



Human Molecular Genetics

Interactome: gateway into systems biology

Michael E. Cusick, Niels Klitgord, Marc Vidal and David E. Hill
Hum. Mol. Genet. 14:171-181, 2005. First published 14 Sep 2005;
doi:10.1093/hmg/ddi335

The full text of this article, along with updated information and services is available online at http://hmg.oxfordjournals.org/cgi/content/full/14/suppl_2/R171

References

This article cites 134 references, 54 of which can be accessed free at http://hmg.oxfordjournals.org/cgi/content/full/14/suppl_2/R171#BIBL

Cited by

This article has been cited by 14 articles at 25 August 2008 . View these citations at http://hmg.oxfordjournals.org/cgi/content/full/14/suppl_2/R171#otherarticles

Reprints

Reprints of this article can be ordered at http://www.oxfordjournals.org/corporate_services/reprints.html

Email and RSS alerting

Sign up for email alerts, and subscribe to this journal's RSS feeds at <http://hmg.oxfordjournals.org>

**PowerPoint®
image downloads**

Images from this journal can be downloaded with one click as a PowerPoint slide.

Journal information

Additional information about Human Molecular Genetics, including how to subscribe can be found at <http://hmg.oxfordjournals.org>

Published on behalf of

Oxford University Press
<http://www.oxfordjournals.org>

Interactome: gateway into systems biology

Michael E. Cusick^{1,*}, Niels Klitgaard¹, Marc Vidal^{1,2} and David E. Hill^{1,2}

¹Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, USA and ²Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Received August 18, 2005; Revised and Accepted September 1, 2005

Protein–protein interactions are fundamental to all biological processes, and a comprehensive determination of all protein–protein interactions that can take place in an organism provides a framework for understanding biology as an integrated system. The availability of genome-scale sets of cloned open reading frames has facilitated systematic efforts at creating proteome-scale data sets of protein–protein interactions, which are represented as complex networks or ‘interactome’ maps. Protein–protein interaction mapping projects that follow stringent criteria, coupled with experimental validation in orthogonal systems, provide high-confidence data sets immanently useful for interrogating developmental and disease mechanisms at a system level as well as elucidating individual protein function and interactome network topology. Although far from complete, currently available maps provide insight into how biochemical properties of proteins and protein complexes are integrated into biological systems. Such maps are also a useful resource to predict the function(s) of thousands of genes.

SYSTEMATIC MAPPING OF INTERACTOME NETWORKS

Most gene products mediate their function within complex networks of interconnected macromolecules. Studies in model organisms suggest that complex macromolecular networks have topological and dynamic properties that reflect biological phenomena (1,2). Thus, an understanding of biological mechanisms and disease processes demands a ‘systems’ approach that goes beyond one-at-a-time studies of single components to more global analyses of the structure, function and dynamics of the networks in which macromolecules function.

We consider the full interactome network as the complete collection of all physical protein–protein interactions that can take place within a cell. Construction of comprehensive sets of protein–protein interactions, interactomes, requires the creation of genome-scale resource collections of open reading frames (ORFeomes) cloned so as to facilitate protein expression, generated iteratively based on improved gene predictions and experimental verification and capturing all expressed isoforms (splice variants and polymorphisms). ORFeomes, as faithful representations of the encoded proteome, provide the starting material for carrying out high-throughput interaction studies that are then validated by orthogonal interaction methods. The resulting interactome maps are regarded as ‘framework’ information; and by integrating

other functional genomic and proteomic data sets, increasingly detailed and reliable biological models can be generated (3).

Model organisms have provided the basis for a systematic characterization of physical protein–protein interactions (‘interactome’ mapping). Initial efforts focused on defined biological processes or ‘modules’ for the yeast *Saccharomyces cerevisiae* and the worm *Caenorhabditis elegans* (4,5). Subsequently, proteome-scale interactome mapping projects for eukaryotes have been carried out in yeast, worm and fly (6–10). Current estimates for the complete yeast interactome suggest ~28 000 potential protein interactions, on the basis of experimental and computational analyses (6,7,11–15) along with incorporating literature-curated interactions such as those collected in the MIPS databases (16). So far, the worm and fly interactome maps each contain approximately 5000 high-quality putative interactions derived primarily from high-throughput yeast two-hybrid (Y2H) screens (8–10). These two data sets demonstrate the feasibility of interactome mapping projects for metazoans, and they also illustrate the power of integrating multiple approaches to model biological networks (17). However, to fully understand human biology and the molecular mechanisms underlying diseases such as cancer, systematic experimental mapping of the human interactome itself is necessary.

Although completed genome sequences provide lists of tens of thousands of predicted unique proteins (~25 000 for the

*To whom correspondence should be addressed. Tel: +1 6176323802; Fax: +1 6176325739; Email: Michael_cusick@dfci.harvard.edu; marc_vidal@dfci.harvard.edu; david_hill@dfci.harvard.edu

human proteome, disregarding splice variants and post-translational modifications), the sequences by themselves do not provide an understanding of the underlying principles of cellular systems. Proteome-scale information is also required at structural, functional and dynamic levels. This information should encompass various molecular networks, such as regulatory, biochemical or protein–protein interaction networks. The initial challenge is the generation of comprehensive network maps, generally depicted as nodes (e.g. proteins, RNAs, DNA binding sites or metabolites) linked by edges corresponding to molecular interactions (e.g. protein–protein interactions, enzymatic reactions, DNA–protein, etc.). For each network map, individual nodes and edges need to be perturbed systematically to help in understanding the logic of molecular networks involved in any biological processes of interest. As biological systems are highly dynamic and fluid, information on where and when nodes appear or disappear on where and when edges take place and on the rewiring of the network, as sub-networks appear or disappear during developmental and cell cycle stages, needs to be obtained.

Here, we review recent progress in interactome mapping, emphasizing the need for high-confidence, experimentally derived data sets to drive the construction and use of these maps as frameworks for integrating other genome-scale information, such as genetic interactions, expression profiling and phenotypic analyses.

FUNCTIONALITY AND MULTIFUNCTIONALITY

A significant hindrance to a comprehensive understanding of human biology, encompassing both the individual parts and the integrated whole, is the limited information available for most human genes, beyond the completed DNA sequence of the euchromatic portion of the genome (18). For example, in the most recent compilation of 25 356 human genes listed at NCBI, only 15 088 contain a predicted PfamA domain (19), leaving ~40% with no annotated functional element. In contrast, in *S. cerevisiae* nearly all genes have an assigned Gene Ontology (GO) term (20), with most of these GO annotations arising from experimental, not predicted, assignments. However, functional annotation assigned solely on the presence of conserved domain signatures has severe limitations, because many predicted domains have no or limited functional annotation(s), and because most proteins can contain multiple domains. The situation becomes even more complicated for proteins that have multiple functions and multiple classes of interacting partners, the multifunctional ‘moonlighting proteins’ (21). The problems with multifunctional proteins are that the apparent main function of the protein may not explain an observed phenotype, activity, or genetic or physical interaction; thus, determining all the multiple functions of a protein or its interacting partners may be elusive.

