FEBS Letters 579 (2005) 1802–1807 FEBS 29329

Minireview

Re-analysis of data and its integration

Lars Juhl Jensen, Lars M. Steinmetz*

European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany

Accepted 1 February 2005

Available online 11 February 2005

Edited by Robert Russell and Giulio Superti-Furga

Abstract To understand a biological process it is clear that a single approach will not be sufficient, just like a single measurement on a protein – such as its expression level – does not describe protein function. Using reference sets of proteins as benchmarks different approaches can be scaled and integrated. Here, we demonstrate the power of data re-analysis and integration by applying it in a case study to data from deletion phenotype screens and mRNA expression profiling.

Keywords: Benchmarking; Data integration; Network analysis; Gene expression; Deletion phenotype

1. Introduction

Biologists have begun to compare the output from one set of high-throughput experiments to another, such as the overlap between deletion phenotypes and protein–protein interactions [1], subcellular localization and expression level [2], and mRNA expression and protein abundance [3]. For such studies *Saccharomyces cerevisiae* has emerged as the de facto standard organism, with numerous genome-scale data sets having been published on gene expression [4–7], deletion phenotypes [8–11], protein–protein interactions [12–15], protein–DNA interactions [16,17], protein abundance [3], and subcellular localization of proteins [18,19].

One surprise finding of systematic comparisons of high-throughput data has been the low overlap between mRNA expression screens and deletion phenotype screens, both of which are being applied as screens to identify new candidate genes in a variety of organisms [20]. When deletion phenotype screens were compared to mRNA expression screens in yeast, the proposed genes identified by phenotype agreed surprisingly poorly with those suggested based on equivalent expression data; the overlap was only 17% for sporulation [9], 7% for growth on non-fermentable carbon sources [11], and even lower for growth in galactose, high pH, high salt and sorbitol [10]. Finally, the number of genes with a fitness defect that showed differential expression in response to DNA damaging agents was no larger than expected by chance [8].

*Corresponding author.
E-mail address: larsms@embl.de (L.M. Steinmetz).

Among high-throughput data, often poor agreement is observed between experiments of the same type. It therefore remains unclear whether the observed differences are biologically relevant or if they are simply a result of a high error rate on either the expression and/or the phenotype data. To address this, we here analyze expression and phenotype data for sporulation and respiration as a case study, and use this process to illustrate the importance of data re-analysis and integration in general for characterizing components of a system.

2. High-throughput experiments: reproducable yet different

High-throughput data sets have often been pointed out to suffer from high error rates, which make it difficult to draw firm conclusions from them. For example, one study has pointed out a poor agreement between yeast genes identified from different mitotic cell cycle expression time series [23]. However, in case of the mitotic cell cycle it was recently shown that the disagreement is largely due to the analysis of the data rather the data themselves: (1) Different methods were used for the original analysis of each data set, which obviously causes discrepancies. (2) Most methods developed for reanalyzing the data turned out to perform worse on benchmark sets than, the methods originally used. (3) Finally, most analyses proposed more genes as being periodically expressed than the data sets supported, which causes a large number of different genes unrelated to the cell cycle to be suggested for each data set. Reanalyzing all data using the best performing algorithms and applying more stringent cutoffs considerably improved the agreement between experiments [24].

The time courses published for the yeast meiosis/sporulation [5,21] have not been reanalyzed to nearly the same extend as the mitotic cell cycle data. For that reason we have picked this system as a case study to illustrate the benefit of data reanalysis and integration. When comparing the lists of differentially expressed genes obtained from their time courses on the *S. cerevisiae* strains SKI and W303, Primig et al. [21] found 915 of the ~1600 genes suggested in each experiment to be identified in both strains. They attributed this relatively poor agreement to strain specific differences. Chu et al. [5] independently generated an expression timeseries in *S. cerevisiae* SKI.

We reanalyzed the three time courses, by simply ranking the genes according to their root-mean-square of log-ratios. Fig. 1A shows a Venn diagram based on the top-300 ranking genes from each data set. Of the 300 genes suggested by each

