doi:10.1093/nar/gkj038

FLIGHT: database and tools for the integration and cross-correlation of large-scale RNAi phenotypic datasets

David Sims[1,2], Borisas Bursteinas[2], Qiong Gao[2], Marketa Zvelebil[2] and Buzz Baum[1,*]

[1]Morphogenesis Group and [2]Bioinformatics and Systems Biology Group, Ludwig Institute for Cancer Research, UCL Branch, Courtauld Building, 91 Riding House Street, London, W1W 7BS, UK

Received August 15, 2005; Revised and Accepted September 29, 2005

ABSTRACT

FLIGHT (www.flight.licr.org) is a new database designed to help researchers browse and cross-correlate data from large-scale RNAi studies. To date, the majority of these functional genomic screens have been carried out using *Drosophila* cell lines. These RNAi screens follow 100 years of classical *Drosophila* genetics, but have already revealed their potential by ascribing an impressive number of functions to known and novel genes. This has in turn given rise to a pressing need for tools to simplify the analysis of the large amount of phenotypic information generated. FLIGHT aims to do this by providing users with a gene-centric view of screen results and by making it possible to cluster phenotypic data to identify genes with related functions. Additionally, FLIGHT provides microarray expression data for many of the *Drosophila* cell lines commonly used in RNAi screens. This, together with information about cell lines, protocols and dsRNA primer sequences, is intended to help researchers design their own cell-based screens. Finally, although the current focus of FLIGHT is *Drosophila*, the database has been designed to facilitate the comparison of functional data across species and to help researchers working with other systems navigate their way through the fly genome.

INTRODUCTION

The sequencing of the first metazoan genomes revealed the existence of numerous previously unidentified genes. Until the discovery of dsRNA-mediated interference (RNAi), the functional analysis of these genes seemed a daunting challenge. Now, using RNAi, the mRNA of any gene specified by sequence can be targeted for destruction, leading to depletion of the corresponding protein (1). Genomic sequence information can therefore be used to design RNAi libraries that induce a partial loss of function phenotype for every gene in the genome. Significantly, screens of this type can be carried out in a matter of weeks in a variety of organisms and systems—many of which were previously refractory to loss-of-function genetics. This approach was pioneered in *Caenorhabditis elegans*, where a number of genome-scale screens have been carried out (2–5). Since then, high-throughput RNAi screens have proved particularly successful in *Drosophila* cell culture, where protein knock-down can be induced by the addition of long dsRNAs to the culture medium. The ease and speed of RNAi screens in *Drosophila* cell culture and the relatively low level of redundancy in the *Drosophila* genome (6) have induced many researchers to turn to this simple system to identify genes controlling fundamental cell biological processes. As a result, there has been a year-upon-year increase in the number studies using RNAi in *Drosophila* cell culture (Figure 1)—a trend that appears likely to continue as dsRNA libraries and screening platforms become more widely available.

This flood of functional data from diverse assays brings with it new challenges, which FLIGHT has been specifically designed to meet. First, because RNAi can result in sequence-specific off-target effects, it is important to record the dsRNA sequences used in each RNAi experiment. Thus, where possible, FLIGHT links dsRNA sequences and RNAi experiments. Second, although classical genetic screens rarely reach saturation, RNAi screens can generate comprehensive and relatively unbiased phenotypic datasets, which contain useful functional information about every gene targeted. Consequently, FLIGHT endeavours to capture the full details of all genes tested in screens, so that it is easy to access and analyse data that passes without comment during publication. Third, the majority of results from RNAi screens usually remain unverified, so that data are likely to contain many false positives and false negatives. In order to make confident

[*]To whom correspondence should be addressed. Tel: +44 207 878 4044; Fax: +44 207 878 4040; Email: b.baum@ucl.ac.uk