STANDARDIZED ORF COLLECTIONS

The nearly 250 complete genome sequences (18 eukaryotic genomes plus 230 microbial genomes) and the 234 eukaryotic

genome sequences in progress or nearing completion, as of June 1, 2005, constitute an enormous, albeit admittedly substantially untapped, wealth of biological information (<http://www.ncbi.nlm.nih.gov/Genomes/>). However, for nearly all sequenced genomes, the exact gene count as well as the precise exon–intron structure for each protein-coding gene remains incomplete, and with available approaches perhaps indeterminable, even for the well-annotated *C. elegans* and *S. cerevisiae* genomes (8,22–25). Consequently, concerted efforts to experimentally verify every protein-coding gene in the genome are still required (26). Large-scale cDNA cloning efforts have provided significant number of genes for further manipulation (27–35). However, full-length cDNAs are generally not immediately suitable for protein expression, mainly because of the presence of 5′ and 3′ untranslated regions flanking the ORF (26). For protein expression at proteome scale, systematic and comprehensive cloning of full-length ORFs, ORFeome projects as opposed to cDNA cloning projects, is specifically required, particularly for eukaryotic organisms. Furthermore, ORFeome cloning using high-throughput systems such as Gateway recombinational cloning creates resource collections that can be readily manipulated as ‘standardized parts’ for any desired expression system (26). ORFeome projects are underway and ongoing for several eukaryotic organisms, yeast, worm, plants, mouse and humans (8,27,29,36–42). ORFeome cloning also provides valuable experimental verification of gene predictions, especially for those gene predictions with minimal or no prior evidence (8). Although ORFeome cloning projects obviously in and of themselves do not establish gene function, they are indispensable for subsequent proteome-scale investigations that do lead to functional annotation (26).

PROTEINS, COMPLEXES AND NETWORKS

A common theme pervading biological investigation is that most proteins generally function as components of complexes that contain other macromolecules to carry out specific biological processes, and networks of interactions connect multiple, different cellular processes (43–46). In these two concepts are embedded several implications. First, if the function of any one protein is known, then identification of its interacting partners will predict function for some or all of the partners, the ‘guilt-by-association’ principle (47). Second, any given protein may participate in multiple complexes, each with a discrete function (48,49). Third, communication between processes involves protein–protein interactions that connect complexes (2,50). Fourth, the function of a novel protein complex can be predicted following the ‘majority rule’ principle, whereby the most common function held by a majority of members defines the overall function of the complex (50–53), and by extension, the function of any unknown component. Fifth, biological networks exhibit emergent properties that are best understood after many or most network connections are identified (54–57). Sixth, reliance on incomplete annotation or incomplete networks can lead to biased or possibly erroneous conclusions (57–60).

With 30% or more of human genes lacking functional annotation, and with just a few thousand human genes well characterized, one pressing need is to obtain comprehensive and unbiased data sets of all potential binary and complex membership interactions. Currently, two experimental methodologies are used for generating genome-scale protein interaction maps at high-throughput. They are high-throughput yeast two-hybrid (HT-Y2H) (61–63) and analysis of protein complexes by affinity purification and mass spectrometry (AP–MS) (64,65). Yeast two-hybrid (Y2H), as a binary assay captures direct protein–protein interactions, whereas AP–MS identifies components of stable complexes. Both assays individually can provide useful information on protein function by employing guilt-by-association and majority rule principles. Some information on the dynamic nature of interactions can be obtained when Y2H and AP–MS are combined (66) or when interaction data is supplemented by data from expression profiling and phenotypic analyses (3,17).

An alternative to experimental determination of protein interactions is prediction by various computational genomics approaches (67–70). Computational genomics utilizes information on individual protein interactions taken from publicly available databases such as BIND (71), MIPS (72) and HPRD (73,74), relies on annotations defined by Pfam (19), GO (20), and combines these data with sequence similarities within genomes and orthologies across genomes for *in silico* prediction of protein–protein interactions (75–80).

IDENTIFICATION OF PROTEIN INTERACTIONS BY YEAST TWO HYBRID

The Y2H system, as originally described by Fields and Song (81), and the many derivatives of it that have been developed (62), is the most commonly utilized system for identifying binary protein–protein interactions, particularly at proteome scale (6,7,9,10,82–84). A key feature of Y2H is that it can be employed as an unbiased investigation into potential binary protein–protein interactions.

The canonical Y2H system consists of a separable, DNA binding domain (DB) from a transcriptional activator protein (yeast Gal4 or bacterial *lexA*) fused to protein ‘X’, generally referred to as the ‘bait’, and a separable transcriptional activation domain (AD) fused to protein ‘Y’, termed the ‘prey’. When DB-X and AD-Y are co-expressed in the nucleus of yeast cells, X-Y protein–protein interactions reconstitute a functional transcription factor that activates one or more reporter genes (Fig. 1A). Two outstanding virtues of Y2H are: (i) DNA, not protein, is manipulated to study both bait and prey (85) so ORFeome resources are readily employed (8,83); and (ii) it is readily adapted to high-throughput methods (61,83,85–88).

Standard Y2H generally underestimates the number of interactions, because forced subcellular localization of bait and prey in the yeast nucleus may preclude certain interactions from taking place, a particular instance being interactions involving integral membrane proteins. For interactions that require specific post-translational modifications, unless the enzymes responsible for such modification happen to be present in the yeast nucleus, the interaction may not be

detectable by Y2H. For these reasons and others, the canonical Y2H has an inherent false-negative rate that limits the number of potential protein–protein interactions (62). However, alternative two-hybrid systems (62,89) can be employed for classes of proteins refractory to the standard Y2H. Although none of these alternative systems have yet been adapted for proteome-scale use, the availability of cloned ORFeomes will facilitate the eventual shift to high-throughput.

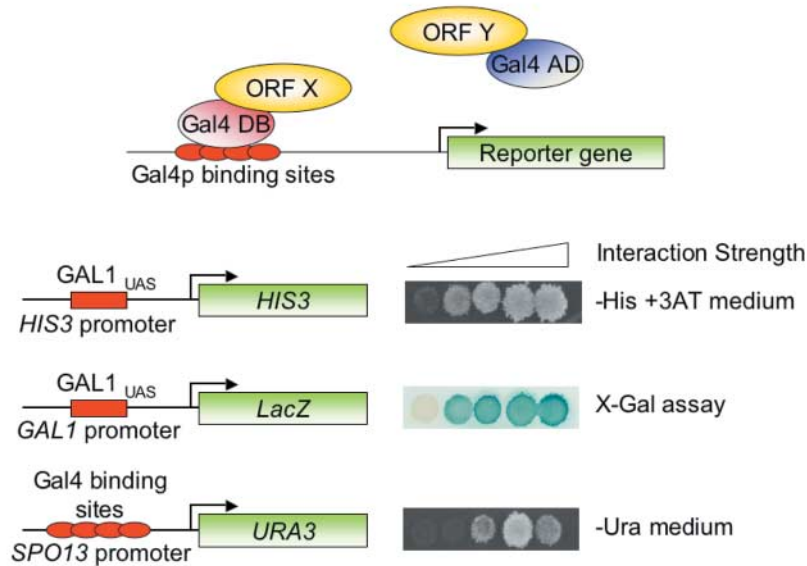
TECHNICAL FALSE-POSITIVES IN Y2H

Both experimental and computational methods for identifying protein–protein interactions will exhibit some degree of false-positives. We consider false-positives to be of two distinct classes. Biological false-positives are those in which the interaction can be confirmed by multiple, different methods, but the two proteins are never present in the same cell or subcellular compartment at the same time. These false-positives are nearly impossible to unequivocally identify using interaction assays alone. The second class is technical false-positives that can occur in any experimental system. Early HT-Y2H experiments might have been compromised by a high rate of technical false-positives owing to specific features of the particular Y2H system available at that time. In later and ongoing HT-Y2H experiments the technical false-positive rate is substantially reduced by, among other improvements, incorporating multiple reporter genes to measure transcription activation, employing different DNA sequences for binding by DB in the promoters of the reporter genes, using low copy number vectors, and, importantly, retesting interacting pairs in fresh yeast (62,83,85,90–92). In addition, any two-hybrid system based on transactivation of reporter genes, not just the Y2H, will experience ‘auto-activators’, where the DB-X construct activates gene expression in the absence of any AD-Y (Fig. 1B). Strong auto-activators can be removed directly before any AD-Y is added (93). Additional auto-activators arise owing to acquisition of mutations in the bait during propagation of bait-containing yeast cells. These latent auto-activators are much harder to identify, as the presence of AD-Y gives the appearance of an interaction when in fact it is the DB-X construct alone that auto-activates the Y2H reporter genes, irrespective of any AD-Y that is present (62).

