Integrating 'omic' information: a bridge between genomics and systems biology

Hui Ge[1], Albertha J.M. Walhout[1,2] and Marc Vidal[1]

[1]Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School, SM858, 44 Binney Street, Boston, MA 02115, USA
[2]Current address: University of Massachusetts Medical School, Program in Gene Function and Expression, 364 Plantation Street, Worcester, MA 01605, USA

The availability of genome sequences for several organisms, including humans, and the resulting first-approximation lists of genes, have allowed a transition from molecular biology to 'modular biology'. In modular biology, biological processes of interest, or modules, are studied as complex systems of functionally interacting macromolecules. Functional genomic and proteomic ('omic') approaches can be helpful to accelerate the identification of the genes and gene products involved in particular modules, and to describe the functional relationships between them. However, the data emerging from individual omic approaches should be viewed with caution because of the occurrence of false-negative and false-positive results and because single annotations are not sufficient for an understanding of gene function. To increase the reliability of gene function annotation, multiple independent datasets need to be integrated. Here, we review the recent development of strategies for such integration and we argue that these will be important for a systems approach to modular biology.

Since the advent of molecular biology, biological questions have been approached mainly by studying the function(s) of individual genes and gene products, one or a few at a time. This reductionist approach proved to be extremely fruitful, leading to the discovery of an impressive number of biological principles. Despite the considerable success of molecular biology, many fundamental biological questions remain unanswered. Most importantly, because the majority of gene products function together with other gene products, biological processes should be considered as complex networks of interconnected components. In other words, for any biological process, one might consider a 'modular approach' in which the behavior and function of the corresponding network are studied as a whole, in addition to studying some of its components individually [1].

The availability of complete genome sequences [2–6], along with gene predictions [7], has resulted in the development of technologies that allow the assignment of genes to particular biological modules. For example, standardized high-throughput (HT) assays have been developed to analyze the transcriptome and the proteome of model organisms and humans. Other more recent functional genomic and proteomic ('omic') approaches include protein–protein, protein–DNA or other 'component–component' interaction mapping (interactome mapping), systematic phenotypic analyses (phenome mapping) and transcript or protein localization mapping (localizome mapping). Omic approaches have already been applied to many biological processes, leading to large lists of genes potentially involved in the corresponding modules (for detailed reviews see Refs [8–13]. One important advance was the development and implementation of computational methods by which genes or proteins that behave similarly under various experimental conditions can be grouped [14–18].

All omic approaches have intrinsic caveats. For example, information can be missing because of the occurrence of false negatives, and information can be misleading because of the presence of false positives. Thus, data obtained from any single omic approach should be interpreted cautiously [10,19–22]. In addition, data emerging from any single omic approach can only provide crude indications of gene or protein function. It has been proposed that these limitations can be overcome by integrating data obtained from two or more distinct approaches [19,23]. Such integration should not only improve functional annotations but also help to formulate biological hypotheses (Fig. 1). For example, an interaction network of proteins whose genes are similarly expressed under various experimental conditions and show overlapping loss-of-function phenotypes is more likely to be relevant *in vivo* than any other network for which this additional information is not available. In addition, the coexpression and phenotypic patterns might indicate functional and dynamic aspects of the corresponding network.

In this review, we discuss recent investigations of the relationships between datasets obtained using distinct omic approaches and the use of such integrated data to improve the analysis of biological systems (Fig. 2). In summary, we will review strategies for investigating the

Corresponding author: Marc Vidal (marc_vidal@dfci.harvard.edu).

 0168-9525/$ - see front matter doi:10.1016/j.tig.2003.08.009

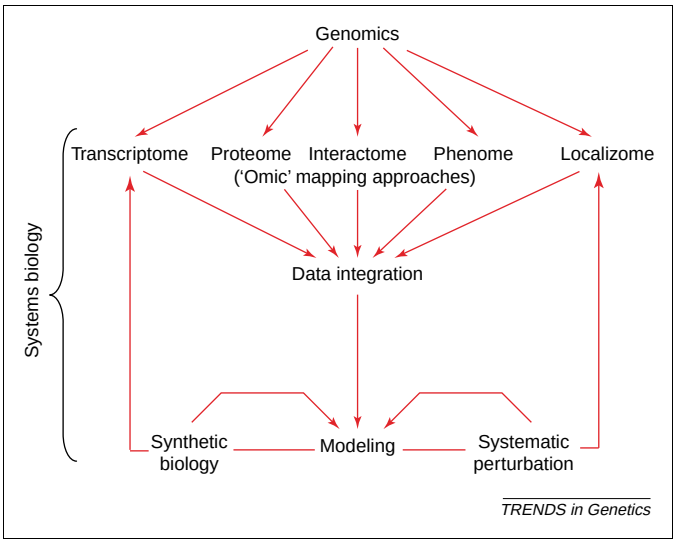

Fig. 1. Integrating 'omic' information. With the availability of complete genome and transcriptome sequences (genomics), functional genomic and proteomic (or 'omic') approaches are used to map the transcriptome (complete set of transcripts), proteome (complete set of proteins), interactome (complete set of interactions), phenome (complete set of phenotypes) and localizome (localization of all transcripts and proteins [64]) of a given organism. Integrating omic information should help to reduce the problems caused by false positives and false negatives obtained from single omic approaches, lead to better functional annotations for gene products and functional relationships between them, and allow the formulation of increasingly relevant biological hypotheses. Computational methods can then be used to model biological processes based on integrated data. The resulting models can be tested either by 'synthetic biology' (*de novo* design and generation of biological modules based on suspected network properties) or by systematic perturbations, or both. Systems biology strategies can thus be viewed as a combination of omic approaches, data integration, modeling and synthetic biology.

potential relationships between datasets obtained from different omic approaches and how biological hypotheses can be formulated based on the integrated data.

The basics of omics

Transcriptome profiling was one of the first omic approaches to be developed [8]. Using microarrays [24], DNA chips [25] or serial analysis of gene expression (SAGE) [26], the relative abundance of transcripts can be monitored simultaneously for thousands of genes under various experimental conditions. Transcriptome profiling experiments can be used to identify genes that are potentially involved in particular modules. For example, by sporulating yeast cells and recording the transcriptome profiles, transcripts that are upregulated or downregulated could be identified and the corresponding genes postulated to function in the sporulation module [27,28]. On a genome-wide scale, combining data from several unrelated expression profiling experiments can result in more detailed and informative module assignments. This was first demonstrated by profiling nearly all *Saccharomyces cerevisiae* transcripts under ≈300 different experimental and genetic conditions and integrating the data into a transcriptome 'compendium' [29]. A similar compendium was generated for *Caenorhabditis elegans* by combining data from more than 550 different microarray experiments [30]. Clustering algorithms were used to group genes with similar expression profiles, and these groups were visualized as 'mountains' in a 'topomap'.