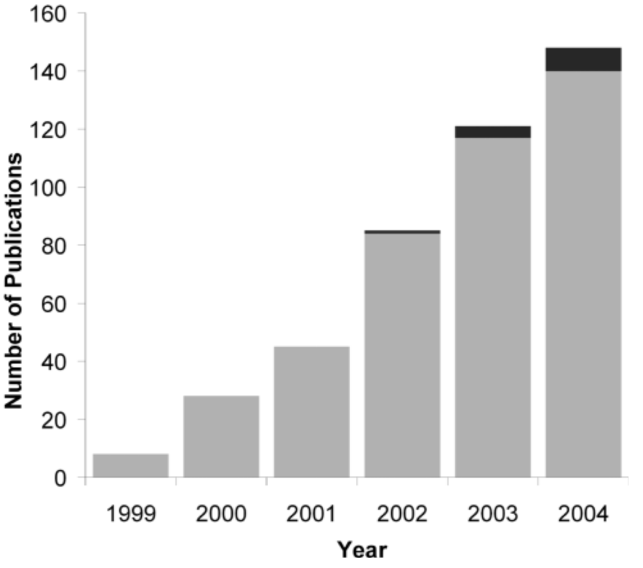


Figure 1. A chart to show the year-upon-year increase in *Drosophila* RNAi studies. The lower (light grey) area of the chart was generated by searching PubMed for published articles that include the keywords 'Drosophila' and 'RNAi' or 'RNA-interference' in the title or abstract. The upper (dark grey) area denotes the number of large-scale *Drosophila* RNAi screens currently included in FLIGHT.

inferences about gene function based on data from RNAi screens, it is therefore important to be able to combine data from multiple screens, and to integrate it with sequence, gene expression and protein interaction data. FLIGHT provides biologists with tools to facilitate this type of analysis.

To encourage more groups to design and carry out their own RNAi screens in *Drosophila* cell culture, FLIGHT also gives users access to normalized microarray expression data for many of the *Drosophila* cell lines commonly used in RNAi screens, along with information about the origin and maintenance of those cell lines. This, together with a repository of RNAi protocols and dsRNA primer sequences should facilitate the design of novel RNAi experiments. Finally, protein homology datasets have been included to facilitate the comparison of functional data across species.

THE DATABASE

User interface

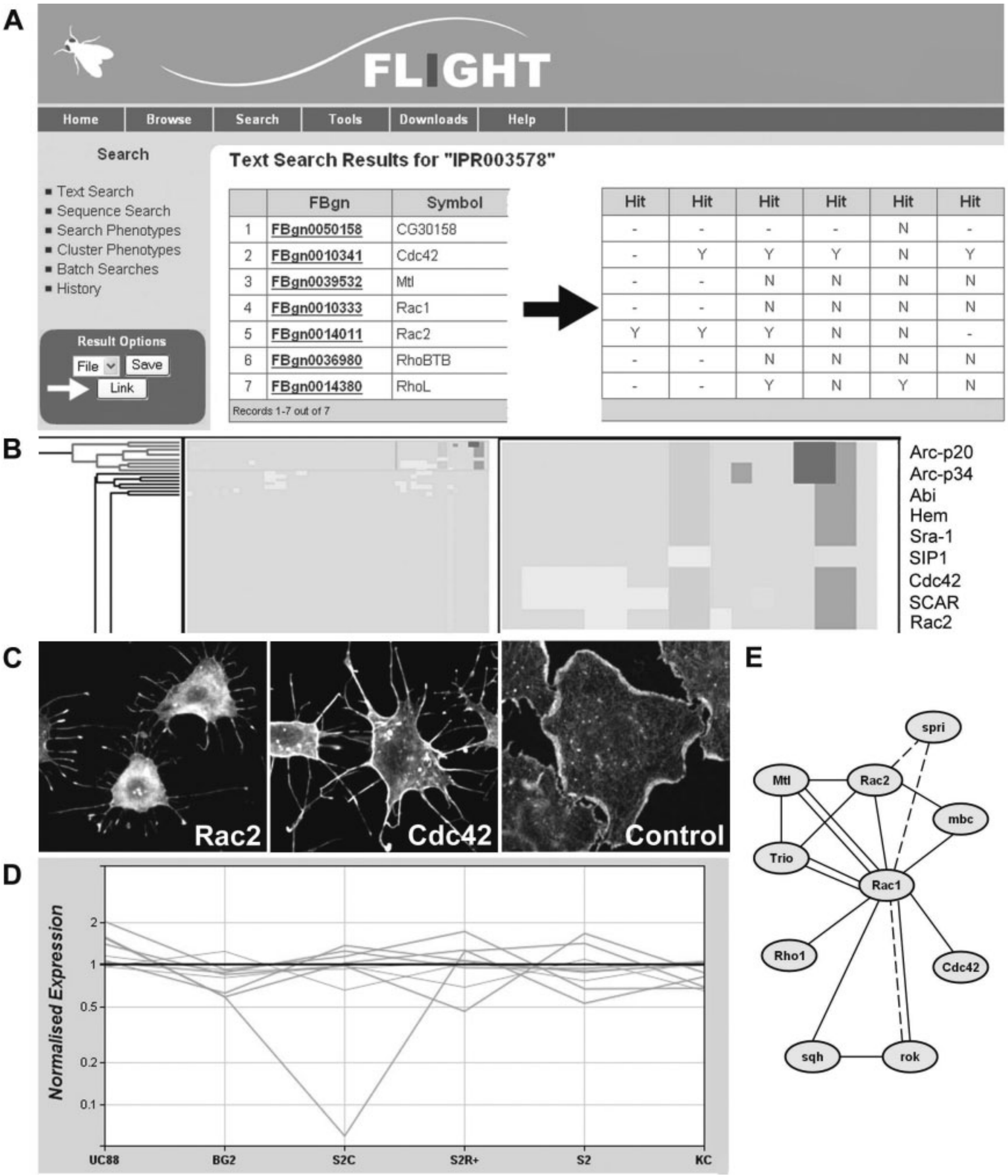
Within FLIGHT (www.flight.licr.org) users can browse through lists of RNAi hits, protocols and *Drosophila* cell lines. Alternatively, they can use text or sequence-based searches to navigate within the database. A user may wish to begin their search with a single gene of interest to identify the corresponding RNAi phenotypes or microarray expression pattern. FLIGHT contains gene identifiers from FlyBase (7), WormBase (8), SGD (9), MGD (10) and Genew (11), along with protein data from UniProt (12). This enables users to search for genes from man, mouse, yeast or worm as well as fly. If a search is initiated using a non-fly gene, lists of putative *Drosophila* orthologues will be returned from the three different homology datasets contained in the database: HomoloGene (13), InParanoid (14) and in-house reciprocal best hit BLAST data. Once a particular *Drosophila* gene has been selected, the user is directed to the FLIGHT entry for that gene. By default, this will provide users with a summary of available RNAi and microarray expression data. By navigating down from the RNAi summary data, users can then access information about the paper from which these data were derived, the RNAi assay used, primary phenotypic data, together with details of the annotated RNAi phenotype. For microarray data, users can similarly navigate down from the initial summary to view the expression of their gene of interest across conditions in selected microarray experiments. Normalized Affymetrix expression data are plotted graphically on a customisable chart, which allows users to alter the conditions that are displayed. Each data point is given a score (present, marginal or absent) as an indication of its reliability, which is represented by different coloured points on the on the gene expression profile. From the primary gene entry page, users can also quickly access Gene Ontology (15), InterPro (16), homology and protein interaction information. In addition, to help researchers quickly design new RNAi experiments in *Drosophila* cell culture, FLIGHT links each gene to a corresponding set of dsRNA primer sequences taken from the literature and from two commercially available RNAi libraries; the first designed by Cenix for Ambion, the second developed by UCSF and distributed by Open Biosystems.