The initial genome-wide Y2H studies contained significant technical false-positives (86), most likely because the influence of auto-activation was not recognized. The high proportion of false-positives does not necessarily diminish the impact of these studies (86,94,95), but does reflect the challenges faced in moving from highly focused, reductionist approaches to global approaches that attempt to interrogate entire genomes and proteomes in an unbiased way that can capture all potential interactions.

Several computational methods have been developed in an attempt to reduce the impact of DB-X auto-activators. Auto-activators appear as ‘sticky’ or promiscuous baits that have many interaction partners, which often do not share any common functional annotation. Therefore, criteria such as cut-offs for maximal number of interactors for a given bait protein or common functional annotation among

A High Specificity Version of Yeast Two Hybrid system



B Auto-activators in Y2H: source of false positives

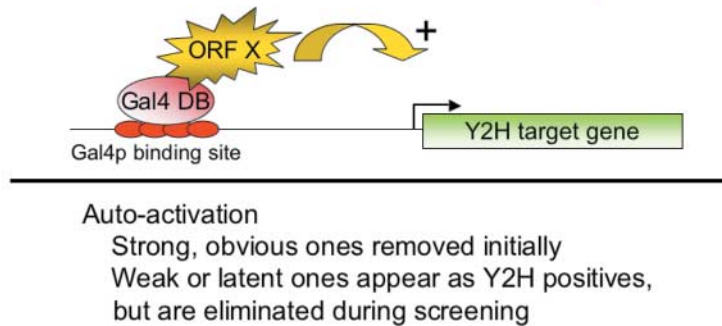


Figure 1. Stringent Y2H screening strategy. (A) Through use of multiple, single-copy reporter genes, low copy plasmids for expression of bait and prey in yeast, and retesting of all positives, Y2H achieves increased stringency leading to reproducibly real interactions. (B) Removal of auto-activators. Auto-activating baits are the major source of false-positives. Strong auto-activators can be removed prior to addition of prey plasmids. Latent ones can arise during manipulation of yeast; testing for latent auto-activators is performed in parallel with Y2H screening.

interactors (9,82,96–98) have been used to eliminate putative auto-activators. This strategy reduces the size and complexity of the resulting interactome network, as some of the baits or preys with large numbers of interactors or interactors that do not share common functional annotation may still represent real interactions. Alternatively, auto-activators can be, and should be, removed experimentally by first testing for prey-independent growth on selective media to remove strong auto-activators (93), and by identifying and removing ‘latent’ auto-activators during the screening process (83,91,99). Once technical false-positives are removed, the quality of the resulting data set is significantly improved (10,83,100).

THOUSANDS OF INTERACTIONS: ON THE WAY TO PROTEOME-SCALE INTERACTOMES

Large-scale Y2H screens for *Helicobacter pylori* (82), *S. cerevisiae* (6,7), *Drosophila melanogaster* (9), *C. elegans*

(10) and *Homo sapiens* (63,83) have produced thousands of interaction pairs. However, for each species the results covered only a small fraction of the expected number of interactions. The two separate studies with *S. cerevisiae* identified over 3000 potential protein–protein interactions, but these constitute only 10–15% of the expected total (101,102). Likewise, the worm interactome containing over 4000 interactions comprises less than 5% of the expected total and similarly for the fly interactome (54).

The two groups that conducted HT-Y2H for yeast used the same ~6000 ORFs as baits, but there was less than 15% overlap in detected interaction pairs (6,7). The low overlap has raised the concern that Y2H is inherently ‘noisy’ (98,102). Even more, each of the two screens had a less than 13% overlap with a literature set of high-confidence interactions compiled from single-gene analyses (2,7). Although the genome-wide *Drosophila* screen reported ~5000 high-quality interactions (9), two smaller, focused

screens for *Drosophila* showed minimal overlap with this larger study (103,104), echoing the situation seen for the two yeast studies. This low overlap suggests that high-throughput screens should be repeated more than once to recover the maximum amount of interaction pairs.

It is assumed that the published literature on protein interactions encompasses the most comprehensive and best-annotated information available. However, the various curated data sets compiled from the literature show limited overlap with each other (2,74,83,105). The limited overlap among various data sets is likely due to several factors, including uneven sampling, high false-negative rates in experimental and computational protocols, and irreproducibility within and between different experimental and computational approaches (74).

VALIDATION OF Y2H BY ORTHOGONAL ASSAYS

Although many 'interesting and novel' interacting pairs have been obtained in the various global Y2H screens, any particular result should be viewed with caution until validated by another distinct method. For the yeast, bacterial and fly screens, computational methods were used to gather sets of high-confidence interactions (96,101,106). Although computational methods can qualify potential interactions, still direct tests of bait/prey interaction in an independent, orthogonal assay are required.

For the *C. elegans* and human HT-Y2H studies, stringent criteria were imposed before and during the screen (10,83). By rigorously eliminating auto-activators and by systematically retesting all interaction pairs, experimentally derived high-confidence Y2H data sets were obtained. To further validate Y2H interaction pairs experimentally, a co-affinity purification (co-AP) strategy was used, taking advantage of the respective *C. elegans* and human ORFeome collections (Fig. 2) (8,83). The orthogonal co-AP assay requires that the interacting pairs, expressed in mammalian cells (as opposed to yeast nuclei with the Y2H), form stable complexes that can be isolated and analyzed by immunoassay. With both the worm and human data sets, the collection of high-confidence Y2H interactions showed an ~80% validation rate by co-AP (10,83). A recent effort at examining protein complexes biochemically based on the *H. pylori* Y2H data set achieved a comparable cross-validation rate (107). Thus, experimental orthogonal approaches demonstrate that these HT-Y2H interaction data sets contain mostly highly reliable interactions.

IDENTIFICATION OF INTERACTING PROTEINS BY MASS SPECTROMETRY

There are two basic strategies for determination of protein membership in a complex, direct (purification of a stable complex and elucidation of the components of the complex by mass spectrometry) and co-AP (purification of a complex by virtue of an affinity tag placed on one of its components, then elucidation of the components of the complex by mass spectrometry). Direct analysis can identify novel members of

stable complexes (108,109) but faces the difficult task of achieving sufficient purification of target complexes without loss of components and with minimal contamination. A protein purified with one organelle or complex but demonstrated to be normally associated with another organelle might represent a contaminant, or might actually represent a protein involved in intracellular communication between organelles and complexes (108). For example, a recent AP-MS study of human Par proteins confirmed previous interactions and identified over 50 novel interactions, some of which are between distinct complexes (110). The identification of increasing numbers of shared components involving complexes of different function (108) highlights the challenge to assigning function based on co-purification strategies. Improved proteomics coverage of complexes and organelles, coupled with unbiased Y2H studies to identify possible binary interactions, is likely to validate more 'shared components' and the ensuing network of interactions as well as provide evidence for functional annotation.