Interaction networks that describe modules of interest can be obtained by systematically identifying protein–protein, protein–DNA or protein–RNA interactions. The yeast two-hybrid system (Y2H) [31] was among the first methods to be adapted for HT protein–protein interaction mapping. Typically, HT Y2H can be performed at a module scale by using most or all proteins already known to function in this module as 'baits' and identifying new binding partners [10]. Examples of module-scale interactome maps include those for *S. cerevisiae* splicing factors, RNA polymerase III and Sm-like proteins [32–34], and those for *C. elegans* proteins involved in vulval development, the 26S proteasome and the DNA damage response (DDR) [17,35,36]. Interactome mapping has also been extended to the whole *S. cerevisiae* proteome [37,38] and the data obtained have been used to reconstitute modular and proteome-scale interaction networks. Another approach to identify protein–protein interactions that has been adapted to an HT format involves the combination of immunoprecipitations with mass spectrometry (IP-MS). This method has been applied at a proteome scale for *S. cerevisiae* [39,40]. The overlap between HT Y2H and IP-MS proteome-scale interactome datasets is relatively small at this stage [41], suggesting that both approaches have intrinsic caveats and generate false negatives and/or false positives.

The phenome can be defined as a collection of phenotypic information observed upon perturbations of large numbers of genes. *S. cerevisiae* deletion mutants have been generated by homologous recombination for ≈96% of the predicted open reading frames (ORFs) [42]. Using this resource, ≈1500 genes have been identified as essential for viability [43] and numerous nonessential genes have been found to be required for particular modules such as nonhomologous end joining [44], sporulation [45] and damage recovery [46]. In *C. elegans*, hypomorphic phenotypes can be generated by a method referred to as RNA-mediated interference (RNAi) [47]. Using HT RNAi, phenotypes such as lethality and sterility have been scored at a genome scale [18,48–51].

Phenome data can be organized in two-dimensional matrices comprised of large numbers of genes on the one hand, and sets of loss-of-function phenotypes on the other. As mentioned above, it is possible to use clustering methods similar to those used in expression profiling to group genes according to their loss-of-function phenotypes. The resulting 'phenoclusters' can provide information about both the involvement of genes in particular modules and the functional relationships that might exist between them [17,18,43,46].

Global correlations between omic data

It is not immediately apparent how relative mRNA abundance obtained from transcriptome profiling can be integrated with the binary all-or-none information of protein–protein interactions obtained from interactome mapping. As a first attempt, several groups, using different strategies, investigated the potential relationships between such distinct omic datasets [20–22,52–55] (Fig. 3). Two types of transcriptome profiling datasets were used for these investigations: clusters of coexpressed genes across subsets of particular conditions [20–22] and

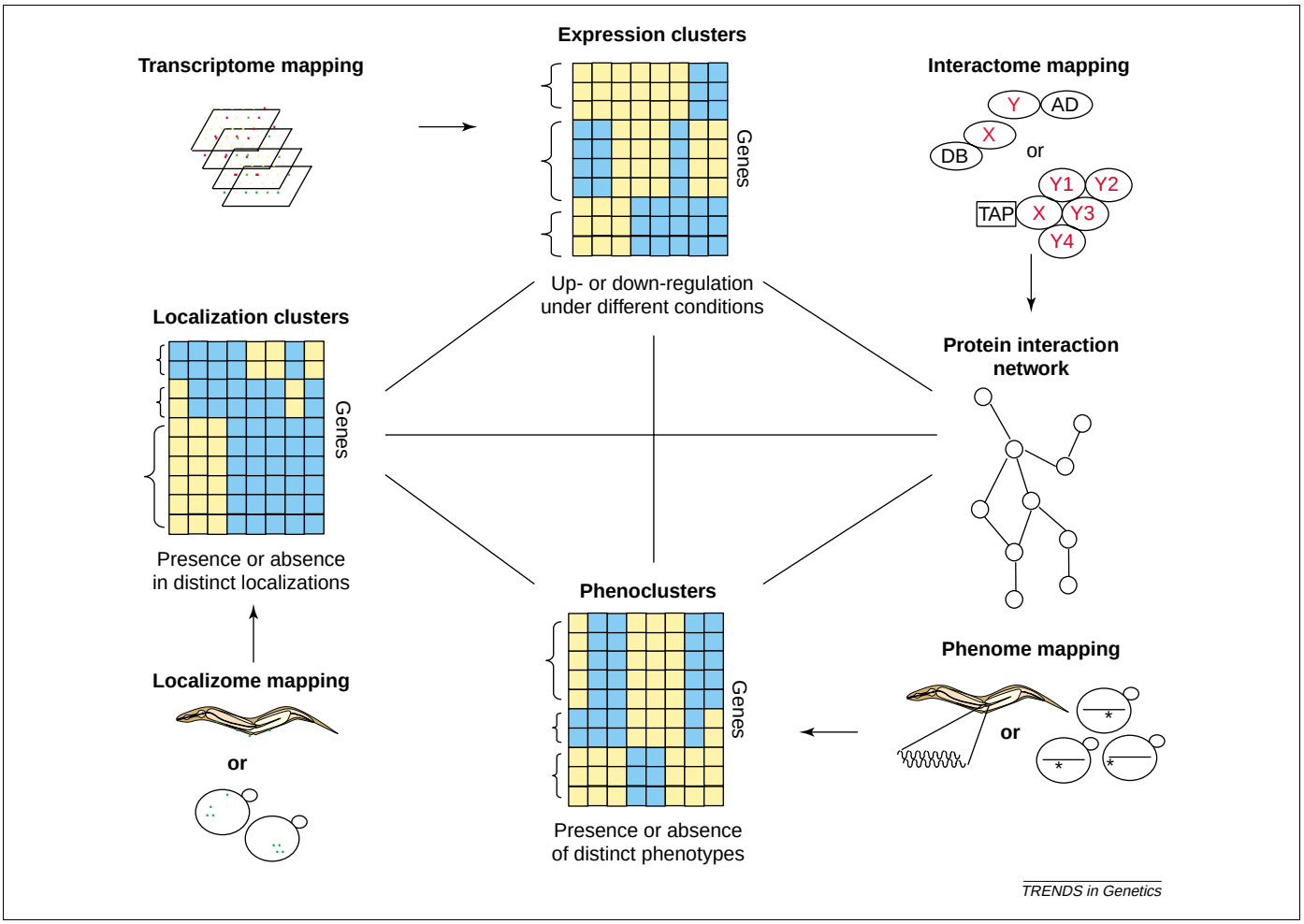

Fig. 2. Integration of transcriptome, interactome, phenome and localizome data. Functional genomic and proteomic data generated by various high-throughput (HT) approaches can be organized into clusters or networks. Transcriptome, interactome, phenome and localizome data can be integrated either two by two or all at once. Examples are shown for *Saccharomyces cerevisiae* and *Caenorhabditis elegans*.

