Computational identification of cellular networks and pathways

Florian Markowetz[ab] and Olga G. Troyanskaya*[ab]

DOI: 10.1039/b617014p

In this article we highlight recent developments in computational functional genomics to identify networks of functionally related genes and proteins based on diverse sources of genomic data. Our specific focus is on statistical methods to identify genetic networks. We discuss integrated analysis of microarray datasets, methods to combine heterogeneous data sources, the analysis of high-dimensional phenotyping screens and describe efforts to establish a reliable and unbiased gold standard for method comparison and evaluation.

Computational functional genomics

In recent years, increasing quantities of high-throughput biological data have become available. Yet even in well studied model organisms like yeast, the functions of significant numbers of genes remain unknown, as do the interactions between gene products and their contributions to the highly structured networks of information flow that can be found in the cell. The inference of such cellular networks using computational and statistical methods is a prospering area of research in computational biology. Computational analysis of high-throughput data that assess functional relationships between gene products may be key to inferring cellular networks and pathways on a large scale. Such predictions can advance experimental studies by providing specific hypotheses for targeted experimental testing.

In this article, we highlight general problems and research strategies in this area by discussing examples in two central areas of computational functional genomics: first, the integration of different datasets and diverse sources of genomic data, and second, approaches to infer the inner organization of a cell by probing its reaction to external stimuli and perturbations. We mostly discuss methods developed in the context of the yeast *Saccharomyces cerevisiae*, because this model organism is widely used as a platform for the development of both high-throughput experimental techniques and computational methods. Readers further interested may also find other reviews[1–5] helpful that cover topics outside the special focus of this article.

Gene networks from microarray data

A prominent data source for assigning gene function is gene expression measurements by microarrays[6,7] that provide a global view of gene activity in a cell. Biological processes result from the concerted action of interacting molecules. This general observation suggests a simple idea, which has already motivated the first approaches to clustering expression profiles[8,9] and is still widely used in functional genomics. It is called the guilt-by-association heuristic: if two genes show similar expression profiles,

[a]*Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA. E-mail: ogt@cs.princeton.edu*
[b]*Department of Computer Science, Princeton University, Princeton, NJ 08544, USA. E-mail: ogt@cs.princeton.edu*

Florian Markowetz

Florian Markowetz studied Mathematics and Philosophy at the University of Heidelberg, Germany, and worked at the German Cancer Research Center (DKFZ) in Heidelberg. He obtained a PhD from the Free University Berlin, Germany, working at the Max Planck Institute for Molecular Genetics, Berlin. His research focused on inferring cellular signaling pathways from RNA interference screens. In July 2006 he joined Olga Troyanskaya's group in Princeton as a postdoctoral fellow and is now working on methods for data integration for pathway prediction.

Olga Troyanskaya

Olga Troyanskaya is an assistant professor in the Department of Computer Science and the Lewis-Sigler Institute for Integrative Genomics at Princeton University, where she leads a bioinformatics and functional genomics group. Her research interests include biological data integration, visualization of genomic data, microarray analysis, and prediction of protein function and biological pathways. Troyanskaya received a PhD in biomedical informatics from Stanford University.

they may follow the same regulatory regime.

Coexpression or relevance networks are a simple statistical model derived from the guilt-by-association heuristic.[10,11] They are constructed by computing a similarity score for each pair of genes, *e.g.* the correlation or mutual information between expression profiles. If similarity is above a certain threshold, the pair of genes gets connected in the graph, if not, it remains unconnected. Coexpression networks derived in this way agree well with functional similarity,[11] and many coexpression relationships are conserved over evolution.[10] This makes the assessment of coexpression relationships a key building block for most approaches to infer cellular networks and pathways.

From coexpression to mechanistic explanations

A central problem of coexpression analysis is that even high similarity of expression tells us little about the underlying biological mechanisms. For example, from the similarity of expression profiles alone, we cannot distinguish between direct and indirect relationships. To approach this problem, more advanced statistical models have been suggested, which build on the concept of conditional independence.[2,12] The idea is simple: if the relationship between gene A and gene B is not direct, but mediated through a gene (or possibly a group of genes) C, then A and B are independent given C. In other words: C explains the correlation between A and B.