Batch searches

FLIGHT has been designed to allow users to analyse the results of high-throughput experiments in the context of other large-scale datasets. Thus, searches can be carried out using lists of genes, and gene lists generated with the database can be saved and combined. In addition, gene lists from non-fly species can be converted into a *Drosophila* gene list using any of the homology datasets described above. Lists of *Drosophila* genes can then be subjected to a variety of batch searches using the 'Link' function. Linking enables users to cross-reference datasets. For example, a user interested in Rho GTPases could start by searching the database with a relevant InterPro motif (e.g. IPR003578) to call up a list of *Drosophila* Rho GTPases (Figure 2A). Using the 'Link' function the user can then view which of the GTPases have been identified as hits in published RNAi screens (Figure 2A). In addition, for each gene in the family users can view the pattern of gene expression across a variety of *Drosophila* cells lines (Figure 2D), examine the full set of known physical and genetic interactors (Figure 2E), and obtain a list of the corresponding set of previously identified mutations in flies.

Phenotype annotation and clustering

In recent years, phenotypic clustering has proved useful within individual RNAi screens in *C.elegans* (5,17). One of the primary goals of FLIGHT is to use a similar principle to facilitate the comparison of data from multiple RNAi screens. Consequently, a Java-based PHenotype Annotation Tool (PHAT) and a Java application for clustering and visualization of phenotypic data (Shuffle) have been developed to compliment FLIGHT. PHAT has been designed to annotate RNAi screen hits using one or more annotation terms from a controlled vocabulary. For each relevant term, phenotypes



Text Search Results for "IPR003578"

	FBgn	Symbol
1	FBgn0050158	CG30158
2	FBgn0010341	Cdc42
3	FBgn0039532	Mtl
4	FBgn0010333	Rac1
5	FBgn0014011	Rac2
6	FBgn0036980	RhoBTB
7	FBgn0014380	RhoL