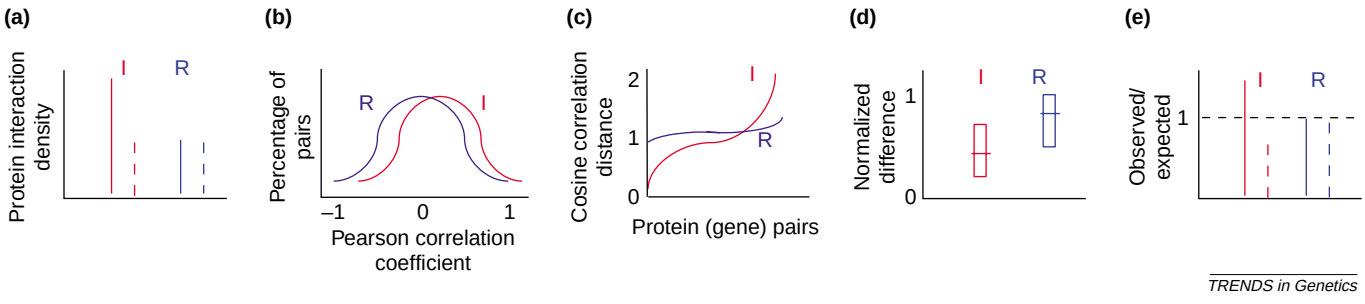

Fig. 3. Strategies developed to integrate transcriptome and interactome datasets. Different indices have been used to correlate protein interaction datasets with expression profiles. In each panel, the red line represents an interactome dataset (I) and the blue line represents a random dataset (R). (a) Intracluster protein interaction density (PID) (solid lines) versus intercluster PID (dashed lines) [20–22]. PID is defined as the ratio of the number of protein–protein interaction pairs observed over the total number of possible pairwise combinations. The intracluster pairs refer to combinations of proteins encoded by genes belonging to common expression clusters, and the intercluster pairs refer to combinations of proteins encoded by genes belonging to different clusters. The average intracluster PID is significantly higher than that of intercluster PID for an interactome dataset, whereas average intracluster and intercluster PIDs are similar for a set of random protein pairs. (b) Distributions of Pearson correlation coefficients [52]. On average, the Pearson correlation coefficients of transcript abundance corresponding to interacting protein pairs are significantly higher (which means better correlation) for interactome datasets than for sets of random protein pairs. (c) Distributions of cosine correlation distance [55]. On average, the cosine correlation distances of transcript abundance corresponding to protein pairs are significantly lower (which means better correlation) for interactome datasets than for sets of random protein pairs. (d) Distributions of normalized difference in box-plot representation [54]. The box and the line through the box depict the distribution and median of normalized differences, respectively. The normalized difference is a measure of similarity of abundance of a pair of transcripts. On average, the normalized differences of transcript abundance corresponding to protein pairs are significantly lower (which means higher similarity) for interactome datasets than for sets of random protein pairs. (e) The ratios of observed to expected number of protein pairs whose mRNA expressions are positively correlated (solid lines) or negatively correlated (dashed lines) [53]. For interactome datasets, the observed expected ratio for positively correlated pairs is significantly higher than 1, whereas this ratio for negatively correlated pairs is significantly lower than 1. The ratios are both very close to 1 for sets of random protein pairs. This indicates a higher percentage of positively correlated pairs in interactome datasets than in random sets.

transcriptome compendia [52–55]. Three types of interactome datasets were used: (1) large collections of potential protein–protein interactions obtained from proteome-wide HT Y2H or IP-MS projects (HT data); (2) collections of individual protein–protein interactions identified by reductionist approaches [low-throughput (LT) data] [56,57]; and (3) randomly generated protein pairs.

In a study of omic data integration, HT, LT and combined HT-LT yeast interactome datasets were compared with three modular transcriptome datasets related to cell cycle, sporulation or environmental stresses [20–22]. First, a protein interaction density (PID) value was calculated as the ratio of the number of observed protein–protein interaction pairs over the total number of possible pairwise combinations for a given set of gene products. PIDs were compared between sets of protein pairs encoded by genes belonging to common transcriptome clusters (or 'intracluster' pairs) and sets of protein pairs encoded by genes belonging to different clusters (or 'intercluster' pairs). In general, average intracluster PIDs are significantly greater than intercluster PIDs for interactome datasets, whereas the average intracluster and intercluster PIDs are similar for sets of random protein pairs (Fig. 3a). Also, LT interactome data give larger PIDs than HT data [22]. This observation indicates a global correlation between transcriptome and interactome mapping data.

Although seemingly obvious, there are many exceptions to the notion that proteins that function together are encoded by genes that share similar expression profiles. For example, cyclins and cyclin-dependent kinases (CDKs) are known to interact with each other during the cell cycle, even though their transcripts are not found in similar transcriptome clusters. Although clustering of two genes in expression profiles increases the likelihood that their corresponding proteins might interact, a lack of correlation in transcriptome profiles does not necessarily rule out interactions between proteins.

Expression profile compendia were also used to investigate a possible correlation between transcriptome and interactome data. Various indices were developed to compare the overall correlation of expression between pairs of genes corresponding to apparently interacting proteins with that of random gene pairs. Significant differences are observed when the distribution of these indices is plotted and protein–protein interaction datasets are compared with random protein pairs (Fig. 3b–e) [52–55]. HT datasets were found to lie between LT and random datasets, confirming the occurrence of both genuine interactions and false-positive information in HT data. These observations indicate that interacting proteins are more likely to be encoded by genes that share similar expression profiles. This was subsequently used to assess the reliability of the protein interaction dataset detected by HT methods [58].