A first example for conditional independence models is the Gaussian graphical model,[13,14] in which each gene pair is tested for conditional independence given the set of *all other genes* in the data. A different approach is using low-order independence models that search for a *single third gene* to explain the correlation of A and B. One example for this kind of modelling is the algorihm ARACNe, which was used for reverse engineering of regulatory networks in human B cells.[15]

A third type of conditional independence models are Bayesian networks.[16] They are perhaps the most flexible of all conditional independence models and offer the finest resolution of correlation structure. One example for Bayesian networks are Module networks[17] that encode modules (sets) of jointly regulated genes, their common regulators, and the conditions under which regulation occurs. Thus, the method gives a global view of the yeast transcriptional network and specifies regulatory programs of condition-specific regulators and their targets.

Whether sophisticated statistical models can outperform simple relevance networks in terms of prediction accuracy in real-world problems, where the observations are very noisy and only a small number of experimental repetitions are available, is still an open problem. A recent study[18] compares the accuracy of relevance networks, Gaussian graphical models and Bayesian networks to reconstruct the Raf pathway, a signalling network in human immune system cells, for both simulated and laboratory data. The higher computational costs of estimating statistical models more sophisticated than relevance networks were found to be only justified when the data was obtained from gene perturbation experiments. For data obtained by passive observations Gaussian graphical models and Bayesian networks offer no significant improvement over relevance networks. Future studies will show whether these observations are general or whether they might change with more and better data.

Integrated analysis of microarray data sets

One way to augment coexpression based methods is by including many different micorarray data sets in the analysis. Integrated analysis of different data sets can enable broader understanding of gene regulation in the context of specific pathways and can allow the discovery of coexpression relationships too weak to be detected in individual experiments. Such integrated analysis of microarray datasets is challenging because of differences in technology, protocols, and experimental conditions across datasets. Thus, any microarray integration system must be robust to such differences, and should easily adjust to new datasets, perhaps from technologies yet to be developed. Furthermore, in examining microarray results drawn from differing experimental conditions, it is critical to consider functional specificity, *i.e.* which biological processes are active in which experiments. The problem of integrating many high-throughput data sources thus includes a problem of determining functional relevance. The analysis objective is two-fold: (1) to reveal genes that are functionally related, and additionally (2) to identify the biological circumstances under which they relate.

To this end, Huttenhower *et al.*[19] propose a Bayesian framework facilitating the integration of multiple microarray data sets for predicting coexpression based functional networks of proteins. Each predicted functional relationship is provided within the context of a specific biological process. These biological functions of interest can be provided directly by a biologist, or they can be derived automatically from functional catalogs such as the Gene Ontology[20] or MIPS.[21] In addition to predicted functional relationships, the analysis process also provides a functional association score indicating how predictive each input microarray data set is of each biological function.

The examples of coexpression based methods we discussed above show the central role gene expression data plays in inferring cellular networks. Gene coexpression data are an excellent tool for hypothesis generation, yet microarray data alone often lack the degree of specificity needed for accurate gene network prediction. For such purposes, an increase in accuracy is needed that can be achieved through incorporation of heterogeneous functional data in an integrated analysis.

Integrated analysis of complementary data sources

Several challenges must be addressed in integrating diverse data. First, high-throughput data are typically noisy, and the nature and degree of this noise varies widely among experimental techniques and even among individual datasets produced by the same experimental method. Furthermore, results of high-throughput experiments vary substantially in the genes they cover, meaning that there are often a number of missing attributes. A second major challenge is the heterogeneity of the different

evidence types. In addition to variations in reliability, different sources of genomic data come in different representations: for instance, interaction data comes in the form of binary pairwise relationships while gene expression data are high-dimensional continuous measurements, and sequence data are strings of varying length. This heterogeneity makes it difficult to find a unifying representation that allows for data integration.