0014-5793/$30.00
doi:10.1016/j.febslet.2005.02.006

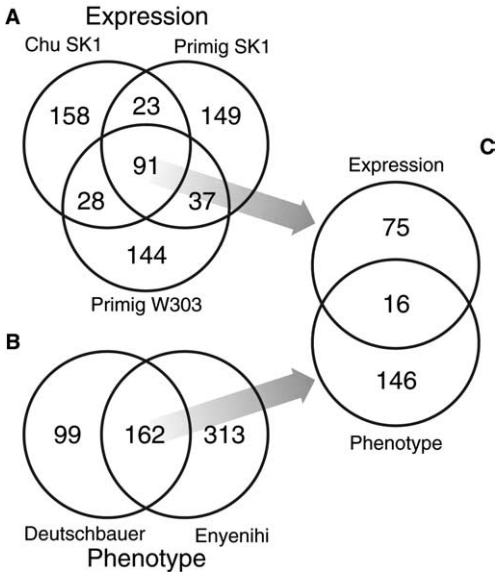

Fig. 1. Agreement of meiosis/sporulation-related gene sets identified from high-throughput experiments. Venn diagrams illustrating the agreement between expression time courses and mutant phenotype screens. (A) Comparison of the top-300 ranking genes in each of the sporulation expression time courses by Chu et al. [5] and Primig et al. [21]. Equally good agreement is observed among the three sets. (B) The systematic screens for sporulation deficient deletion strains by Deutschbauer et al. [9] and Enyenihi et al. [22] also show good agreement. (C) The core expression and phenotype gene sets obtained from the two other Venn diagrams show poor agreement. Different types of experiments thus identify different subsets of sporulation-related genes.

method, 91 genes are confirmed by both of the other time series and ~150 by at least one other time series.

To evaluate overlaps between data sets, we split the actual overlaps into two numbers: (1) the percentage of overlap that remains after subtracting what would be expected at random, and (2) the random expectation. Both of these percentages were calculated relative to the smaller set; the random expectation corresponds to the proportion of genes in the larger set out of all genes in the genome. The reason behind this approach is to correct for the increase in agreement that comes by chance, when assessing overlaps between large data sets.

For the ~150 genes identified from at least two time series, the overlap corresponds to 46% agreement (plus 4% expected at random) as compared to 32% agreement (plus 25% expected at random) obtained for a comparison of the two original data sets published by Primig et al. [21]. Compared to the original analyses, the agreement between the sporulation expression experiments is thus clearly improved by selecting a smaller, more conservative set of genes, as was also observed for the various expression time series on the mitotic cell cycle [24].

In addition to expression timeseries, phenotype screens were performed for sporulation. In this case strains were monitored that each lacked one gene product in the genome because of a gene deletion. This process identified genes that when deleted cause a defect in sporulation. Fig. 1B shows a comparison of two deletion strain screens for genes involved in sporulation [9,22]. As one of the two screens involved visual inspection of the strains rather than quantitative measurements [22], it was not possible to check if reanalysis of the data would improve the agreement. However, the agreement is already quite good with 52% (plus 10% expected at random) of the genes identified by Deutschbauer et al. [9] being confirmed by Enyenhi et al. [22]. We thus generally observed good agreement between different high-throughput experiments of the same type, be they microarray expression time series or phenotype screens of deletion mutants.

There can be little doubt that the vast majority of the 91 genes that occur in the top-300 list for all three expression time courses are in fact transcriptionally regulated during sporulation. Similarly, it is safe to assume that the 162 genes identified in both phenotype screens are important for *S. cerevisiae* to properly sporulate. Yet, the agreement between the two sets is remarkably poor as shown in Fig. 1C: only 16 genes are present in both of these high-confidence sets, which corresponds to 15% (plus 3% expected at random) of the genes identified from expression data being confirmed by deletion phenotype. Although this is much lower than observed between different experiments of the same type, it should be noted that 16 genes is higher than random expectation (hypergeometric test, $P < 10^{-8}$).

In order to assess the generality of these findings on meiosis/sporulation, we extended our analysis to an entirely different biological system, namely yeast mitochondria. From analysis of data sets of deletion phenotype and mRNA expression under fermentable and non-fermentable conditions, the same picture emerges. Fig. 2 shows that there is hardly any correlation between the genes that show a phenotype (specific growth defect on non-fermentable carbon source) and the genes that change in expression (growth on non-fermentable vs. fermentable carbon source).