The observations described above suggest that transcriptome and interactome data can be correlated for the unicellular *S. cerevisiae*. For multicellular organisms, such a correlation could be obscured by the fact that, although pairs of genes encoding potentially interacting proteins might show similar expression profiles, their transcripts or proteins can localize to different parts of the body. Recent experiments suggest that correlations between the transcriptome and interactome can be extended to *C. elegans*, at least when a particular tissue is considered [59]. A module-scale interactome map was obtained for proteins encoded by germline-enriched transcripts, and the data were compared to the transcriptome compendium described above [30]. Gene pairs corresponding to interacting proteins were more likely to belong to common expression clusters than would be expected from a random distribution. It remains to be investigated whether such transcriptome–interactome correlation can be extended to other *C. elegans* tissues, to the whole animal, or to other multicellular organisms.

Global relationships have also been examined for other pairwise combinations of omic data, such as interactome and phenome. The relationship between interactome and phenome mapping data has been investigated at both module and genome scales. Just as can be observed for transcriptome and interactome data, one can hypothesize that genes that share similar loss-of-function phenotypes are more likely to encode interacting proteins than random gene pairs. In a module-scale study, HT Y2H and RNAi analyses were combined to identify genes involved in the DNA damage response (DDR) in *C. elegans* [17]. Proteins corresponding to potential *C. elegans* orthologs of DDR genes in other organisms were subjected to Y2H interactome screens and potential protein–protein interactions were identified. Subsequently, HT Y2H interactors were subjected to RNAi and examined for DDR-related phenotypes. Approximately 10% of the interactors tested exhibited DDR-related phenotypes and two phenoclusters were established, distinguishing potential functions in either DNA damage checkpoints or DNA repair. A majority of the interactions identified belong to the intracluster category, whereas a few are intercluster, which provided supporting evidence for a correlation between the interactome and the phenome.

In a separate study, an interactome map was generated by the HT Y2H approach for the *C. elegans* germline [59]. When phenome data were integrated with interactome data, it was found that proteins encoded by genes that exhibit an embryonic lethal phenotype in HT RNAi assays were more likely to interact with each other than with nonessential proteins. This example also indicated that interactome and phenome mapping data can be correlated.

Correlation between genome-scale interactome and phenome data has recently been described for *S. cerevisiae*. In a Y2H protein–protein interaction network, proteins with large numbers of potential interaction partners, or 'hubs', tend to be essential [60,61]. Such a correlation between the centrality of location in an interactome network and the necessity of function might have important implications for our understanding of the proteome as a whole.

As discussed above, genes whose mRNA levels change in response to particular experimental conditions can be identified by transcriptome mapping, whereas genes essential for a particular process can be identified by phenome mapping. To gain a better understanding for a given biological module, it is important to investigate

potential correlations between the transcriptome and the phenome and what can be learned from such correlations.

The relationship between transcriptome and phenome datasets has been explored in the context of particular modules. In a study aimed at identifying yeast genes involved in sporulation, about half of 261 genes required for sporulation [45] belong to a set of ≈1000 genes (<20% of all yeast genes) found to be transcriptionally regulated during sporulation [28]. The statistical significance of this finding suggests that transcriptome and phenome data can be correlated. However, it should be pointed out that in a different attempt to integrate transcriptome and phenome datasets, the correlation was less clear. Deleting genes whose mRNA abundance changes in response to exposure to mutagens was no more likely to result in mutagen sensitivity than deleting nonresponsive genes [46]. In other words, the fact that expression of a gene is responsive to mutagen exposure was not predictive of whether or not it contributes to recovery from the mutagen. One possibility to reconcile these different observations is that potential correlations between transcriptome and phenome data might be module-specific.

Recently, HT RNAi was performed for ≈86% of predicted genes of *C. elegans*. It was reported that genes that exhibit nonviable RNAi phenotypes (lethality or sterility) are enriched in two mountains in the transcriptome topomap [51]. This suggests that genes whose loss-of-functions result in similar phenotypes tend to be coregulated, which can be viewed as a transcriptome–phenome correlation at a genome-wide scale.

Taken together, global and module-scale correlations have been detected between various combinations of omic data. In the next section, we review how biological hypotheses can be formulated based on such integrated omic information.