In two distinct studies of the yeast proteome by AP-MS (11,12) there was limited overlap among the common complexes identified, and the reproducibility was approximately 70%. Interestingly, 98 complexes were previously characterized, whereas 134 complexes were novel (11), which demonstrates the ability of AP-MS to identify novel interactions. The fact that interactions previously characterized in the processes of DNA damage response and signal transduction were not all recovered could be indicative of an experimental or technical bias in the basic methodology (11). Alternatively, because many complexes involve very transient interactions and/or individual components are not readily detectable owing to low expression, AP-MS will underestimate the extent of complex co-membership. The lack of overlap between the two studies, along with the identification of specific proteins shared among distinct complexes, has also raised the issue of false-positives. However, a systematic analysis suggests that a majority of novel and shared components are likely to be biologically relevant (108), which means that AP-MS is a reliable method for identifying novel components of complexes and for assigning putative function based on guilt-by-association and majority rule criteria.

COMPLEMENTARITY OF AP-MS AND Y2H

AP-MS presumably identifies interactions that occur in the native cellular environment, provided that temporal and spatial expression of the baits is normal, although purification of complexes can lead to both loss of real interactions and gain of spurious ones. In contrast, with Y2H all interactions occur in the heterologous environment of the yeast nucleus. With AP-MS expression of proteins in their normal cell/tissue may allow identification of post-translational modifications, but the heterologous expression in Y2H generally precludes this. However, AP-MS may miss weak, transient associations, whereas Y2H is a better choice for such interactions (86,92). Y2H identifies specific, binary protein-protein interactions, which are not readily identified in AP-MS of complexes. Y2H is amenable to high-throughput methods for identifying mutations that

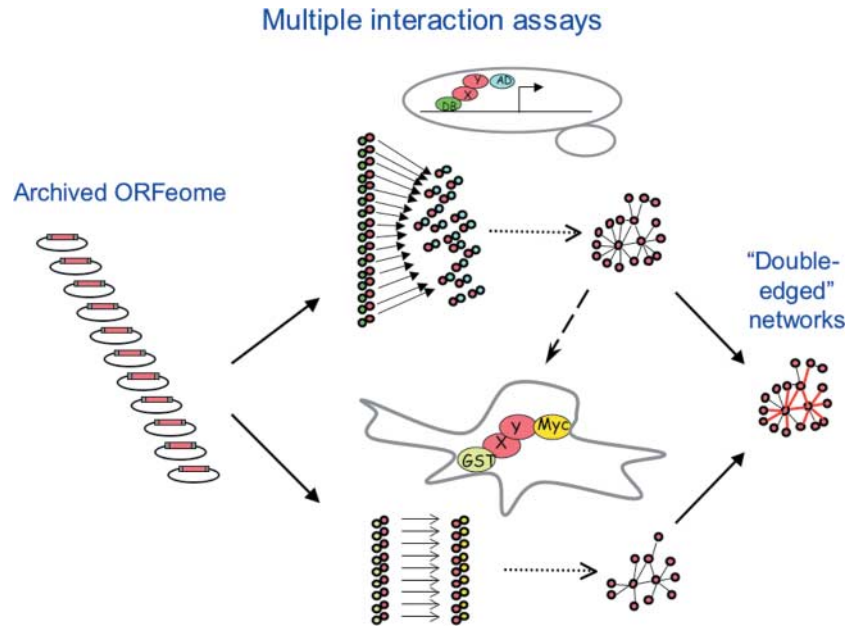


Figure 2. ORFeome resources used for interaction mapping: ORFeome collections are first used in HT-Y2H to generate a set of protein–protein interactions. ORF clones corresponding to all unique bait and prey from Y2H are then utilized in an orthogonal assay, co-AP from 293T cells to validate Y2H interactions (10). Adapted with permission from Brasch *et al.* (26).

disrupt interactions (90,111) and for interrogating protein domains (86), whereas AP–MS is less suitable for these two functions. These comparisons demonstrate that comprehensive identification of all the individual protein–protein interactions that collectively make up all interactions within any given complex requires both Y2H and AP–MS approaches (66). Integration of data from the two approaches can also serve to increase confidence in either data set, and provide further evidence of functional annotation for unknown proteins.

MULTIFUNCTIONALITY: REALITY DISGUISED AS FALSE-POSITIVES

As described already, technical false-positives can be eliminated from high-throughput screens by rigorous testing and validation. However, even when technical false-positives are fully removed there can still remain interactions that are reproducibly real but do not appear to make sense based on existing functional annotation of either or both/all interactors. These may well be biological false-positives that arise from the forced out-of-context expression of baits and preys in Y2H or from overexpression of affinity-tagged proteins in AP–MS. Nevertheless, because an interaction does not appear to make sense does not necessarily mean that it is a false-positive. There is a growing list of moonlighting proteins performing multiple, apparently unrelated functions. Many moonlighting proteins are metabolic enzymes with additional functional activity, although numerous signaling, structural and nucleic acid binding proteins are also multifunctional (21,112).

The existence of moonlighting proteins argues that Y2H and AP–MS results that initially appear to be biologically

irrelevant should not be dismissed out-of-hand. Instead, novel interactions found by unbiased approaches such as Y2H and AP–MS may provide clues on new functional annotations for well-characterized proteins and potential multiple functions for previously uncharacterized proteins. An uncharacterized protein may connect different functions as either a moonlighting protein or through interactions with other uncharacterized proteins (105). Thus, simple reliance on conserved domains and majority rule assessments may lead to incomplete functional annotation of both characterized and uncharacterized proteins.

BUILD AS YOU GO: INTERACTOME WALKING

In ‘interactome walking’ interaction partners identified from an initial screen are subsequently used as baits in secondary and tertiary rounds of screening, a strategy readily accomplished by accessing cloned ORFeome collections. Interactome walking has been implemented for the DNA replication network of *Bacillus subtilis* (113), the TGF- β signaling pathway in *C. elegans* (114) and the network about the human huntingtin protein associated with Huntington Disease (115). Interactome walking has the added benefit that advance knowledge of the interactors or components of a particular target protein, complex or organelle is not necessary to conduct the screen. By letting the experiment direct the screen, rather than the experimenter, interactome walking can identify novel interactions and interactors that serve as critical links between discrete functions (113,114). Identified links can serve to enhance the annotation of a known protein or allow one to assign novel functions based on the connections observed (105).