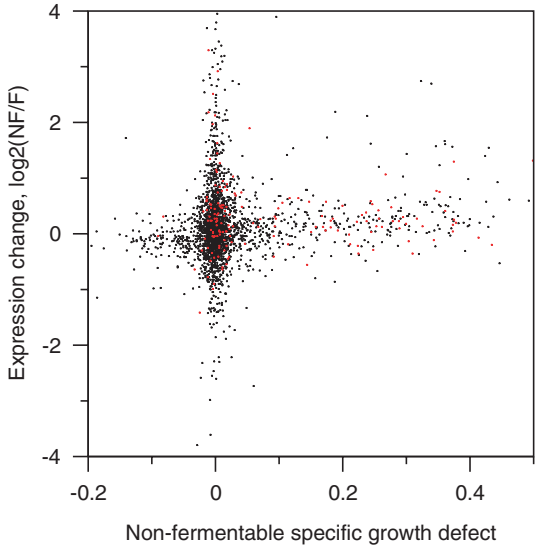

Fig. 2. Correlation of expression change and deletion phenotype under the same perturbation. Expression change was calculated as the $\log_2$(NF/F), where F and NF are the expression level under fermentable (F) and non-fermentable (NF) growth conditions, respectively (data from Prokisch et al. [25]). Deletion phenotype was measured as the difference in pooled growth rate between F and NF conditions – larger values correspond to a greater defect under NF conditions (data from Steinmetz et al. [11]). In red are genes whose protein products are known to localize to the mitochondrial organelle. The genes with the largest expression change tend not to have a deletion phenotype, and the genes with the largest deletion phenotype tend not to be expression regulated.

It would thus appear that data on phenotype and expression often disagree. As the agreement between multiple experiments of the same type is much better (Fig. 1), this cannot be explained by false predictions due to a high error rate on one (or both) data types. Instead, the obvious explanation is that the two assays disagree because they measure different properties of the biological system, which suggests that disagreement should also be expected for other data types than the ones considered here. Moreover, it implies that a complete biological systems in general cannot be identified using only a single high-throughput experimental technique.

3. Agreement with current biological notion

Given that the different high-throughput experimental methods are able to identify different parts of a biology system, it is natural to ask which method agrees better with the current conception of biology. To answer this, we compared the gene sets suggested by each method with a gold standard. For benchmarking, two lists of genes were compiled from the GeneOntology annotation in SGD [26]: one consisting of 191 genes with the terms "meiosis" or "sporulation" and another of 325 with the term "mitochondrion".

Benchmarking the high-throughput experimental data against known sporulation-related genes reveals that the two phenotype data sets agree marginally better with curated biological knowledge than the three expression data sets (Fig. 3A). Each individual experiment can only reliably detect 30–40% of the known sporulation genes. The figure also shows that each expression experiment only supports a reliable prediction for 300–500 genes, as the curves are parallel to the random expectation curve from this point on. Nonetheless, Chu et al. proposed a list of more than 1100 genes regulated in response to sporulation [5], and Primig et al. suggested ~1600 genes based on each of their two time courses [21].

In contrast, for the mitochondrial system, expression and phenotype experiments do not agree equally well with GeneOntology annotation. As already hinted at by Fig. 2, genes that encode mitochondrial localizing proteins are neither specifically expressed when cells are grown on a non-fermentable carbon source, nor are they among the genes that show the largest change in expression during the diauxic shift (Fig. 3B). Conversely, about 30% of the known mitochondrial genes result in detectable growth defects under non-fermentable conditions (Fig. 3B).

The analysis shows that phenotype data appears to generally agree well with curated biological knowledge, while expression data only agrees in some cases (Fig. 3). One part of the explanation is likely that phenotype data have long been used to assign gene function, which is not generally the case for expression data. More expression regulation may occur than in functional [20]. Moreover, because it is not necessary to regulate the expression of all subunits of a complex in order to control assembly and thus activity of a complex, there may be many genes involved in a process that need not change expression level and would not be detected by expression assays [27].

4. Analyzing proteins in network context

Large-scale screens for protein–protein and protein–DNA interactions provide an entirely different type of data, which creates a context for integrating predictions made by different high-throughput methods. So far, genome wide chromatin-IP screens for transcription factor binding sites [16,17] and most