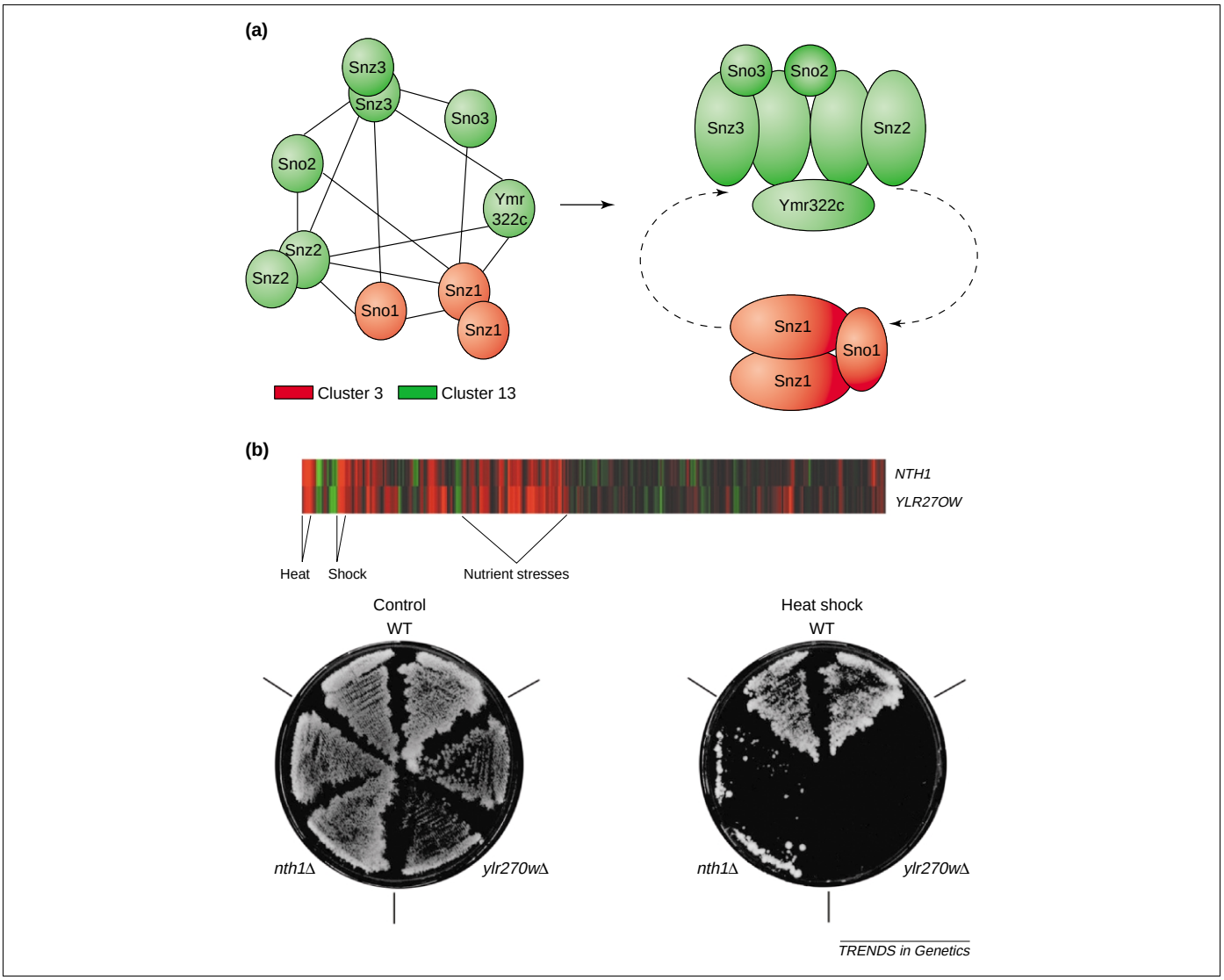

Fig. 4. Formulation of hypotheses based on integrated omic data. (a) An interaction network of Snz–Sno proteins of *Saccharomyces cerevisiae*. The nodes represent proteins and the lines represent yeast two-hybrid (Y2H) interactions. The red nodes represent proteins that correspond to genes in one transcriptome cluster, whereas the green nodes represent proteins that correspond to genes belonging to a different cluster. The existence of two stable complexes can be hypothesized based on the integrated data. Reproduced, with permission, from [20] (http://www.nature.com) using data from [21,22]. (b) The genes *NTH1* and *YLR270W* have similar expression profiles (upper panel). Red indicates upregulation and green indicates downregulation. mRNA expressions of both genes are upregulated during heat shock and other forms of stress. Deletions of *NTH1* and *YLR270W* each confer similar heat-shock sensitive phenotypes (lower panel). Reproduced, with permission, from Ref. [55].

Formulation of biological hypotheses based on integrated omic data

In modular-scale approaches, the observation that the transcripts corresponding to a pair of interacting proteins belong to a common expression cluster is considered as supporting evidence that the interaction might be genuine [20–22]. One example was provided in the context of the Snz–Sno complex implicated in stress resistance [20–22,62] (Fig. 4a). In the Y2H HT interactome map, a network of interacting Snz–Sno proteins can be observed, which suggests the existence of a multiprotein complex. This interactome information was integrated with transcriptome

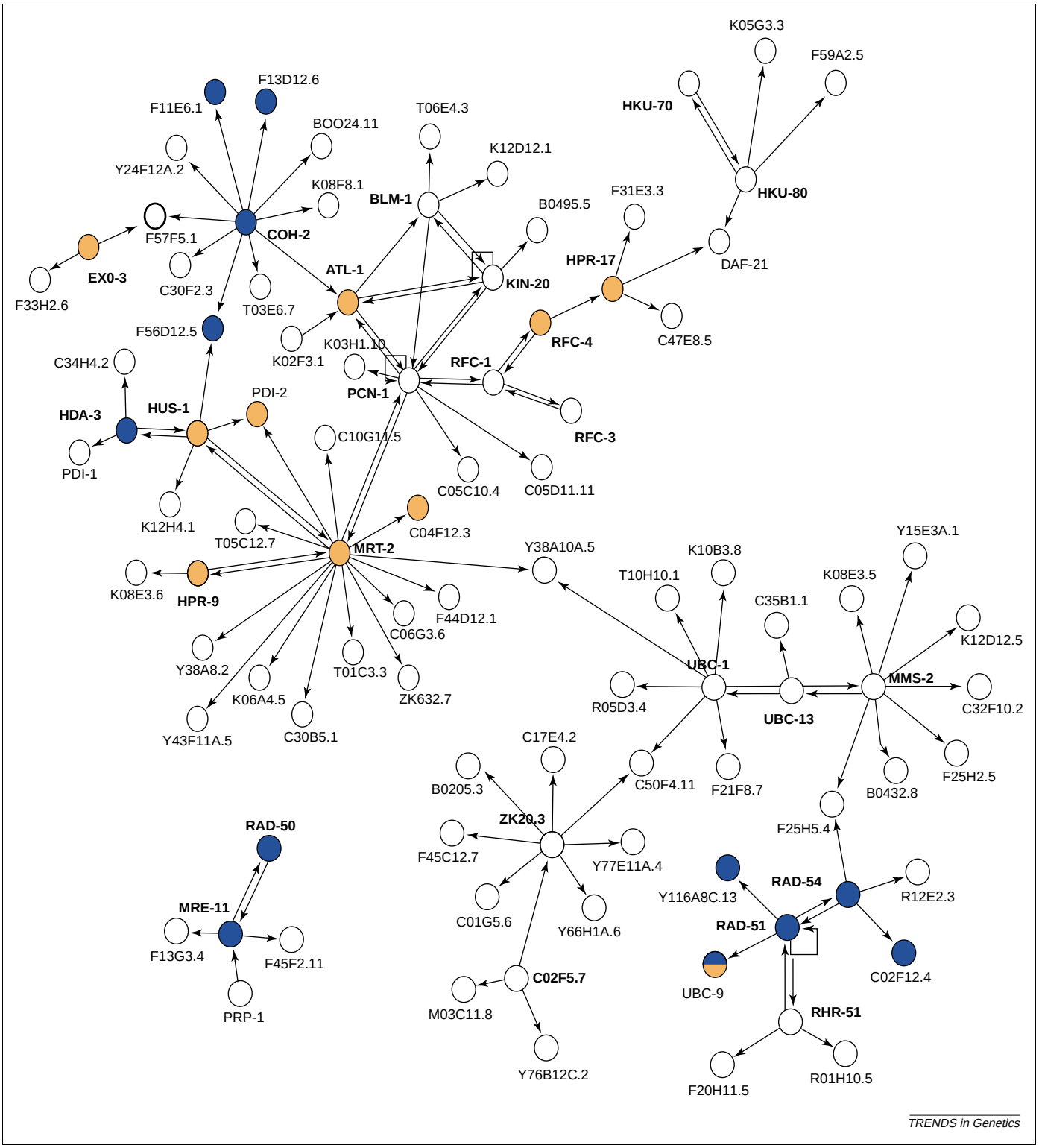

Fig. 5. Part of a DNA damage response (DDR) protein interaction map for *Caenorhabditis elegans*. Arrows represent yeast two-hybrid (Y2H) interactions and nodes (circles) represent proteins. Blue and orange nodes indicate products of genes from the DNA repair and checkpoint phenoclusters, respectively. *mrt-2* and C04F12.3 belong to a common phenocluster, and their products physically interact in the Y2H system. Reproduced, with permission, from Ref. [17] ©2002 American Association for the Advancement of Science.