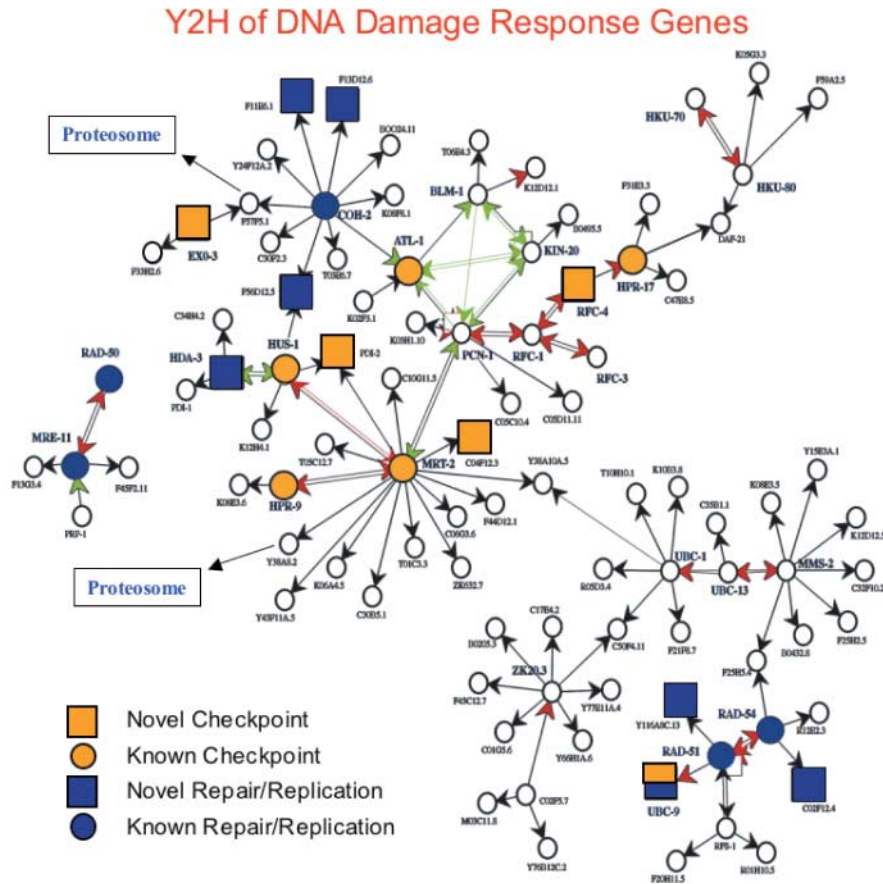


Figure 3. Framework of Y2H interactions for *C. elegans* DNA damage response genes: a high degree of interconnectivity and numerous novel interactions among DNA damage response genes were identified by Y2H. Phenotypic testing provided evidence for assigning new functions to novel interactors. Adapted with permission from Boulton *et al.* (129).

COMPUTATIONAL ASSESSMENT OF HIGH-THROUGHPUT DATA SETS

Efforts to assess biological relevance and/or reliability of HT-Y2H and HT-AP-MS interactome data sets have utilized phylogenetic conservation, subcellular localization patterns, co-expression correlations, and network topology (94–96,98,101,102,108,116–118). The basic conclusions from these efforts are that the interactome data sets are of reasonably good quality for both Y2H and AP-MS, that the interactions or complexes identified generally recapitulate biologically relevant processes, and that previously unsuspected complexes containing unknown proteins can be identified (101). Computational methods can increase confidence in unbiased Y2H and AP-MS data sets through correlation with other data. However, apparently uncorrelated interactions should not be discarded because they may still be biologically relevant.

As none of the compiled data sets of experimentally derived protein–protein interactions constitutes a truly comprehensive data set for any eukaryotic organism (71–74,83,105), and given the high false-negative rate intrinsic to Y2H and AP-MS, computational prediction of interactions has also been undertaken. Computational prediction efforts have utilized

conserved protein and DNA sequences, functional annotations based on GO, co-localization or homologous interactions in other species, and are complemented by mathematical modeling employing strategies such as Bayesian networks (75–78,80,119–122). The daunting challenges faced by purely computational approaches are to effectively incorporate the nearly 40% of human genes with minimal or no annotation, and to incorporate multifunctional proteins, particularly when only one functional activity is established. Recent analyses of multifunctional proteins demonstrate that computational predictions can accurately identify multiple functions but also have an inherent false-positive rate (123). Computational prediction utilizing subcellular localization data avoids potential false-positives but can lead to increased false-negatives when a protein is multifunctional and/or active in multiple subcellular compartments or complexes (48,49,108).

PROTEIN INTERACTION NETWORKS: THE INTERACTOME

The most comprehensive interaction maps currently available are actually compilations of high-throughput Y2H and AP-MS appended to an assemblage of literature-reported

interactions curated in various interaction databases (10,74,83). The overlap between different curated data sets is quite small, as is the overlap between literature-curated data sets and high-throughput approaches (74,83). In literature-curated collections the number of interactions involving novel or unknown genes is quite low, particularly for human proteins (83). As nearly 40% of human genes do not have a PfamA or GO annotation, the current compilations do not adequately capture the entire repertoire of possible functions. This shortfall defines 'inspection bias', where characterized proteins get repeatedly examined and uncharacterized proteins are left out. Although only a fraction of the expected number of interactions for any organism examined to date have been identified (74,124,125), the proteome-scale screens done so far have nevertheless connected hundreds of uncharacterized proteins to well-characterized partners (52,126–128), thereby circumventing inspection bias somewhat and providing preliminary evidence for functional annotation.

Regardless of the data source, be it AP-MS or Y2H, or genome-scale or smaller scale, many interconnections between distinct biological processes are noted (6–10,82,129). Many of these interconnections involve novel or otherwise unknown proteins, thereby suggesting functions for these proteins. For example, an Y2H study in worm (129) identified many novel DNA damage response genes (Fig. 3). Moreover, proteins of known function are found in interaction clusters that are of completely different function, suggesting potentially new functions for these 'known' proteins (48,49,130,131). Some of the interacting proteins were partially characterized previously and have putative functions in line with the function of the known cluster (113,132), but most are uncharacterized. In addition, every global-scale Y2H screen has identified novel interactions not previously observed by focused examination of individual protein clusters.

SYSTEMATIC MAPPING AND INTEGRATION OF DATA

Although current high-confidence interactome data sets are very incomplete, they still provide a framework onto which other types of biological information can be hung. Integrating other types of data (133), either data from alternative PPI approaches (101,134,135), or data from altogether different functional genomics approaches (17,120,129,136–138), or both (79,96,102), with interactome maps is the most effective way to assign function to uncharacterized proteins that are components of the network. Another value of the framework is that it provides a network for cellular modeling (139), which further enhances functional annotation.

Combining experimental and computational approaches for identifying protein–protein interactions have substantially expanded the known universe of human protein–protein interactions. The latest contribution is a recently completed Y2H screen testing over 7000 human full-length ORFs (83) in all possible pair-wise interactions, and applying the stringent criteria described earlier to reduce the number of false-positives. The screen identified 2754 interactions among 1549 proteins, and included more than 300 novel connections to over 100 human disease genes. Combined with a

set of ~4700 high-confidence literature-curated interactions, a human interactome map of nearly 7500 high-confidence interactions is now available, constituting a framework human interactome for further investigation into systems biology as well as functional annotation of known and unknown proteins.

Even though existing interactome maps are still far from complete (57), some of their topological properties are already evident. For example, current interactome networks appear to have a 'scale-free' or power law degree distribution, that is, most proteins interact with few partners, whereas a few proteins, the 'hubs', interact with many partners (1,2). Power law topology might relate to genetic robustness (140), as knockouts of genes encoding hubs are approximately three times more likely to confer lethality than those of non-hubs (1,2). Furthermore, the dynamics of interactions mediated by hub proteins points to a modular organization of the yeast proteome coordinated by rapidly evolving 'inter-module' or 'date' hubs (2,141). As interactome maps expand, particularly in an unbiased manner that captures all possible interactions that can take place, novel hypotheses of how network dynamics and topology relate to biological function will be generated. Finally, correlation of interactome data with other functional genomic data, such as expression and phenotypic profiling, will provide a clearer understanding of the functional relationships underlying biological processes (3,17).

ACKNOWLEDGEMENTS

We thank members of the Vidal Laboratory, our colleagues and collaborators, and the participants of the ORFeome Meeting for discussions; W. Zundel for critical reading of the manuscript; E. Benz, S. Korsmeyer, D. Livingston, P. McCue, J. Song, B. Rollins and the DFCI Strategic Planning Initiative for support. We regret not being able to include all relevant published work and apologize to those investigators whose studies were not included due to limitations in length and space. This work was supported by the High-Tech Fund of the DFCI (S. Korsmeyer), and by an Ellison Foundation grant, a Keck Foundation grant, and an 'interactome mapping' grant from NIH/NHGRI and NIH/NIGMS awarded to M.V. This article is dedicated to the memory of S. Korsmeyer.

Conflict of Interest statement. None declared.

REFERENCES

- Jeong, H., Mason, S.P., Barabasi, A.L. and Oltvai, Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
- Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., Cusick, M.E., Roth, F.P. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, **430**, 88–93.
- Ge, H., Walhout, A.J. and Vidal, M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.*, **19**, 551–560.
- Fromont-Racine, M., Rain, J.C. and Legrain, P. (1997) Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.*, **16**, 277–282.

5. Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N. and Vidal, M. (2000) Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*, **287**, 116–122.
6. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
7. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
8. Reboul, J., Vaglio, P., Rual, J.F., Lamesch, P., Martinez, M., Armstrong, C.M., Li, S., Jacotot, L., Bertin, N., Janky, R. *et al.* (2003) *C. elegans* ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.*, **34**, 35–41.
9. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.
10. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
11. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
12. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutillier, K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
13. Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, **23**, 324–328.
14. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
15. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
16. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S. and Weil, B. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.
17. Gunsalus, K.C., Ge, H., Schetter, A.J., Goldberg, D.S., Han, J.D., Hao, T., Bertin, N., Li, N., Huang, J., Chuang, L.S. *et al.* (2005) Predictive models of molecular machines involved in *C. elegans* early embryogenesis. *Nature*, **436**, 861–865.
18. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.
19. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
20. Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
21. Jeffery, C.J. (2003) Moonlighting proteins: old proteins learning new tricks. *Trends Genet.*, **19**, 415–417.
22. Lamesch, P., Milstein, S., Hao, T., Rosenberg, J., Li, N., Sequerra, R., Bosak, S., Doucette-Stamm, L., Vandenhaute, J., Hill, D. *et al.* (2004) *C. elegans* ORFeome version 3.1: increasing the coverage of ORFeome resources with improved gene predictions. *Genome Res.*, **14**, 2064–2069.
23. Wei, C., Lamesch, P., Arumugam, M., Rosenberg, J., Hu, P., Vidal, M. and Brent, M.R. (2005) Closing in on the *C. elegans* ORFeome by cloning TWINSCAN predictions. *Genome Res.*, **15**, 577–582.
24. Oshiro, G., Wodicka, L.M., Washburn, M.P., Yates, J.R., III, Lockhart, D.J. and Winzler, E.A. (2002) Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res.*, **12**, 1210–1220.
25. Kessler, M.M., Zeng, Q., Hogan, S., Cook, R., Morales, A.J. and Cottarel, G. (2003) Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. *Genome Res.*, **13**, 264–271.
26. Brasch, M.A., Hartley, J.L. and Vidal, M. (2004) ORFeome cloning and systems biology: standardized mass production of the parts from the parts-list. *Genome Res.*, **14**, 2001–2009.
27. Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C., Nguyen, M. *et al.* (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science*, **302**, 842–846.
28. Ota, T., Suzuki, Y., Nishikawa, T., Otsuki, T., Sugiyama, T., Irie, R., Wakamatsu, A., Hayashi, K., Sato, H., Nagai, K. *et al.* (2004) Complete sequencing and characterization of 21 243 full-length human cDNAs. *Nat. Genet.*, **36**, 40–45.
29. Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H. *et al.* (2003) Collection, mapping, and annotation of over 28 000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
30. Carninci, P., Waki, K., Shiraki, T., Konno, H., Shibata, K., Itoh, M., Aizawa, K., Arakawa, T., Ishii, Y., Sasaki, D. *et al.* (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.*, **13**, 1273–1289.
31. Strausberg, R.L. and Riggins, G.J. (2001) Navigating the human transcriptome. *Proc. Natl Acad. Sci. USA*, **98**, 11837–11838.
32. Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S., Wagner, L., Shenmen, C.M., Schuler, G.D., Altschul, S.F. *et al.* (2002) Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc. Natl Acad. Sci. USA*, **99**, 16899–16903.
33. Imanishi, T., Itoh, T., Suzuki, Y., O'Donovan, C., Fukuchi, S., Koyanagi, K.O., Barrero, R.A., Tamura, T., Yamaguchi-Kabata, Y., Tanino, M. *et al.* (2004) Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *PLoS Biol.*, **2**, E162.
34. Seki, M., Narusaka, M., Kamiya, A., Ishida, J., Satou, M., Sakurai, T., Nakajima, M., Enju, A., Akiyama, K., Oono, Y. *et al.* (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science*, **296**, 141–145.
35. The MGC Project Team. (2004) The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res.*, **14**, 2121–2127.
36. Hudson, J.R., Jr, Dawson, E.P., Rushing, K.L., Jackson, C.H., Lockshon, D., Conover, D., Lanciault, C., Harris, J.R., Simmons, S.J., Rothstein, R. *et al.* (1997) The complete set of predicted genes from *Saccharomyces cerevisiae* in a readily usable form. *Genome Res.*, **7**, 1169–1173.
37. Reboul, J., Vaglio, P., Tzellas, N., Thierry-Mieg, N., Moore, T., Jackson, C., Shin-i, T., Kohara, Y., Thierry-Mieg, D., Thierry-Mieg, J. *et al.* (2001) Open-reading-frame sequence tags (OSTs) support the existence of at least 17 300 genes in *C. elegans*. *Nat. Genet.*, **27**, 332–336.
38. Messina, D.N., Glasscock, J., Gish, W. and Lovett, M. (2004) An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome Res.*, **14**, 2041–2047.
39. Gong, W., Shen, Y.P., Ma, L.G., Pan, Y., Du, Y.L., Wang, D.H., Yang, J.Y., Hu, L.D., Liu, X.F., Dong, C.X. *et al.* (2004) Genome-wide ORFeome cloning and analysis of *Arabidopsis* transcription factor genes. *Plant Physiol.*, **135**, 773–782.
40. Collins, J.E., Wright, C.L., Edwards, C.A., Davis, M.P., Grinham, J.A., Cole, C.G., Goward, M.E., Aguado, B., Mallya, M., Mokrab, Y. *et al.* (2004) A genome annotation-driven approach to cloning the human ORFeome. *Genome Biol.*, **5**, R84.
41. Rual, J.F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.O. *et al.* (2004) Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.*, **14**, 2128–2135.
42. Bonaldo, M.F., Bair, T.B., Scheetz, T.E., Snir, E., Akabogu, I., Bair, J.L., Berger, B., Crouch, K., Davis, A., Eyestone, M.E. *et al.* (2004) 1274 full-open reading frames of transcripts expressed in the developing mouse nervous system. *Genome Res.*, **14**, 2053–2063.
43. Alberts, B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.