Records 1-7 out of 7

Hit	Hit	Hit	Hit	Hit	Hit
-	-	-	-	N	-
-	Y	Y	Y	N	Y
-	-	N	N	N	N
-	-	N	N	N	N
Y	Y	Y	N	N	-
-	-	N	N	N	N
-	-	Y	N	Y	N

Figure 2. Examples of ways to use FLIGHT and Shuffle. (**A**) To identify the full set of RhoGTPases in *Drosophila*, a list was generated using a text search with the InterPro motif 'IPR003578'. By clicking the 'Link' button on the left hand menu (small white arrow) users can perform a range of further searches using this gene list. The table on the right hand side illustrates the output seen following a 'Link' search for hits in selected RNAi screens using this gene list. (**B**) Panel shows a screen shot from Shuffle in which a cluster of related RNAi phenotypes has been highlighted by selecting one node of the tree. (**C**) Within FLIGHT detailed phenotypic data can be retrieved by linking to individual gene records. Screen images of morphological phenotypes for two Rho GTPases are shown along with an appropriate control. (Reprinted from Current Biology, Vol. 13, Kunda *et al*., pp. 1867–1875, 2003, with permission from Elsevier.) (**D**) Using the 'Link' function, the normalized expression of each gene in a list can be graphically displayed. This enables the expression profiles of different genes to be compared across a panel of *Drosophila* cell lines. (**E**) Genetic interactions (bold lines) and protein interactions (dashed lines) from publicly available datasets can be graphically displayed for a set of genes from a list (nodes).

are scored, relative to wild-type, on a scale from −3 to +3. This creates a quantitative phenotypic profile for each gene across different annotation terms, which can be used for hierarchical clustering within Shuffle. When clustering genes using Shuffle, users can apply a range of standard distance measures and tree-construction algorithms across one or more RNAi screens. Shuffle then generates an interactive, colour-coded dendrogram depicting the clustering results (Figure 2B). As genes with similar RNAi phenotypes tend to be functionally related (18), clusters of genes generated using these phenotypic profiles are likely to represent functional groups (Figure 2B). Users can extend their analysis by searching FLIGHT using gene lists generated within Shuffle. The availability of this tool on the website also makes it possible for users to continuously refine their analysis of published data as more becomes available.

Shuffle has also been used to estimate the correlation between the phenotypic profiles of all possible gene pairs within the database. The resulting correlation matrix is employed in the website function 'Find Similar Phenotypes' to enable users to search quickly for genes that are functionally related to their specific genes of interest.

Designing novel RNAi screens

As reagents, such as libraries of dsRNAs, become more widely available, RNAi screening will be limited primarily by assay development. The power, simplicity and scalability of the technique mean that it is applicable to a wide range of cell biological problems. With this in mind, FLIGHT has been designed to facilitate the design of new RNAi experiments. The website provides information about the origin, morphology and gene expression of many key *Drosophila* cell lines, allowing researchers to assess the suitability of particular cell lines for their chosen assay. FLIGHT also provides a repository of pre-designed dsRNA primer sequences, as well as links to the E-RNAi dsRNA primer design tool and its associated dsRNA library (19). Finally, authors are also encouraged to submit the results of their RNAi screens to FLIGHT. This will both facilitate analysis and help communicate their results to the wider community.

Current contents

FLIGHT is continually updated but currently includes a total of 3357 RNAi hits from 22 screens in *Drosophila melanogaster*. It contains normalized Affymetrix microarray expression information for six of the principal fly cell lines used in RNAi screens (S2, S2c, S2R+, Kc, BG2 and UC88), and details of over 42 000 pre-designed *Drosophila* dsRNA primer sequences. Finally, FLIGHT includes 44 187 protein interactions from high-throughput studies in *Drosophila* (20–22) and yeast (23–26), along with genetic interactions from FlyBase (7).

Discussion and future prospects

FLIGHT is designed to meet the increasing need to manage, navigate and cross-correlate data from the rapidly increasing number of RNAi screens. In addition, FLIGHT is designed to facilitate the cross-referencing of results from large-scale RNAi screens with data from yeast two-hybrid and microarray expression studies. The website provides interactive tools for annotating and clustering phenotypic data, and contains data designed to help users design, carry out and analyse their own RNAi screens.

Although the focus of the database is *Drosophila*, several large-scale RNAi screens have also been carried out in *C.elegans* (2–5) and the first human siRNA screens are beginning to be published (27,28). In the coming months, these datasets will be incorporated into the database, along with data from knock-out experiments in yeast. This data will be mapped via protein homology to putative fly orthologues. Finally, additional microarray expression data will shortly be added to the website. In the future, more complete integration of phenotypic, expression and protein interaction data should provide novel insights into the topology of networks that control different aspects of cell behaviour across cell types and across species.

