner et al., 2006). From the systematic and comparative analyses of these large-scale epistatic interaction networks, systems biologists found the following general properties of epistatic interactions: (1) Epistatic interactions are enriched for between-pathway interactions (Kelley and Ideker, 2005). (2) The evolutionary conservation of epistatic interactions among species is much lower than that of protein–protein interactions or functional associations (Roguev et al., 2008; Tischler et al., 2008). The low evolutionary conservation of epistatic interactions suggests that simple interologs – interactions between orthologs – mapping (Yu et al., 2004) would not be effective for discovery of epistatic interactions. Yeast interologs have proved to be highly effective in the discovery of new protein–protein interactions in humans (Lehner and Fraser, 2004), but may not be true for the discovery of human epistatic interactions.

The discovery of new epistatic interactions, however, can be facilitated by the functional gene network. As seen in Fig. 4B, if the majority of known genetic modifiers (epistatic interactors) for a particular phenotypic gene (e.g., a disease gene) participates in the same pathway (i.e., they are well-interconnected in the gene network), one may easily predict new modifiers for the same phenotypic gene from network neighbors of the known modifiers. Using this approach, a total of 31 novel genetic modifiers for three worm signal transduction genes have been identified with 7.3-folds enrichment compared to the previous screens that were based on semi-random candidates (Lee et al., 2010b).

5. Conclusions

The pleiotropy of genes and the complexity of phenotypes are all emergent properties of interactive activities among genes. Thus obtaining the map for a gene network is an important goal in modern systems biology. Among many possible modeling views of a gene network, the probabilistic functional view provides advantages in efficient integration of heterogeneous genomics data to construct a more accurate and comprehensive model. In addition, the network model itself allows opportunities for a higher level of data integration to generate more confident biological hypotheses. To date, the probabilistic functional view has been successfully used in constructing network models for a unicellular microbe yeast, a simple multicellular animal worm, and the reference plant Arabidopsis. These earlier accomplishments suggest an optimistic future development for gene network models for humans and crops. The network-based approaches in medical and agricultural researches may bring a new paradigm to a more predictive and cost-effective genetic analysis to clinically and economically important phenotypes such as complex diseases in humans and stress resistance in food or energy crops. The initial goal of systems biology was to learn more about the whole organismal system. After all, we may learn more about individual genes from the systemic model, just like the way we learn about a person via the analysis of his social network.

Acknowledgments

I thank Moonhee Lee for help editing the manuscript. This work was supported by grants from the National Research Foundation of Korea (NRF) funded by the Korea government (MEST) (No. 20100017649, 20100001818, 20090087951, 20100015754) and POSCO TJ Park Junior Faculty Fellowship.