Data integration by kernel matrices and Bayesian networks

Myers *et al.*[22] use a Bayesian network for data integration, which readily captures the variation in reliability of different input types and accommodates missing information. The final output of the Bayesian integration is a completely connected, probabilistic graph of proteins. Each edge connecting two proteins is weighted by the posterior probability that these two proteins are functionally related given all the different data sources.

A different approach[23–25] to data integration uses (non)linear similarity measures called kernel functions, which were shown to be successfully applicable in a wide range of data analysis problems.[26] For kernel-based data integration each data type is summarized in a quadratic kernel matrix that has as entries the pairwise similarity values between proteins. Data integration is then performed by computing a weighted combination of these individual kernel matrices, where the weights indicate the relative importance of each data type. If two proteins show a combined kernel value above a certain threshold they get connected in the inferred network.

Models must be specific for the biological target context

A robust framework for integration of diverse data types is only the first step toward predicting biological networks. An equally challenging task is: given a map of protein or gene interactions from an integration of multiple data types, how does one group functionally related proteins together into process-specific networks? Expert information, such as proteins already known to be involved in the process of interest, should be used to direct the search and prediction process. Thus, Myers *et al.*[22] adopt a query-based model that allows a user to specify the biological area of interest, which can then be used to extract the relevant biological predictions. This is a first step to make pathway prediction methods applicable in many biological studies: the computational methods must be user-driven, or developed in such a way that a biologist can quickly extract the relevant information for his or her area of interest; and secondly, they must be biologically context-sensitive, meaning any prediction or search should be specifically optimized for the target biological context.

Genetic interactions from gene perturbation screens

Functional genomics has a long tradition of inferring the inner working of a cell through analysis of its response to various perturbations. Observing cellular features after knocking out or silencing a gene can reveal which genes are essential for an organism or for a particular pathway.

There are several perturbation techniques suitable for large-scale analysis in different organisms, including RNA interference[27] and gene knock-outs.[28] In most studies, perturbation effects are measured by single reporters like viability or growth.[29,30] Genetic interactions are then derived from perturbation screens by comparing the phenotypes of two single gene perturbations with the phenotype of the double gene perturbation. One example for a genetic interaction is epistasis, where one gene is masking the effect of another gene.[31] Another genetic interaction is synthetic lethality, where two genes with a viable phenotype show a lethal phenotype in a double perturbation.[32,33] Synthetic lethal interactions can be interpreted as two genes contributing to two alternative pathways: the cell can survive if one of these pathways is blocked, but not if both are affected.[34] Epistasis and synthetic lethality are just two examples of a broad range of possible genetic interactions. Drees *et al.*[35] define nine modes of genetic interactions for a quantitative phenotype that can be described by inequality constraints between the phenotypic values. They show that all modes of genetic interactions can be identified in agar-invasion phenotypes of mutant yeast.

Recent studies use phenotypes defined by high-dimensional readouts like gene expression profiles,[28,36] metabolite concentrations,[37] sensitivity to cytotoxic or cytostatic agents,[38] or morphological features of the cell.[39] Van Driessche *et al.*[40] use expression time-courses as phenotypes and partly reconstruct a developmental pathway in *Dictyostelium discoideum* by epistasis analysis. Such high-dimensional phenotypic profiles promise a comprehensive view on the function of genes in a cell, but only limited work has been done so far to adapt statistical and computational methodology to the specific needs of large-scale and high-dimensional phenotyping screens.

Phenotypic profiles offer only indirect information on gene interactions

A key obstacle to inferring genetic networks from high-dimensional perturbation screens is that phenotypic profiles generally offer only indirect information on how genes interact. Cell morphology or sensitivity to stresses are global features of the cell, which are hard to relate directly to the genes contributing to them. Gene expression phenotypes also offer an indirect view of pathway structure due to the high number of non-transcriptional regulatory events like protein modifications. For example, when silencing a kinase we might not be able to observe changes in the activation states of other proteins involved in the pathway. The only information we may get is that genes downstream of the pathway show expression changes. Thus, phenotypic profiles may provide only indirect information about information flow and pathway structure.