ACKNOWLEDGEMENTS

The authors are grateful to Cenix, Ambion and Nico Stuurman and Ron Vale from UCSF for providing dsRNA primer sequence information, and FlyCHIP, Jing Wang and the BBSRC for help generating the microarray data. The authors would like to thank Anne Ridley for comments on the manuscript, Michael Stumpf for assistance with statistical methods, Ewan Mollison for help with data entry, Gavin Craig for compiling cell line data and Boda design for the FLIGHT logo. The research was funded in part by a studentship from the BBSRC and through funding from MAIN, an EU funded project. Funding to pay the Open Access publication charges for this article was provided by The Royal Society.

Conflict of interest statement. None declared.

REFERENCES

1. Fire,A., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
2. Kamath,R.S., Fraser,A.G., Dong,Y., Poulin,G., Durbin,R., Gotta,M., Kanapin,A., Le Bot,N., Moreno,S., Sohrmann,M. *et al*. (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature*, **421**, 231–237.
3. Maeda,I., Kohara,Y., Yamamoto,M. and Sugimoto,A. (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Curr. Biol*., **11**, 171–176.
4. Piano,F., Schetter,A.J., Mangone,M., Stein,L. and Kemphues,K.J. (2000) RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Curr. Biol*., **10**, 1619–1622.
5. Sonnichsen,B., Koski,L.B., Walsh,A., Marschall,P., Neumann,B., Brehm,M., Alleaume,A.M., Artelt,J., Bettencourt,P., Cassin,E. *et al*. (2005) Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature*, **434**, 462–469.
6. Sutcliffe,J.E. and Brehm,A. (2004) Of flies and men; p53, a tumour suppressor. *FEBS Lett*., **567**, 86–91.
7. Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res*., **33**, D390–D395.
8. Chen,N., Harris,T.W., Antoshechkin,I., Bastiani,C., Bieri,T., Blasiar,D., Bradnam,K., Canaran,P., Chan,J., Chen,C.K. *et al*. (2005) WormBase: a comprehensive data resource for Caenorhabditis biology and genomics. *Nucleic Acids Res*., **33**, D383–D389.
9. Christie,K.R., Weng,S., Balakrishnan,R., Costanzo,M.C., Dolinski,K., Dwight,S.S., Engel,S.R., Feierbach,B., Fisk,D.G., Hirschman,J.E. *et al*. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res*., **32**, D311–D314.
10. Eppig,J.T., Bult,C.J., Kadin,J.A., Richardson,J.E., Blake,J.A., Anagnostopoulos,A., Baldarelli,R.M., Baya,M., Beal,J.S., Bello,S.M. *et al*. (2005) The Mouse Genome Database (MGD): from genes to mice—a community resource for mouse biology. *Nucleic Acids Res*., **33**, D471–D475.
11. Wain,H.M., Lush,M.J., Ducluzeau,F., Khodiyar,V.K. and Povey,S. (2004) Genew: the human gene nomenclature database, 2004 updates. *Nucleic Acids Res*., **32**, D255–D257.
12. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
13. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.

14. O'Brien,K.P., Remm,M. and Sonnhammer,E.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.*, **33**, D476–D480.
15. 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.
16. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201–D205.
17. Piano,F., Schetter,A.J., Morton,D.G., Gunsalus,K.C., Reinke,V., Kim,S.K. and Kemphues,K.J. (2002) Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C.elegans*. *Curr. Biol.*, **12**, 1959–1964.
18. Kiger,A.A., Baum,B., Jones,S., Jones,M.R., Coulson,A., Echeverri,C. and Perrimon,N. (2003) A functional genomic analysis of cell morphology using RNA interference. *J. Biol.*, **2**, 27.
19. Arziman,Z., Horn,T. and Boutros,M. (2005) E-RNAi: a web application to design optimized RNAi constructs. *Nucleic Acids Res.*, **33**, W582–W588.
20. 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.
21. 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.
22. 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.
23. 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.
24. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 180–183.
25. 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.
26. 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.
27. Kittler,R., Putz,G., Pelletier,L., Poser,I., Heninger,A.K., Drechsel,D., Fischer,S., Konstantinova,I., Habermann,B., Grabner,H. *et al.* (2004) An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. *Nature*, **432**, 1036–1040.
28. MacKeigan,J.P., Murphy,L.O. and Blenis,J. (2005) Sensitized RNAi screen of human kinases and phosphatases identifies new regulators of apoptosis and chemoresistance. *Nat. Cell Biol.*, **7**, 591–600.