44. Hartwell, L.H., Hopfield, J.J., Leibler, S. and Murray, A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
45. Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proc. Natl Acad. Sci. USA*, **100**, 1128–1133.
46. Spirin, V. and Mirny, L.A. (2003) Protein complexes and functional modules in molecular networks. *Proc. Natl Acad. Sci. USA*, **100**, 12123–12128.
47. Oliver, S. (2000) Guilt-by-association goes global. *Nature*, **403**, 601–603.
48. Bomsztyk, K., Denisenko, O. and Ostrowski, J. (2004) hnRNP K: one protein multiple processes. *Bioessays*, **26**, 629–638.
49. Cohen, P.T. (2002) Protein phosphatase 1–targeted in many directions. *J. Cell Sci.*, **115**, 241–256.
50. Valente, A., Cusick, M.E., Fagerstrom, R.M., Hill, D.E. and Vidal, M. (2005) Coordinated functionality in the topology of the yeast interactome, in preparation.
51. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. *Yeast*, **18**, 523–531.
52. Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.
53. Vázquez, A., Flammini, A., Maritan, A. and Vespignani, A. (2003) Global protein function prediction from protein–protein interaction networks. *Nat. Biotechnol.*, **21**, 697–700.
54. Vidal, M. (2005) Interactome modeling. *FEBS Lett.*, **579**, 1834–1838.
55. Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
56. Papin, J.A., Hunter, T., Palsson, B.O. and Subramaniam, S. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell. Biol.*, **6**, 99–111.
57. Han, J.-D.J., Dupuy, D., Bertin, N., Cusick, M.E. and Vidal, M. (2005) Effect of sampling on topology predictions of protein–protein interaction networks. *Nat. Biotechnol.*, **23**, 839–844.
58. Middendorf, M., Ziv, E. and Wiggins, C.H. (2005) Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network. *Proc. Natl Acad. Sci. USA*, **102**, 3192–3197.
59. Stumpf, M.P., Wiuf, C. and May, R.M. (2005) Subnets of scale-free networks are not scale-free: sampling properties of networks. *Proc. Natl Acad. Sci. USA*, **102**, 4221–4224.
60. Thomas, A., Cannings, R., Monk, N.A. and Cannings, C. (2003) On the structure of protein–protein interaction networks. *Biochem. Soc. Trans.*, **31**, 1491–1496.
61. Walhout, A.J. and Vidal, M. (2001) Protein interaction maps for model organisms. *Nat. Rev. Mol. Cell. Biol.*, **2**, 55–62.
62. Vidal, M. and Legrain, P. (1999) Yeast forward and reverse 'n'-hybrid systems. *Nucleic Acids Res.*, **27**, 919–929.
63. Colland, F., Jacq, X., Trouplin, V., Mougou, C., Groizeleau, C., Hamburger, A., Meil, A., Wojcik, J., Legrain, P. and Gauthier, J.M. (2004) Functional proteomics mapping of a human signaling pathway. *Genome Res.*, **14**, 1324–1332.
64. Bauer, A. and Kuster, B. (2003) Affinity purification-mass spectrometry. Powerful tools for the characterization of protein complexes. *Eur. J. Biochem.*, **270**, 570–578.
65. Ferguson, P.L. and Smith, R.D. (2003) Proteome analysis by mass spectrometry. *Annu. Rev. Biophys. Biomol. Struct.*, **32**, 399–424.
66. Scholtens, D., Vidal, M. and Gentleman, R. (2005) Local modeling of global interactome networks. *Bioinformatics*, **21**, 3458–3551.
67. Marcotte, E.M. (2000) Computational genetics: finding protein function by nonhomology methods. *Curr. Opin. Struct. Biol.*, **10**, 359–365.
68. Xia, Y., Yu, H., Jansen, R., Seringhaus, M., Baxter, S., Greenbaum, D., Zhao, H. and Gerstein, M. (2004) Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.*, **73**, 1051–1087.
69. Galperin, M.Y. and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.
70. Droit, A., Poirier, G.G. and Hunter, J.M. (2005) Experimental and bioinformatic approaches for interrogating protein–protein interactions to determine protein function. *J. Mol. Endocrinol.*, **34**, 263–280.
71. Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
72. Pagel, P., Kovac, S., Oesterheld, M., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Mark, P., Stumpflen, V., Mewes, H.W. *et al.* (2005) The MIPS mammalian protein–protein interaction database. *Bioinformatics*, **21**, 832–834.
73. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjana, V., Muthusamy, B., Gandhi, T.K., Gronborg, M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
74. Ramani, A.K., Bunesco, R.C., Mooney, R.J. and Marcotte, E.M. (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, **6**, R40.
75. Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S. and Vidal, M. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein–protein interactions or 'interologs'. *Genome Res.*, **11**, 2120–2126.
76. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. *Genome Res.*, **14**, 1107–1118.
77. Brown, K.R. and Jurisica, I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.
78. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003) A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*, **302**, 449–453.
79. Zhang, L.V., Wong, S.L., King, O.D. and Roth, F.P. (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, **5**, 38.
80. Lehner, B. and Fraser, A. (2004) A first-draft human protein–interaction map. *Genome Biol.*, **5**, R63.
81. Fields, S. and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature*, **340**, 245–246.
82. Rain, J.C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V. *et al.* (2001) The protein–protein interaction map of *Helicobacter pylori*. *Nature*, **409**, 211–215.
83. 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. *et al.* (2005) Towards a proteome-scale map of the human interactome network. *Nature*, Epub ahead of print.
84. Uetz, P. and Pankratz, M.J. (2004) Protein interaction maps on the fly. *Nat. Biotechnol.*, **22**, 43–44.
85. Walhout, A.J. and Vidal, M. (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods*, **24**, 297–306.
86. Ito, T., Ota, K., Kubota, H., Yamaguchi, Y., Chiba, T., Sakuraba, K. and Yoshida, M. (2002) Roles for the two-hybrid system in exploration of the yeast protein interactome. *Mol. Cell. Proteom.*, **1**, 561–566.
87. Buckholz, R.G., Simmons, C.A., Stuart, J.M. and Weiner, M.P. (1999) Automation of yeast two-hybrid screening. *J. Mol. Microbiol. Biotechnol.*, **1**, 135–140.
88. Uetz, P. and Hughes, R.E. (2000) Systematic and large-scale two-hybrid screens. *Curr. Opin. Microbiol.*, **3**, 303–308.
89. Obrdlik, P., El-Bakkoury, M., Hamacher, T., Cappellaro, C., Vilarino, C., Fleischer, C., Ellerbrok, H., Kamuzinzi, R., Ledent, V., Blaudez, D. *et al.* (2004) K⁺ channel interactions detected by a genetic system optimized for systematic studies of membrane protein interactions. *Proc. Natl Acad. Sci. USA*, **101**, 12242–12247.
90. Vidal, M. (1997) The reverse two-hybrid system. In Bartels, P. and Fields, S. (eds), *The Yeast Two-Hybrid System*. Oxford University Press, New York, pp. 109–147.
91. Vidalain, P.O., Boxem, M., Ge, H., Li, S. and Vidal, M. (2004) Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods*, **32**, 363–370.
92. James, P., Halladay, J. and Craig, E.A. (1996) Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics*, **144**, 1425–1436.
93. Walhout, A.J. and Vidal, M. (1999) A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res.*, **9**, 1128–1134.
94. Patil, A. and Nakamura, H. (2005) Filtering high-throughput protein–protein interaction data using a combination of genomic features. *BMC Bioinformatics*, **6**, 100.

