Minireview

Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets

Vincent Detours[a,*], Jacques E. Dumont[a], Hugues Bersini[b], Carine Maenhaut[a]

[a] *IRIBHM, Free University of Brussels, Bldg C, Campus Erasme, 808 route de Lennik, B-1070 Brussels, Belgium*
[b] *IRIDIA, Free University of Brussels, 50 Av. F. Roosevelt, B-1050 Brussels, Belgium*

Received 16 April 2003; revised 23 April 2003; accepted 23 April 2003

First published online 16 May 2003

Edited by Richard Marais

Abstract Data analysis – not data production – is becoming the bottleneck in gene expression research. Data integration is necessary to cope with an ever increasing amount of data, to cross-validate noisy data sets, and to gain broad interdisciplinary views of large biological data sets. New Internet resources may help researchers to combine data sets across different gene expression platforms. However, noise and disparities in experimental protocols strongly limit data integration. A detailed review of four selected studies reveals how some of these limitations may be circumvented and illustrates what can be achieved through data integration.

Key words: Bioinformatics; Microarray;
Serial analysis of gene expression; Data integration

1. Why integration of expression data?

Functional genomics is a new field of research emerging from full genome sequencing and from new technologies which make it possible to quantify mRNA transcripts on a genome-wide scale in any cell or tissue type. The major gene expression platforms, Serial Analysis of Gene Expression (SAGE) [1] and cDNA microarrays [2], were first proposed in 1995, followed by oligonucleotide arrays in 1996 [3]. Since then, the publication rate in the field has grown exponentially, reaching 2000 papers for microarrays alone in 2001 [4]. Applications encompass drug development [5], yeast biology [6], vaccine design [7], cancer research [8], etc. – they are too numerous to be listed here.

Because they are readily available in electronic format, sequence data and expression data open the door to data and knowledge integration on a scale unprecedented in the history of biology. Molecular biology would benefit if expression data produced by different groups on different systems could be compared. For example, in cancer research the integration of high-throughput expression data sets presents an exciting opportunity to transcend the frontiers in terms of cancer types that have traditionally fragmented the field. What is common between cancers? Do metastatic cells arise from the same processes – and thus share potential drug targets – in different tumors? Is it possible to classify cancers on the basis of their expression profile? Such questions are coming within reach.

High-throughput gene expression data sets are subject to noise and error [9,10]. This is compounded by the statistical difficulties raised by massive multiple hypothesis testing when identifying differentially expressed genes [10]. Thus, independent validation of differential expression is required before drawing conclusions from microarray or SAGE experiments. Polymerase chain reaction (PCR)-based protocols or Northern blots used for this purpose measure expression on a per-gene basis. Only a dozen among hundreds of putatively differentially expressed genes are validated in a typical study. The inherent power of high-throughput technology is not matched by high-throughput validation [11,12]. Comparing data sets produced by different groups on different platforms could increase confidence in expression results for many more genes than is tractable with classical validation [11,12].

Although validation of expression results could be improved, a dozen genes of potential interest are identified and confirmed in a typical microarray study. Hundreds of such studies are published every year. This is much more information than the research community can possibly follow up with detailed experimental studies. Furthermore, it is doubtful whether investigators producing data on a high-throughput scale squeeze all of the information buried in it [13]. This is information overload [14]: data analysis, not data generation, is becoming the main bottleneck. Re-use and integration could help researchers to form the comprehensive views of existing data needed to better prioritize experimental efforts.

This minireview addresses integration of expression data from a practical perspective. The major resources available to match the probes of heterogeneous platforms are presented. Next, we review what is known about results reproducibility, hence comparability, within and across platforms. Finally, we analyze the design of four particularly successful studies, and derive strategies to overcome noise and platform heterogeneity issues that may prevent the integration of expression data.

*Corresponding author.
E-mail address: vdetours@ulb.ac.be (V. Detours).

Abbreviations: GO, Gene Ontology; MAGE-ML, Microarray Gene Expression Mark-up Language; MIAME, Minimum Information About a Microarray Experiment; PPI, protein–protein interaction; PCR, polymerase chain reaction; SAGE, Serial Analysis of Gene Expression

0014-5793/03/$22.00 © 2003 Published by Elsevier Science B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/S0014-5793(03)00522-2

2. Data, standards, and tools

Large public gene expression databases are already, or will soon be, operational (Table 1), and many groups publish expression data on their web sites. Hundreds of data sets are readily available. *Cell* and journals of the *Nature* publishing group require authors to make their expression data publicly available, more journals are considering a similar move as databases become more user-friendly [15].