information obtained from exposing yeast cells to various stress conditions. Because two potential subunits reside in a single expression cluster, whereas the other five subunits reside in another expression cluster, one can hypothesize the existence of two stable complexes. Thus, the Y2H interactions between the two potential complexes might represent false positives or, alternatively, the multiprotein complex might exist transiently, because the clustering analysis does not preclude partial overlaps of expression.

Likewise, transcriptome and interactome correlations established using either Pearson correlation coefficients or cosine correlation distances can be used to make functional predictions for individual proteins [55]. For example, a potential interaction between Nth1p and the product of YLR-270W can be found in the Y2H HT dataset for *S. cerevisiae*. The two corresponding genes show a relatively high degree of transcriptome correlation (cosine correlation distance = 0.17) (Fig. 4b). Because *NTH1* is known to function in thermotolerance as well as resistance to other forms of stress, YLR270W could be hypothesized to mediate a similar function. Indeed, phenotypic analyses revealed that deletion of either gene results in very similar inabilities to recover from heat shock treatments, which supports the functional annotations for *YLR270W*.

Interactome–phenome correlation analyses also suggest that a combination of these two approaches can help to formulate hypotheses. In an integrated interactome and phenome map of the *C. elegans* DDR described above [17], relatively strong functional predictions could be made based on the protein–protein interactions whose corresponding genes share similar RNAi phenotypes. For example, the product of C04F12.3, an uncharacterized gene, was identified as an HT Y2H interactor of the checkpoint protein MRT-2 (Fig. 5). In addition, C04F12.3 exhibited a checkpoint defective phenotype in HT RNAi similar to that of *mrt-2*. Thus, C04F12.3 is very likely to mediate functions related to checkpoint integrity. Altogether, 12 worm DDR orthologs and 11 novel DDR genes were annotated using this integrated approach. Damage recovery modules for *S. cerevisiae* have also been studied by interactome and phenome integration [46] (Fig. 6).

Just as expression clusters can be used to find putative protein–protein interactions of higher confidence, it will be valuable to determine to what extent more refined phenoclusters – that is, those generated with large numbers of scored phenotypes – will be useful to rank potential protein–protein interactions. For example, time-lapse differential interference contrast (DIC) microscopy was performed to score 47 distinct phenotypes during *C. elegans* early embryogenesis for 161 genes that give an embryonic lethal HT RNAi phenotype, and phenoclusters were obtained based on the digitized data [18]. These phenoclusters were used to interrogate the relative likelihood of biological relevance of Y2H interactions identified for the *C. elegans* germline module. In a few cases, striking similarities were observed between the two phenotypic profiles of genes encoding potentially interacting proteins [59], greatly increasing the confidence of functional relationships between these pairs of genes.

Further analysis of transcriptome and phenome data can also lead to the formulation of hypotheses. For example, the genes known by genetic means to be essential for yeast sporulation revealed two subgroups [45] (see above). In one subgroup, expression was not responsive to sporulation and most of these genes were found to encode general cellular factors. In the other subgroup, genes were transcriptionally responsive to the sporulation process and many of those were found to encode proteins specifically required for sporulation [45]. Therefore, uncharacterized genes that are both essential for and transcriptionally responsive to sporulation can be hypothesized to be sporulation-specific factors.

Integration of multiple omic approaches can also be performed based on the principles and methodologies established for the integration of pairs of omic datasets. In a recent study that combined transcriptome, interactome and phenome datasets for the *C. elegans* germline, it was demonstrated that genes encoding pairs of interacting proteins show a relatively high likelihood of exhibiting similar RNAi phenotypes and expression profiles [59]. These interactions were proposed to have a relatively high likelihood of relevance to germline biology (Fig. 7). With the increasing availability of various omic maps, more studies are likely to employ multidimensional integration of omic data. Over time, it should become increasingly clear whether a global correlation of omic datasets applies to different systems and modules and how biological hypotheses can be formulated based on data integration.

It is very important to consider the visualization aspect for an optimal implementation of information integration. Currently, several bioinformatic tools are being developed to provide visualization of integrated omic data. For example, expression correlation of two genes encoding potentially interacting proteins can be visualized in web-accessible protein–protein interaction networks (http://vidal.dfci.harvard.edu/interactomedb/interactome4.pl) by coloring coregulated genes and gene products found by transcriptome profiles [59]. The correlation can also be viewed by aligning their respective individual expression profiles across many different conditions, and the conditions under which their expression correlation is most apparent can be extracted [55,59]. Finally, because omic datasets are being constantly updated, visualization tools should allow the incorporation of constantly evolving data.

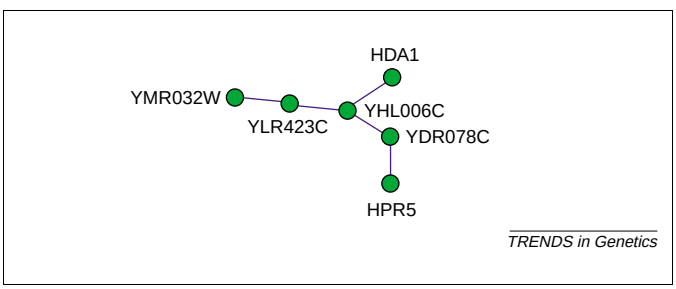

Fig. 6. A protein interaction network identified by interactome–phenome integration might reveal a damage recovery module in *Saccharomyces cerevisiae*. The nodes represent proteins and the lines represent yeast two-hybrid (Y2H) interactions. All of the six proteins were identified as required for methyl methanesulfonate (MMS) resistance in phenotypic analyses. Reproduced, with permission, from Ref. [46] (http://www.aacr.org).