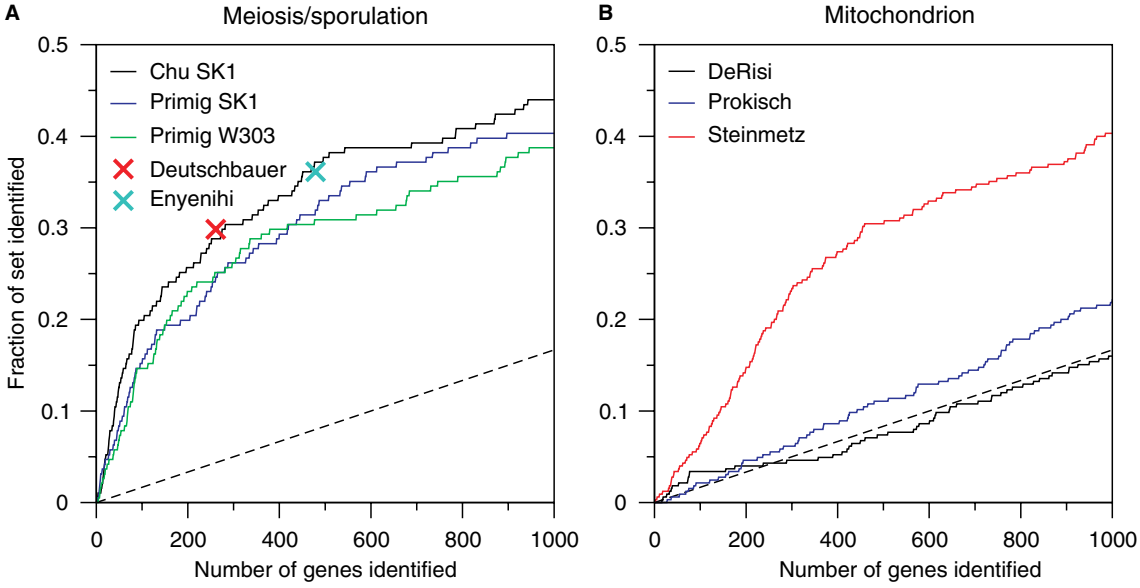

Fig. 3. Benchmark of high-throughput data. (A) The three expression time courses as well as the two deletion phenotype screens were benchmarked against known meiosis/sporulation genes from SGD [26]. The three expression data sets [5,21] are shown as curves since the genes could be ranked. The phenotype screens [9,22] are shown as single points since only lists with a fixed number of genes were available. The dotted line shows the expectation from random sampling. (B) Two expression time courses and one deletion phenotype screen relevant to mitochondria were benchmarked against known mitochondrion genes from SGD [26]. As expression data sets, we used the comparison of expression under fermentable and non-fermentable conditions by Prokisch et al. [25] as well as the time series by DeRisi et al. [4] that measures expression changes during the diauxic shift when yeast cells shift from fermentation to respiration. The Prokisch et al. data set and the phenotype data set [11] are the same as in Fig. 2.

large-scale protein–protein interaction screens [12–15] have been performed on yeast, although large yeast two-hybrid screens have also been published for both *Drosophila melanogaster* [28] and *Caenorhabditis elegans* [29]. Moreover, several methods have been developed for transferring interaction evidence between species based on homology/orthology [29–31].

The data obtained from interaction screens have, perhaps more than any other type of experiment, been criticized for being highly error prone. Indeed, several groups have estimated the rate of false positives to be in the order of 50% using several independent criteria for evaluation [32–38]. However, the reliability of individual interactions can be assessed using topology-based quality scores that rely on the local connectivity [39,27], thus allowing many of the erroneous interactions to be removed. The quality of an interaction set can be further improved by filtering interactions based on subcellular localization information [27], or by only considering interactions within a well defined system [27].

In addition to being important in their own right, interaction data are crucial for the interpretation of large-scale data, because they provide a network context for proteins identified by high throughput approaches. Simple examples include the two sporulation-related binary complexes shown in Fig. 4A, which both consist of a mixture of genes identified in expression and phenotype screens. In addition to linking proteins already identified by one or the other screen, interaction networks also allow the discovery of additional proteins that may have been missed by all assays (e.g., Hrr25p, Fig. 4B). Of the four proteins in the septin complex (Fig. 4C) that are part of neither the expression nor the phenotype core set, two are among the top-300 most regulated genes in at least one of the three experiments (Fig. 4C, lighter shades of blue). This illustrates how protein–protein interaction data can be used to integrate other types of experimental data, thereby allowing high-confidence predictions to be made for entire complexes based on weaker evidence from individual components. The approach is equally applicable to other types of functional modules, e.g., based on associations derived from genomic context methods or literature mining [31], and can be used for the integrating of many other types of data than expression and phenotype data.