A recent approach especially designed to learning from indirect information and high-dimensional phenotypes are Nested Effects Models[41] that reconstruct features of the internal organization of the cell from the nested structure of observed perturbation effects. Perturbing some genes may have an influence on a global process, while perturbing others affects sub-processes of it. Imagine, for

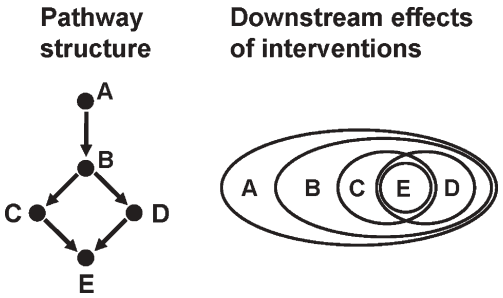

Fig. 1 Nested effects models. Phenotypic profiles generally offer only indirect information on how genes interact. Still, patterns of perturbation effects observable downstream of a target pathway can reveal features of the pathway topology[41]. Cutting the signal flow at an early point (A) shows a bigger set of downstream effects than cutting it at a later point (*e.g.* E). Branches in the pathway correspond to disjoint subsets of effects (B and C,D), while joint regulation can be seen by the overlap of perturbation effects (C,D and E). In real data, these subset relations are obscured by noise and must be recovered using a probabilistic model.

example, a signaling pathway activating several transcription factors. Blocking the entire pathway will affect all targets of the transcription factors, while perturbing a single downstream transcription factor will only affect its direct targets, which are a subset of the phenotype obtained by blocking the complete pathway. Fig. 1 shows a schematic plot of how the position of perturbed genes in a pathway corresponds to a nested structure of observed effects. The application of Nested Effects Models is exploratory and could provide a good starting point for more detailed analysis of gene function.

Method evaluation and gold standards

Individual high-throughput datasets are typically noisy, but effective integration can yield precise predictions without sacrificing valuable information in the data. All of these methods require a gold standard, which is a trusted representation of the functional information one might hope to discover. Such a standard, coupled with an effective means of evaluation, can be used to assess the performance of a method and serves as a basis for comparison with existing approaches. Beyond methods for predicting protein function or interactions, evaluation against gold standards can be used to directly measure the quality of a single genomic dataset, a necessary step in developing and validating new experimental technology.

Current evaluation approaches are inconsistent and biased

Myers *et al.*[42] report a study of proposed standards and approaches to evaluation of functional genomic data. They find that current approaches are inconsistent, making reported results incomparable and often biased in such a way that the resulting evaluation cannot be interpreted even in a qualitative sense. The majority of current evaluation approaches are performed without regard to which biological processes are represented in the set of correctly predicted examples, and thus they are often unknowingly skewed toward particular processes. To address these problems, Myers *et al.*[42] develop an expert-curated functional genomics standard based on the Gene Ontology by letting a panel of experts select GO terms with enough specificity that predictions based on them could be used to formulate detailed biological hypotheses, which could be confirmed or refuted by laboratory experiments.

Summary and outlook

Recovering networks of interactions between genes and gene products is a key challenge in present-day molecular biology. To achieve this task, high-throughput experimental techniques must be combined with statistical modelling and computational analysis. As more large-scale functional data become available, integrated analysis techniques will become more and more important.

To understand the complexity of living cells we need to build models including all levels of cellular organisation: we must draw information from the genome, the transcriptome and the proteome. Computational inference on parts of the system will not provide the mechanistic insights functional genomics is seeking for. However, these models will still be fragmentary if they do not include (and predict) phenotypical changes of interventions perturbing the normal course of action in the cell. Thus, an important future research direction will be to combine data sources like protein–protein interaction or transcription factor–DNA binding screens, which tell us directly about the interactions between biological molecules, with functional data from perturbation screens that carry only indirect information of gene and protein interactions.

To develop such methodology, an expert curated gold standard is necessary for an accurate understanding of how well computational methods for cellular network prediction perform. Representative evaluation of computational approaches and high throughput experimental technologies is imperative to harness the full potential of biological data in the post-genome era.