95. Poyatos, J.F. and Hurst, L.D. (2004) How biologically relevant are interaction-based modules in protein networks? *Genome Biol.*, **5**, R93.
96. Bader, J.S., Chaudhuri, A., Rothberg, J.M. and Chant, J. (2004) Gaining confidence in high-throughput protein interaction networks. *Nat. Biotechnol.*, **22**, 78–85.
97. Okada, K., Kanaya, S. and Asai, K. (2005) Accurate extraction of functional associations between proteins based on common interaction partners and common domains. *Bioinformatics*, **21**, 2043–2048.
98. Deane, C.M., Salwinski, L., Xenarios, I. and Eisenberg, D. (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteom.*, **1**, 349–356.
99. Zhong, J., Zhang, H., Stanyon, C.A., Tromp, G. and Finley, R.L., Jr. (2003) A strategy for constructing large protein interaction maps using the yeast two-hybrid system: regulated expression arrays and two-phase mating. *Genome Res.*, **13**, 2691–2699.
100. Lehner, B., Semple, J.I., Brown, S.E., Counsell, D., Campbell, R.D. and Sanderson, C.M. (2004) Analysis of a high-throughput yeast two-hybrid system and its use to predict the function of intracellular proteins encoded within the human MHC class III region. *Genomics*, **83**, 153–167.
101. Bader, G.D. and Hogue, C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.
102. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P. (2002) Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403.
103. Formstecher, E., Aresta, S., Collura, V., Hamburger, A., Meil, A., Trehin, A., Reverdy, C., Betin, V., Maire, S., Brun, C. *et al.* (2005) Protein interaction mapping: A *Drosophila* case study. *Genome Res.*, **15**, 376–384.
104. Stanyon, C.A., Liu, G., Mangiola, B.A., Patel, N., Giot, L., Kuang, B., Zhang, H., Zhong, J. and Finley, R.L., Jr. (2004) A *Drosophila* protein–interaction map centered on cell-cycle regulators. *Genome Biol.*, **5**, R96.
105. Jiang, T. and Keating, A.E. (2005) AVID: an integrative framework for discovering functional relationships among proteins. *BMC Bioinformatics*, **6**, 136.
106. Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.
107. Terradot, L., Durnell, N., Li, M., Ory, J., Labigne, A., Legrain, P., Colland, F. and Waksman, G. (2004) Biochemical characterization of protein complexes from the *Helicobacter pylori* protein interaction map: strategies for complex formation and evidence for novel interactions within Type IV secretion systems. *Mol. Cell. Proteom.*, **3**, 809–819.
108. Krause, R., von Mering, C., Bork, P. and Dandekar, T. (2004) Shared components of protein complexes—versatile building blocks or biochemical artefacts? *Bioessays*, **26**, 1333–1343.
109. Taylor, S.W., Fahy, E. and Ghosh, S.S. (2003) Global organellar proteomics. *Trends Biotechnol.*, **21**, 82–88.
110. Brajenovic, M., Joberty, G., Kuster, B., Bouwmeester, T. and Drewes, G. (2004) Comprehensive proteomic analysis of human Par protein complexes reveals an interconnected protein network. *J. Biol. Chem.*, **279**, 12804–12811.
111. Endoh, H., Vincent, S., Jacob, Y., Real, E., Walhout, A.J. and Vidal, M. (2002) Integrated version of reverse two-hybrid system for the post-proteomic era. *Meth. Enzymol.*, **350**, 525–545.
112. Sriram, G., Martinez, J.A., McCabe, E.R., Liao, J.C. and Dipple, K.M. (2005) Single-gene disorders: what role could moonlighting enzymes play? *Am. J. Hum. Genet.*, **76**, 911–924.
113. Noirot-Gros, M.F., Dervyn, E., Wu, L.J., Mervelet, P., Errington, J., Ehrlich, S.D. and Noirot, P. (2002) An expanded view of bacterial DNA replication. *Proc. Natl Acad. Sci. USA*, **99**, 8342–8347.
114. Tewari, M., Hu, P.J., Ahn, J.S., Ayivi-Guedehoussou, N., Vidalain, P.O., Li, S., Milstein, S., Armstrong, C.M., Boxem, M., Butler, M.D. *et al.* (2004) Systematic interactome mapping and genetic perturbation analysis of a *C. elegans* TGF-beta signaling network. *Mol. Cell*, **13**, 469–482.
115. Goezler, H., Lalowski, M., Stelzl, U., Waelter, S., Stroedicke, M., Worm, U., Droege, A., Lindenberg, K.S., Knoblich, M., Haenig, C. *et al.* (2004) A protein interaction network links GIT1, an enhancer of Huntingtin aggregation, to Huntington's Disease. *Mol. Cell*, **15**, 853–865.
116. Deng, M., Zhang, K., Mehta, S., Chen, T. and Sun, F. (2003) Prediction of protein function using protein–protein interaction data. *J. Comput. Biol.*, **10**, 947–960.
117. Sprinzak, E., Sattath, S. and Margalit, H. (2003) How reliable are experimental protein–protein interaction data? *J. Mol. Biol.*, **327**, 919–923.
118. Kemmeren, P. and Holstege, F.C. (2003) Integrating functional genomics data. *Biochem. Soc. Trans.*, **31**, 1484–1487.
119. Ben-Hur, A. and Noble, W.S. (2005) Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**(Suppl. 1), i38–i46.
120. Jansen, R., Greenbaum, D. and Gerstein, M. (2002) Relating whole-genome expression data with protein–protein interactions. *Genome Res.*, **12**, 37–46.
121. Huang, S. (2004) Back to the biology in systems biology: what can we learn from biomolecular networks? *Brief Funct. Genomic Proteom.*, **2**, 279–297.
122. Liu, Y., Liu, N. and Zhao, H. (2005) Inferring protein–protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, **21**, 3279–3285.
123. Gomez, A., Domedel, N., Cedano, J., Pinol, J. and Querol, E. (2003) Do current sequence analysis algorithms disclose multifunctional (moonlighting) proteins? *Bioinformatics*, **19**, 895–896.
124. Bork, P., Jensen, L.J., von Mering, C., Ramani, A.K., Lee, I. and Marcotte, E.M. (2004) Protein interaction networks from yeast to human. *Curr. Opin. Struct. Biol.*, **14**, 292–299.
125. Grigoriev, A. (2003) On the number of protein–protein interactions in the yeast proteome. *Nucleic Acids Res.*, **31**, 4157–4161.
126. Tucker, C.L., Gera, J.F. and Uetz, P. (2001) Towards an understanding of complex protein networks. *Trends Cell Biol.*, **11**, 102–106.
127. Fromont-Racine, M., Mayes, A.E., Brunet-Simon, A., Rain, J.C., Colley, A., Dix, I., Decourty, L., Joly, N., Ricard, F., Beggs, J.D. *et al.* (2000) Genome-wide protein interaction screens reveal functional networks involving Sm-like proteins. *Yeast*, **17**, 95–110.
128. McDermott, J. and Samudrala, R. (2004) Enhanced functional information from predicted protein networks. *Trends Biotechnol.*, **22**, 60–62; discussion 62–63.
129. Boulton, S.J., Gartner, A., Reboul, J., Vaglio, P., Dyson, N., Hill, D.E. and Vidal, M. (2002) Combined functional genomic maps of the *C. elegans* DNA damage response. *Science*, **295**, 127–131.
130. Lehner, B. and Sanderson, C.M. (2004) A protein interaction framework for human mRNA degradation. *Genome Res.*, **14**, 1315–1323.
131. Li, F., Long, T., Lu, Y., Ouyang, Q. and Tang, C. (2004) The yeast cell-cycle network is robustly designed. *Proc. Natl Acad. Sci. USA*, **101**, 4781–4786.
132. Noirot, P. and Noirot-Gros, M.F. (2004) Protein interaction networks in bacteria. *Curr. Opin. Microbiol.*, **7**, 505–512.
133. Vidal, M. (2001) A biological atlas of functional maps. *Cell*, **104**, 333–339.
134. Aloy, P. and Russell, R.B. (2002) Interrogating protein interaction networks through structural biology. *Proc. Natl Acad. Sci. USA*, **99**, 5896–5901.
135. Edwards, A.M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. and Gerstein, M. (2002) Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.*, **18**, 529–536.
136. Ge, H., Liu, Z., Church, G.M. and Vidal, M. (2001) Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nat. Genet.*, **29**, 482–486.
137. Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, F.C. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol. Cell*, **9**, 1133–1143.
138. Walhout, A.J., Reboul, J., Shtanko, O., Bertin, N., Vaglio, P., Ge, H., Lee, H., Doucette-Stamm, L., Gunsalus, K.C., Schetter, A.J. *et al.* (2002) Integrating interactome, phenome, and transcriptome mapping data for the *C. elegans* germline. *Curr. Biol.*, **12**, 1952–1958.
139. Ideker, T. and Lauffenburger, D. (2003) Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol.*, **21**, 255–262.
140. Wagner, A. (2000) Robustness against mutations in genetic networks of yeast. *Nat. Genet.*, **24**, 355–361.
141. Fraser, H.B. (2005) Modularity and evolutionary constraint on proteins. *Nat. Genet.*, **37**, 351–352.