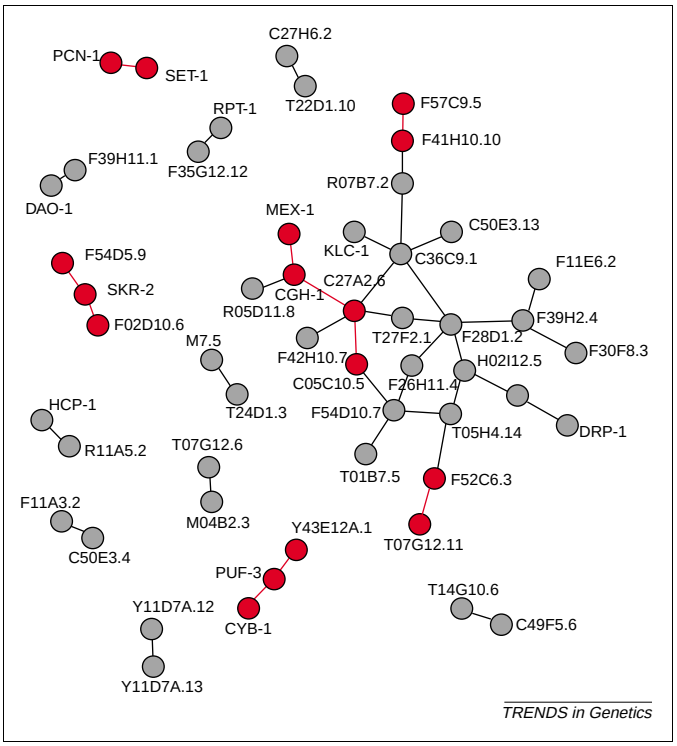

Fig. 7. A core germline interactome map for *Caenorhabditis elegans*. The nodes represent proteins and the lines represent yeast two-hybrid (Y2H) interactions. The interactome map was integrated with transcriptome and phenome data. Red lines indicate interactions between proteins whose corresponding genes have both similar expression profiles and overlapping RNA interference (RNAi) phenotypes. Reproduced, with permission, from Ref. [59].

Perspectives: from integration of omic data to systems biology

So far integration of omic data has been applied mainly to improve functional annotations of individual genes, to evaluate the likelihood of putative protein–protein interactions, or to identify components potentially involved in specific modules. We propose that such data integration can be further applied to examine the topology of biological networks, to provide information on directionality of interactions, and to create wiring diagrams that better depict the functional outcome of component–component relationships. Together, these strategies should facilitate a systems approach to modular biology.

Systems biology can be approached by perturbing the suspected components of a given cellular process, monitoring the responses, integrating the data and modeling the biological process in question [63]. Omic data integration might gradually become indispensable for systems biology (Figs 1,2,8). By applying a single omic approach, the knowledge of a system can be expanded from a single gene to a network of genes, which can be regarded as a basic model for the system. When genes or proteins in this network are systematically disrupted, responses from other parts of the network can be recorded and the data obtained can be incorporated into the basic model. A wiring diagram that depicts the direction of interactions in the network can be constructed to better represent the relationships between the components. The simple example shown in Fig. 8 illustrates how our current limited knowledge of a biological system can be expanded and a model built based on integrated omic information.

Real-life biological systems might contain more components and the wiring diagrams that depict the relationships between these components could be much more complex than currently appreciated. Also, the information available for biological systems is increasing as more omic datasets become available. Thus, HT data integration is needed in systems biology approaches, which should be achieved by the use of computational tools that apply the principles and methodologies discussed here to multiple sets of omic information in a dynamic manner. The biological networks or wiring diagrams modeled in this manner should shed light on the complexity of biological systems.

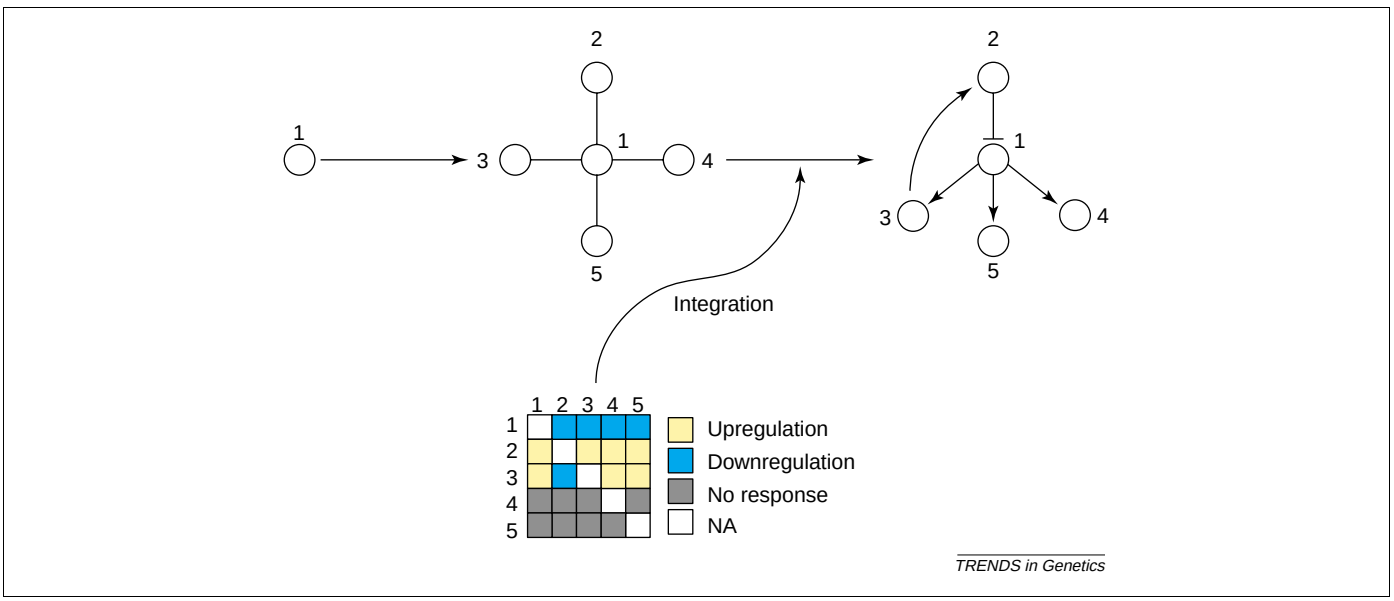

Fig. 8. A systems biology strategy. A simplified systems biology strategy is illustrated. Starting from one known component, more components involved in a module of interest can be identified, for example, by interactome mapping. A network can be constructed to describe these interactions. Perturbation experiments are then systematically performed and responses from the rest of the network are recorded, for example, by transcriptome profiling. The data can be incorporated back into the network and a wiring diagram can be constructed that more clearly depicts the relationships between the components. Abbreviation: NA, not analyzed.