References

Aloy, P., Russell, R.B., 2003. InterPreTS: protein interaction prediction through tertiary structure. Bioinformatics 19, 161–162.
Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I.F., Soboleva, A., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Muertter, R.N., Edgar, R., 2009. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 37, D885–D890.
Boone, C., Bussey, H., Andrews, B.J., 2007. Exploring genetic interactions and networks with yeast. Nat. Rev. Genet. 8, 437–449.
Bowers, P.M., Pellegrini, M., Thompson, M.J., Fierro, J., Yeates, T.O., Eisenberg, D., 2004. Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol. 5, R35.
Christie, K.R., Hong, E.L., Cherry, J.M., 2009. Functional annotations for the *Saccharomyces cerevisiae* genome: the knowns and the known unknowns. Trends Microbiol. 17, 286–294.
Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.Y., Liang, W., Marback, M., Paw, J., San Luis, B.J., Shuteriqi, E., Tong, A.H., van Dyk, N., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.C., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J., Boone, C., 2010. The genetic landscape of a cell. Science 327, 425–431.
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.
Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B., Boone, C., 2009. Systematic mapping of genetic interaction networks. Annu. Rev. Genet. 43, 601–625.
Enright, A.J., Iliopoulos, I., Kyrpides, N.C., Ouzounis, C.A., 1999. Protein interaction maps for complete genomes based on gene fusion events. Nature 402, 86–90.
Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., Taylor, R., Dharsee, M., Ho, Y., Heilbut, A., Moore, L., Zhang, S., Ornatsky, O., Bukhman, Y.V., Ethier, M., Sheng, Y., Vasilescu, J., Abu-Farha, M., Lambert, J.P., Duewel, H.S., Stewart II, , Kuehl, B., Hogue, K., Colwill, K., Gladwish, K., Muskat, B., Kinach, R., Adams, S.L., Moran, M.F., Morin, G.B., Topaloglou, T., Figeys, D., 2007. Large-scale mapping of human protein–protein interactions by mass spectrometry. Mol. Syst. Biol. 3, 89.
Fraser, A.G., Marcotte, E.M., 2004. A probabilistic view of gene function. Nat. Genet. 36, 559–564.
Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dumpelfeld, B., Edelmann, A., Heurtier, M.A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J.M., Kuster, B., Bork, P., Russell, R.B., Superti-Furga, G., 2006. Proteome survey reveals modularity of the yeast cell machinery. Nature 440, 631–636.
Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley Jr., R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J., Rothberg, J.M., 2003. A protein interaction map of *Drosophila melanogaster*. Science 302, 1727–1736.
Hutchins, J.R., Toyoda, Y., Hegemann, B., Poser, I., Heriche, J.K., Sykora, M.M., Augsburg, M., Hudecz, O., Buschhorn, B.A., Bulkescher, J., Conrad, C., Comartin, D., Schleiffer, A., Sarov, M., Pozniakovsky, A., Slabicki, M.M., Schloissnig, S., Steinmacher, I., Leuschner, M., Ssykor, A., Lawo, S., Pelletier, L., Stark, H., Nasmyth, K., Ellenberg, J., Durbin, R., Buchholz, F., Mechtler, K., Hyman, A.A., Peters, J.M., 2010. Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science 328, 593–599.
Huynen, M., Snel, B., Lathe 3rd, W., Bork, P., 2000. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. 10, 1204–1210.
Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y., 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. U S A 98, 4569–4574.
Jenssen, T.K., Laegreid, A., Komorowski, J., Hovig, E., 2001. A literature network of human genes for high-throughput analysis of gene expression. Nat. Genet. 28, 21–28.
Jeong, H., Mason, S.P., Barabasi, A.L., Oltvai, Z.N., 2001. Lethality and centrality in protein networks. Nature 411, 41–42.
Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A., 2002. The KEGG databases at GenomeNet. Nucleic Acids Res. 30, 42–46.
Kelley, R., Ideker, T., 2005. Systematic interpretation of genetic interactions using protein networks. Nat. Biotechnol. 23, 561–566.
Kim, E., Shin, J., Lee, I., 2010. Assessment of effectiveness of the network-guided genetic screen. Mol. BioSyst 6, 1803–1806.

Kim, H.D., Shay, T., O'Shea, E.K., Regev, A., 2009. Transcriptional regulatory circuits: predicting numbers from alphabets. Science 325, 429–432.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., Punna, T., Peregrin-Alvarez, J.M., Shales, M., Zhang, X., Davey, M., Robinson, M.D., Paccanaro, A., Bray, J.E., Sheung, A., Beattie, B., Richards, D.P., Canadien, V., Lalev, A., Mena, F., Wong, P., Starostine, A., Canete, M.M., Vlasblom, J., Wu, S., Orsi, C., Collins, S.R., Chandran, S., Haw, R., Rilstone, J.J., Gandi, K., Thompson, N.J., Musso, G., St Onge, P., Ghanny, S., Lam, M.H., Butland, G., Altaf-Ul, A.M., Kanaya, S., Shilatifard, A., O'Shea, E., Weissman, J.S., Ingles, C.J., Hughes, T.R., Parkinson, J., Gerstein, M., Wodak, S.J., Emili, A., Greenblatt, J.F., 2006. Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. Nature 440, 637–643.

Lalonde, S., Ehrhardt, D.W., Loque, D., Chen, J., Rhee, S.Y., Frommer, W.B., 2008. Molecular and cellular approaches for the detection of protein–protein interactions: latest techniques and current limitations. Plant J. 53, 610–635.

Lee, I., Ambaru, B., Thakkar, P., Marcotte, E.M., Rhee, S.Y., 2010a. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. Nat. Biotechnol. 28, 149–156.

Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M., 2004. A probabilistic functional network of yeast genes. Science 306, 1555–1558.

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G., Marcotte, E.M., 2008. A single gene network accurately predicts phenotypic effects of gene perturbation in *Caenorhabditis elegans*. Nat. Genet. 40, 181–188.

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G., Marcotte, E.M., 2010b. Predicting genetic modifier loci using functional gene networks. Genome Res. 20, 1143–1153.