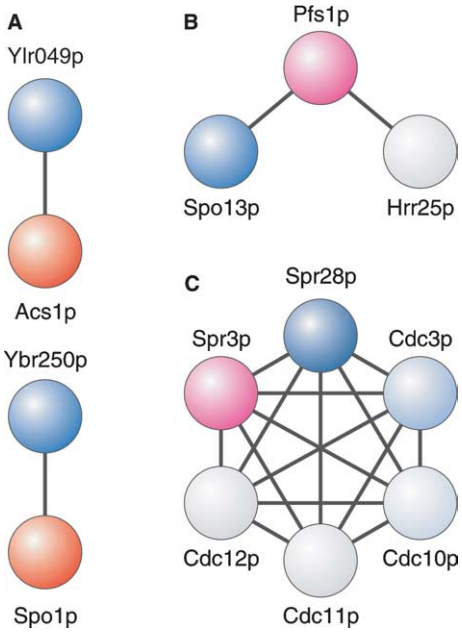

Fig. 4. Sporulation-related protein complexes. Proteins are colored according to sporulation expression change and deletion phenotype. Different shades of blue signify whether the gene was among the top-300 most regulated in one, two, or three (most intense) of the expression time courses. Genes exhibiting a sporulation-deficient deletion phenotype are shown in red, and genes detected by both the phenotype screen and the expression screen are magenta. The genes shown in white were identified as sporulation-related only through their protein–protein interactions. (A) Two examples of binary complexes, each consisting of a subunit of unknown function, identified only by expression time series, and a known sporulation-related gene identified only by deletion phenotype screens. (B) Module consisting of three proteins that play different roles during meiosis/sporulation. Hrr25p is only identified due to its interactions with Pfslp. (C) While the core expression and phenotype gene sets only detect two of the septin ring components as being involved in sporulation, the entire complex can be implicated in this process by integration of expression and phenotype data with protein–protein interactions.

5. Conclusion

Proteins do not function in isolation, rather their activity depends on a multitude of other factors in the cell, such as other proteins, small molecules, and ions. Analyzing proteins in the context of their physical and functional interaction is therefore an important step towards moving from a list of proteins to an understanding of cellular processes. To achieve this it is necessary to integrate complementary datasets and to evaluate the resultant data sets in the context of networks. Data integration will in many cases require re-analysis of the data using common benchmarks and integration schemes. Our case examples show that high-throughput data can be reproducible if analyzed using identical methods. Data sets coming from different approaches, like expression and deletion phenotype screens, may not agree because they measure different aspects of the biological system. For this reason data sets should be integrated to make full use of available complementary information.

Acknowledgment: We acknowledge support to L.M.S. from a grant from the Deutsche Forschungsgemeinschaft (STE1422/2-1) and thank Adam Deutschbauer for critical reading of the manuscript.

References

[1] Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W. (2002) Evolutionary rate in the protein interaction network. Proc. Natl. Acad. Sci. USA 196, 750–752.
[2] Drawid, A., Jansen, R. and Gerstein, M. (2000) Genome-wide analysis relating expression level with protein subcellular localization. Trends Genet. 16, 426–430.
[3] Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. Nature 425, 737–741.
[4] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680–686.
[5] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, L. (1998) The transcriptional program of sporulation in budding yeast. Science 282, 699–705.
[6] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E.,