Acknowledgements

We thank C. Armstrong, G. Cottarel, J. Dekker, B. Deplancke, D. Hill and J. Vandenhaute for critical reading and discussion of the manuscript. Work in this laboratory is supported by grants from NCI, NHGRI and NIGMS awarded to M.V.

References

1 Hartwell, L.H. *et al.* (1999) From molecular to modular cell biology. *Nature* 402, C47–C52
2 Blattner, F.R. *et al.* (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474
3 Goffeau, A. *et al.* (1997) The yeast genome directory. *Nature* 387 (6632 suppl.), 5
4 The C. elegans sequencing consortium, (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282, 2012–2018
5 Adams, M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287, 2185–2195
6 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
7 Mathe, C. *et al.* (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30, 4103–4117
8 Lockhart, D.J. and Winzeler, E.A. (2000) Genomics, gene expression and DNA arrays. *Nature* 405, 827–835
9 Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* 405, 837–846
10 Walhout, A.J.M. and Vidal, M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell Biol.* 2, 55–62
11 Sternberg, P.W. (2001) Working in the post-genomic *C. elegans* world. *Cell* 105, 173–176
12 Hope, I.A. (2001) Broadcast interference – functional genomics. *Trends Genet.* 17, 297–299
13 Zhu, H. and Snyder, M. (2002) 'Omic' approaches for unraveling signaling networks. *Curr. Opin. Cell Biol.* 14, 173–179
14 Eisen, M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95, 14863–14868
15 Tamayo, P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U. S. A.* 96, 2907–2912
16 Brown, M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector. *Proc. Natl. Acad. Sci. U. S. A.* 97, 262–267
17 Boulton, S.J. *et al.* (2002) Combined functional genomic maps of the *C. elegans* DNA damage response. *Science* 295, 127–131
18 Piano, F. *et al.* (2002) Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. *Curr. Biol.* 12, 1959–1964
19 Vidal, M. (2001) A biological atlas of functional maps. *Cell* 104, 333–339
20 Ge, H. *et al.* (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.* 29, 482–486
21 Mrowka, R. *et al.* (2003) Does mapping reveal correlation between gene expression and protein–protein interaction? *Nat. Genet.* 33, 15–16
22 Ge, H. *et al.* (2003) Reply to 'Does mapping reveal correlation between gene expression and protein–protein interaction?'. *Nat. Genet.* 33, 16–17
23 Walhout, A.J.M. *et al.* (1998) A model of elegance. *Am. J. Hum. Genet.* 63, 955–961
24 Schena, M. *et al.* (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470
25 Lockhart, D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* 14, 1675–1680
26 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487
27 Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705
28 Primig, M. *et al.* (2000) The core meiotic transcriptome in budding yeasts. *Nat. Genet.* 26, 415–423
29 Hughes, T.R. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109–126
30 Kim, S.K. *et al.* (2001) A gene expression map for *Caenorhabditis elegans*. *Science* 293, 2087–2092
31 Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* 340, 245–246
32 Fromont-Racine, M. *et al.* (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16, 277–282
33 Flores, A. *et al.* (1999) A protein–protein interaction map of yeast RNA polymerase III. *Proc. Natl. Acad. Sci. U. S. A.* 96, 7815–7820
34 Fromont-Racine, M. *et al.* (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast* 17, 95–110
35 Walhout, A.J.M. *et al.* (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science* 287, 116–122
36 Davy, A. *et al.* (2001) A protein–protein interaction map of the *Caenorhabditis elegans* 26S proteasome. *EMBO Rep.* 2, 821–828
37 Uetz, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* 403, 623–627
38 Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4569–4574
39 Gavin, A.C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415, 141–147
40 Ho, Y. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* 415, 180–183
41 Grunenfelder, B. and Winzeler, E.A. (2002) Treasures and traps in genome-wide data sets: case examples from yeast. *Nat. Rev. Genet.* 3, 653–661
42 Winzeler, E.A. *et al.* (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285, 901–906
43 Giaever, G. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391
44 Ooi, S.L. *et al.* (2001) A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science* 294, 2552–2556
45 Deutschbauer, A.M. *et al.* (2002) Parallel phenotypic analysis of sporulation and postgermination growth in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15530–15535
46 Begley, T.J. *et al.* (2002) Damage recovery pathways in *Saccharomyces cerevisiae* revealed by genomic phenotyping and interactome mapping. *Mol. Cancer Res.* 1, 103–112
47 Fire, A. *et al.* (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* 391, 806–811
48 Fraser, A.G. *et al.* (2000) Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* 408, 325–330
49 Gonczy, P. *et al.* (2000) Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* 408, 331–336
50 Maeda, I. *et al.* (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol.* 11, 171–176
51 Kamath, R.S. *et al.* (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421, 231–237
52 Grigoriev, A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast *Saccharomyces cerevisiae*. *Nucleic Acids Res.* 29, 3513–3519
53 Mrowka, R. *et al.* (2001) Is there a bias in proteome research? *Genome Res.* 11, 1971–1973
54 Jansen, R. *et al.* (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.* 12, 37–46
55 Kemmeren, P. *et al.* (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell* 9, 1133–1143
56 Mewes, H.W. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* 30, 31–34
57 Costanzo, M.C. *et al.* (2001) YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* 29, 75–79
58 Deane, C.M. *et al.* (2002) Protein interactions: two methods for assessment of reliability of high throughput observations. *Mol. Cell. Proteomics* 1, 349–356

59 Walhout, A.J.M. *et al.* (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.* 12, 1952–1958
60 Jeong, H. *et al.* (2001) Lethality and centrality in protein networks. *Nature* 411, 41–42
61 Oltvai, Z.N. and Barabasi, A.L. (2002) Systems biology. Life's complexity pyramid. *Science* 298, 763–764
62 Padilla, P.A. *et al.* (1998) The highly conserved, coregulated Sno and Snz gene families in *Saccharomyces cerevisiae* respond to nutrient limitation. *J. Bacteriol.* 180, 5718–5726
63 Ideker, T. *et al.* (2001) A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* 2, 343–372
64 Kumar, A. *et al.* (2002) Subcellular localization of the yeast proteome. *Genes Dev.* 16, 707–719