Lee, I., Li, Z., Marcotte, E.M., 2007. An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*. PLoS One 2, e988.

Lee, I., Narayanaswamy, R., Marcotte, E.M., 2006. Bioinformatic prediction of yeast gene function. In: Stansfield, I. (Ed.), Yeast Gene Analysis. Elsevier Press.

Lehner, B., Crombie, C., Tischler, J., Fortunato, A., Fraser, A.G., 2006. Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. Nat. Genet. 38, 896–903.

Lehner, B., Fraser, A.G., 2004. A first-draft human protein-interaction map. Genome Biol. 5, R63.

Lehner, B., Lee, I., 2008. Network-guided genetic screening: building, testing and using gene networks to predict gene function. Brief. Funct. Genomic. Proteomic. 7, 217–227.

Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E., Vidal, M., 2004. A map of the interactome network of the metazoan *C. elegans*. Science 303, 540–543.

Li, Z., Lee, I., Moradi, E., Hung, N.J., Johnson, A.W., Marcotte, E.M., 2009. Rational extension of the ribosome biogenesis pathway using network-guided genetics. PLoS Biol. 7, e1000213.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., Visscher, P.M., 2009. Finding the missing heritability of complex diseases. Nature 461, 747–753.

Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., Eisenberg, D., 1999. Detecting protein function and protein–protein interactions from genome sequences. Science 285, 751–753.

McGary, K.L., Lee, I., Marcotte, E.M., 2007. Broad network-based predictability of *Saccharomyces cerevisiae* gene loss-of-function phenotypes. Genome Biol. 8, R258.

Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., Weil, B., 2002. MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30, 31–34.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. U S A 96, 2896–2901.

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

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J., Hoe, K.L., Kim, D.U., Ideker, T., Grewal, S.I., Weissman, J.S., Krogan, N.J., 2008. Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. Science 322, 405–410.

Roguev, A., Wiren, M., Weissman, J.S., Krogan, N.J., 2007. High-throughput genetic interaction mapping in the fission yeast *Schizosaccharomyces pombe*. Nat. Methods 4, 861–866.

Rual, J.F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P., Vidal, M., 2005. Towards a proteome-scale map of the human protein–protein interaction network. Nature 437, 1173–1178.

Simonis, N., Rual, J.F., Carvunis, A.R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F., Cevik, S., Klitgord, N., Fan, C., Braun, P., Li, N., Ayivi-Guedehoussou, N., Dann, E., Bertin, N., Szeto, D., Dricot, A., Yildirim, M.A., Lin, C., de Smet, A.S., Kao, H.L., Simon, C., Smolyar, A., Ahn, J.S., Tewari, M., Boxem, M., Milstein, S., Yu, H., Dreze, M., Vandenhaute, J., Gunsalus, K.C., Cusick, M.E., Hill, D.E., Tavernier, J., Roth, F.P., Vidal, M., 2009. Empirically controlled mapping of the *Caenorhabditis elegans* protein–protein interactome network. Nat. Methods 6, 47–54.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Stapley, B.J., Benoit, G., 2000. Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac. Symp. Biocomput., 529–540.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksoz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., Wanker, E.E., 2005. A human protein–protein interaction network: a resource for annotating the proteome. Cell 122, 957–968.

Tischler, J., Lehner, B., Fraser, A.G., 2008. Evolutionary plasticity of genetic interaction networks. Nat. Genet. 40, 390–391.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M., 2000. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403, 623–627.

Uren, A.G., Kool, J., Matentzoglu, K., de Ridder, J., Mattison, J., van Uitert, M., Lagcher, W., Sie, D., Tanger, E., Cox, T., Reinders, M., Hubbard, T.J., Rogers, J., Jonkers, J., Wessels, L., Adams, D.J., van Lohuizen, M., Berns, A., 2008. Large-scale mutagenesis in p19(ARF)- and p53-deficient mice identifies cancer genes and their collaborative networks. Cell 133, 727–741.

Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S., Koonin, E.V., 2001. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome Res. 11, 356–372.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabasi, A.L., Tavernier, J., Hill, D.E., Vidal, M., 2008. High-quality binary protein interaction map of the yeast interactome network. Science 322, 104–110.

Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M., Gerstein, M., 2004. Annotation transfer between genomes: protein–protein interologs and protein-DNA regulogs. Genome Res. 14, 1107–1118.