Landsman, R.W., Lockhart, D.J. and Davis, R.W. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2, 65–73.
[7] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *S. cerevisiae* by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.
[8] Birrell, G.W., Brown, J.A., Wu, H.I., Giaever, G., Chu, A.M., Davis, R.W. and Brown, J.M. (2002) Transcriptional response of *Saccharomyces cerevisiae* to DNA-damaging agents does not identify the genes that protect against these agents. Proc. Natl. Acad. Sci. USA 99, 8778–8783.
[9] Deutschbauer, A.M., Williams, R.M., Chu, A.M. and Davis, R.W. (2002) Parallel phenotypic analysis of sporulation and postgermination growth in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. USA 99, 15530–15535.
[10] Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A.P., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.D., Flaherty, P., Foury, F., Garfinkel, D., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J.H., Hempel, S., Herman, Z., Jaramillo, D.F., Kelly, D.E., Kelly, S.L., Kotter, P., LaBonte, D., Lamb, D.C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S.L., Revuelta, J.L., Roberts, C.J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D.D., Sookhai-Mahadeo, S., Storms, R.K., Strathern, J.N., Valle, G., Voet, M., Volckaert, G., Wang, C.Y., Ward, T.R., Wilhelmy, J., Winzeler, E.A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J.D., Snyder, M., Philippsen, P., Davis, R.W. and Johnston, M. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418, 387–391.
[11] Steinmetz, L.M., Scharfe, C., Deutschbauer, A.M., Mokranjac, D., Herman, Z.S., Jones, T., M., C.A., Giaever, G., Prokisch, H., Oefner, P.J. and Davis, R.W. (2002) Systematic screen for human disease genes in yeast. Nat Genet 36, 400–404.
[12] Ito, T., Tashiro, K., Muta, S., Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. (2000) A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc. Natl. Acad. Sci. USA 97, 1143–1147.
[13] Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleish, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. Nature 403, 623–627.
[14] Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147.
[15] Ho, Y., Gruhler, A., Hellbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennet, K., Boutiller, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sørensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hougue, C.W.V., Figeys, D. and Tyers, M. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415, 180–183.
[16] Simon, L., Barnett, J., Hannett, N., Harbison, C.T., Ranaldi, N.J., Volkert, T.L., Wyrick, J.J., Zeitlinger, J., Gifford, D.K., Jaakkola, T.S. and Young, R.A. (2001) Serial regulation of transcriptional regulators in the yeast cell cycle. Cell 106, 697–708.
[17] Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, E., Zeitlinger, J., Jennings, E.G., Murray, H., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J.B., Volkert, T.L., Fraenkel, E., Gifford, D.K. and Young, R.A. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 298, 799–804.
[18] Kumar, A., Agarwal, S., Heyman, J.A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K.H., Miller, P., Gerstein, M., Roeder, G.S. and Snyder, M. (2002) Sub-cellular localization of the yeast proteome. Genes Dev. 16, 707–719.
[19] Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S. and O'Shea, E.K. (2003) Global analysis of protein localization in budding yeast. Nature 425, 686–691.
[20] Steinmetz, L.M. and Davis, R.W. (2004) Maximizing the potential of functional genomics. Nat. Rev. Genet. 5, 190–201.
[21] Primig, M., Williams, R.M., Winzeler, E.A., Tevzadze, G.G., Conway, A.R., Hwang, S.Y., Davis, R.W. and Esposito, R.E. (2000) The transcriptional program of sporulation in budding yeast. Nat Genet 26, 415–423.
[22] Enyenihi, A.H. and Saunders, W.S. (2003) Large-scale functional genomic analysis of sporulation and meiosis in *Saccharomyces cerevisiae*. Genetics 163, 47–54.
[23] Shedden, K. and Cooper, S. (2002) Analysis of cell-cycle gene expression in *Saccharomyces cerevisiae* using microarrays and multiple synchronization methods. Nucleic Acids Res. 30, 2920–2929.
[24] de Lichtenberg, U., Jensen, L.J., Fausbøll, A., Jensen, T.S., Bork, P. and Brunak, S. (2005) Comparison of computational methods for the identification of cell cycle regulated genes. Bioinformatics, Doi: 10.1093.
[25] Prokisch, H., Scharfe, C., Camp, D.G. 2nd, White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J. and Rothberg, J.M. (2004) A protein interaction map of *Drosophila melanogaster*. Science 302, 1727–1736.
[26] 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., Hong, E.L., Issel-Tarver, L., Nash, R., Sethuraman, A., Starr, B., Theesfeld, C.L., Andrada, R., Binkley, G., Dong, Q., Lane, C., Schroeder, M., Botstein, D. and Cherry, J.M. (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.
[27] de Lichtenberg, U., Jensen, L.J., Brunak, S. and Bork, P. (2005) Dynamic complex formation during the yeast cell cycle. Science 307, 724–727.
[28] Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., loime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley Jr., R.L., White, K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J. and Rothberg, J.M. (2003) A protein interaction map of *Drosophila melanogaster*. Science 302, 1727–1736.
[29] Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, P.P., Hill, D.E. and Vidal, M. (2004) A map

of the interactome network of the metazoan *C. elegans*. Science 303, 540–543.
[30] 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 regu-logs. Genome Res. 14, 1107–1118.
[31] von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. and Bork, P. (2005) STRING: Known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33, D433–D437.
[32] Aloy, P. and Russell, R.B. (2002) The third dimension for protein interactions and complexes. Trends Biochem. Sci. 27, 633–638.
[33] Bader, G.D. and Hogue, C.W.V. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. Nat Biotechnol 20, 991–997.
[34] 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. Proteomics I, 349–356.
[35] 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.
[36] Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A. and Holstege, C.P. (2002) Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol. Cell 9, 1133–1143.
[37] 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.
[38] Sprinzak, E., Sattath, S. and Margalit, H. (2003) How reliable are experimental protein–protein interaction data. J. Mol. Biol. 327, 919–923.
[39] Saito, R., Suzuki, H. and Hayashizaki, Y. (2003) Construction of reliable protein–protein interaction networks with a new interaction generality measure. Bioinformatics 19, 756–763.