References

1 E. Segal, N. Friedman, N. Kaminski, A. Regev and D. Koller, From signatures to models: understanding cancer using microarrays, *Nat. Genet.*, 2005, **37**, S38–45.
2 N. Friedman, Inferring cellular networks using probabilistic graphical models, *Science*, 2004, **303**(5659), 799–805.
3 P. D'Haeseleer, S. Liang and R. Somogyi, Genetic network inference: from co-expression clustering to reverse engineering, *Bioinformatics*, 2000, **16**(8), 707–726.
4 G. W. Carter, Inferring network interactions within a cell, *Brief Bioinfo.*, 2005, **6**(4), 380–389.
5 J. Mandel, N. M. Palfreyman, J. A. Lopez and W. Dubitzky, Representing bioinformatics causality, *Brief Bioinfo.*, 2004, **5**(3), 270–283.
6 M. Schena, D. Shalon, R. W. Davis and P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 1995, **270**(5235), 467–470.
7 L. Wodicka, H. Dong, M. Mittmann, M. H. Ho and D. J. Lockhart, Genome-wide expression monitoring in Saccharomyces cerevisiae, *Nat. Biotechnol.*, 1997, **15**(13), 1359–67.
8 M. B. Eisen, P. T. Spellman, P. O. Brown and D. Botstein, Cluster analysis and

display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**(25), 14863–14868.
9 P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein and B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell*, 1998, **9**(12), 3273–97.
10 J. M. Stuart, E. Segal, D. Koller and S. K. Kim, A gene-coexpression network for global discovery of conserved genetic modules, *Science*, 2003, **302**(5643), 249–255.
11 C. J. Wolfe, I. S. Kohane and A. J. Butte, Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks, *BMC Bioinfo.*, 2005, **6**, 227.
12 D. Pe'er, Bayesian network analysis of signaling networks: a primer, *Sci. STKE*, 2005, **2005**(281), pl4.
13 J. Schäfer and K. Strimmer, An empirical Bayes approach to inferring large-scale gene association networks, *Bioinformatics*, 2005, **21**(6), 754–764.
14 J. Schafer and K. Strimmer, A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics, *Stat. Appl. Genet. Mol. Biol.*, 2005, **4**(1), Article32.
15 K. Basso, Reverse engineering of regulatory networks in human B cells, *Nat. Genet.*, 2005, **37**(4), 382–90.
16 N. Friedman, Using Bayesian networks to analyze expression data, *J. Comput. Biol.*, 2000, **7**(3–4), 601–620.
17 E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data, *Nat. Genet.*, 2003, **34**(2), 166–176.
18 A. V. Werhli, M. Grzegorczyk and D. Husmeier, Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks, *Bioinformatics*, 2006, **22**(20), 2523–2531.
19 C. Huttenhower, M. Hibbs, C. Myers and O. G. Troyanskaya, A scalable method for integration and functional analysis of multiple microarray data sets, *Bioinformatics*, 2006, **22**(23), 2890–2897.
20 M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat. Genet.*, 2000, **25**(1), 25–29.
21 H. W. Mewes, C. Amid, R. Arnold, D. Frishman, U. Gü ldener, G. Mannhaupt, M. Münsterkötter, P. Pagel, N. Strack, V. Stümpflen, J. Warfsmann and A. Ruepp, MIPS: analysis and annotation of proteins from whole genomes, *Nucleic Acids Res.*, 2004, **32**(Database issue), D41–4.
22 C. L. Myers, D. Robson, A. Wible, M. A. Hibbs, C. Chiriac, C. L. Theesfeld, K. Dolinski and O. G. Troyanskaya, Discovery of biological networks from diverse functional genomic data, *GenomeBiology*, 2005, **6**(13), R114.
23 Y. Yamanishi, J. P. Vert and M. Kanehisa, Protein network inference from multiple genomic data: a supervised approach, *Bioinformatics*, 2004, **20**, I363–I370.
24 K. Tsuda and K. Asai, Selective integration of multiple biological data for supervised network inference, *Bioinformatics*, 2005, **21**(10), 2488–2495.
25 Z. Barutcuoglu, R. E. Schapire and O. G. Troyanskaya, Hierarchical multi-label prediction of gene function, *Bioinformatics*, 2006, **22**(7), 830–836.
26 B. Schoelkopf, A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.
27 A. Fire, Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans, *Nature*, 1998, **391**(6669), 806–811.
28 T. R. Hughes, M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard and S. H. Friend, Functional discovery via a compendium of expression profiles, *Cell*, 2000, **102**(1), 109–126.
29 M. Boutros, Genome-wide RNAi analysis of growth and viability in Drosophila cells, *Science*, 2004, **303**(5659), 832–835.
30 E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M'Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens, M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Veronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmerman, P. Philippsen, M. Johnston and R. W. Davis, Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis, *Science*, 1999, **285**(5429), 901–906.
31 L. Avery and S. Wasserman, Ordering gene function: the interpretation of epistasis in regulatory hierarchies, *Trends Genet.*, 1992, **8**(9), 312–6.
32 A. H. Tong, Systematic genetic analysis with ordered arrays of yeast deletion mutants, *Science*, 2001, **294**(5550), 2364–2368.
33 A. H. Tong, Global mapping of the yeast genetic interaction network, *Science*, 2004, **303**(5659), 808–813.
34 R. Kelley and T. Ideker, Systematic interpretation of genetic interactions using protein networks, *Nat. Biotechnol.*, 2005, **23**(5), 561–566.
35 B. L. Drees, V. Thorsson, G. W. Carter, A. W. Rives, M. Z. Raymond, I. Avila-Campillo, P. Shannon and T. Galitski, Derivation of genetic interaction networks from quantitative phenotype data, *Genome Biol.*, 2005, **6**(4), R38.
36 M. Boutros, H. Agaisse and N. Perrimon, Sequential activation of signaling pathways during innate immune responses in Drosophila, *Dev. Cell*, 2002, **3**(5), 711–22.
37 L. M. Raamsdonk, B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. C. Walsh, J. A. Berden, K. M. Brindle, D. B. Kell, J. J. Rowland, H. V. Westerhoff, K. van Dam and S. G. Oliver, A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations, *Nat. Biotechnol.*, 2001, **19**(1), 45–50.
38 J. A. Brown, G. Sherlock, C. L. Myers, N. M. Burrows, C. Deng, H. I. Wu, K. E. McCann, O. G. Troyanskaya and J. M. Brown, Global analysis of gene function in yeast by quantitative phenotypic profiling, *Mol. Syst. Biol.*, 2006, **2**, 0001.
39 Y. Ohya, J. Sese, M. Yukawa, F. Sano, Y. Nakatani, T. L. Saito, A. Saka, T. Fukuda, S. Ishihara, S. Oka, G. Suzuki, M. Watanabe, A. Hirata, M. Ohtani, H. Sawai, N. Fraysse, J. P. Latge, J. M. Francois, M. Aebi, S. Tanaka, S. Muramatsu, H. Araki, K. Sonoike, S. Nogami and S. Morishita, High-dimensional and large–scale phenotyping of yeast mutants, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**(52), 19015–19020.
40 N. Van Driessche, J. Demsar, E. O. Booth, P. Hill, P. Juvan, B. Zupan, A. Kuspa and G. Shaulsky, Epistasis analysis with global transcriptional phenotypes, *Nat. Genet.*, 2005, **37**(5), 471–477.
41 F. Markowetz, J. Bloch and R. Spang, Non-transcriptional pathway features reconstructed from secondary effects of RNA interference, *Bioinformatics*, 2005, **21**(21), 4026–4032.
42 C. L. Myers, D. R. Barrett, M. A. Hibbs, C. Huttenhower and O. G. Troyanskaya, Finding function: evaluation methods for functional genomic data, *BMC Genomics*, 2006, **7**, 187.