The substantial – and fast growing – number of publicly available data sets is of limited value, however, as long as a number of compatibility issues are not resolved. Investigators use different platforms, different sample preparation protocols, different data formats, and different data normalization algorithms, making it difficult to compare data sets. Data format standardization is being addressed at an international level by the Microarray Gene Expression Data Society [16]. The Minimum Information About a Microarray Experiment (MIAME) guidelines aim at unambiguously interpreting microarray data and at subsequently allowing independent verification [17]. MIAME compliance is mandatory for publication of microarray data in *Cell*, *The Lancet*, *Nature* journals and *Science*. The Microarray Gene Expression Mark-up Language (MAGE-ML) provides the formal infrastructure to exchange and store MIAME-compliant data [16]. Major gene expression public databases are using MAGE-ML, or will use it in the near future (A. Brazma, ArrayExpress, personal communication).

Data storage and exchange are the first steps toward data integration. The data sets to be pooled use different probe sequences and formats: nonamer tags for SAGE, spotted cDNA sequences or oligonucleotides for microarrays. In addition, mutants or orthologs of the same genes may be used in the data sets to be integrated. Thus, a next step is to construct maps between disparate sets of probes in order to compare data sets. Several resources useful to this end are available over the Internet (Table 1). RESOURCERER [18] provides maps between the probe sets of various commercial and noncommercial platforms which encompass human, mouse and rat. SOURCE [19] and EnsMart automate the construction of association tables between various database identification numbers and sequence annotations. For example, the user may submit a list of GenBank IDs to EnsMart and request the corresponding Affymetrix® probe IDs, LocusLink IDs, UniGene IDs, Gene Ontology information, etc. UniGene [20] is a grouping of GenBank mRNA entries by sequence similarity. A set of similar sequences is called a cluster. Maps may be constructed by associating probes that belong to the same UniGene cluster, i.e. with same UniGene ID. LocusLink [21] provides locus-to-sequence and to-UniGene associations useful for cross-species mapping. Once the probes of the data sets under investigation have been related to one another, one can proceed to further comparison.

3. How comparable are expression data sets?

Constructing probe libraries, printing arrays, collecting samples and hybridizing them require many steps. A range of technical options (reviewed in [22]) which may impact the final result are available at each of these steps, and each one of them may incur noise [23]. The complexity of the protocols together with the measurement errors may undermine cross-platform integration. If microarrays and SAGE measure anything objective, however, one may expect that results obtained for mRNA samples collected under one particular biological condition and assessed with one platform carry over to other platforms for biologically comparable samples. To what extent is this verified?

High reproducibility of measurements within a given platform has been reported for cDNA microarrays [24,25], Affymetrix® oligonucleotide microarrays [23,26], and SAGE [27,28]. None of these studies compared samples and hybridization procedures from different laboratories. To the best of our knowledge, how the reported reproducibility carries over to experiments performed in different laboratories has not been systematically investigated. In the case of microarrays, comparison between arrays of the same type cannot detect sequence errors. Up to 30% of the spotted probes did not match the expected gene sequence on arrays from one vendor [29]. This shortcoming highlights the relevance of cross-platform validation. High-throughput experiments lead to useful fingerprinting of tissues. In this case the comprehensiveness of the probe set and the exact identity of the represented genes are less crucial. This is not the case, however, if detailed biological insights are to be obtained.

Surprisingly, little information is available regarding agreement between platforms. A Spearman correlation of 0.8 was found between the 200 most differentially expressed transcripts in an Affymetrix® microarray analysis and the corresponding transcripts in a SAGE library [30]. In another study, SAGE tags present at a frequency of 6 in 76 000 or more were also detected in more than 80% of 43 Affymetrix® microarrays [31]. Taken together, these results suggest a good agreement between these two platforms at medium to high expression levels.

Yuen et al. [32] compared Affymetrix® and cDNA microarrays by focusing on 47 genes which were also tested with quantitative real-time PCR. Both platforms detected 16 out of 17 up- or down-regulated genes, and no non-regulated genes were detected as regulated. Both platforms underestimated fold change, although cDNA arrays did so in a predictable and correctable way. By contrast, Kuo et al. [33] found a poor correlation between Affymetrix® and cDNA microarray data. They used the raw, non-normalized, data generated by spot quantification softwares. Better correlation might have been found with appropriate normalizations. Unlike Yuen et al. [32], the data sets they compared originated from two unrelated laboratories.

The dual-channel cDNA technology [2] measures simultaneously hybridization of spotted probes with the target mRNA preparation and with a reference mRNA preparation. Many options are available to prepare reference mRNA. For example, in a tumor vs. normal tissue set-up, some groups compare expression in tumor with expression in normal tissues from the same individual, others use pooled normal tissues from several individuals, still others favor tightly calibrated custom-made mRNA preparations, depending on the questions asked. This diversity of protocols is an hindrance when comparing data sets.

Overall, the limited number of formal cross-platform comparison studies suggests that data may be comparable at medium and high expression levels. In addition, reference disparities will limit comparison in the case of cDNA arrays. We suggest that any intergroup collaboration should begin with

Table 1
Selected web sites

Major databases for high-throughput gene expression	
ArrayExpress (European Bioinformatics Institute)	http://www.ebi.ac.uk/arrayexpress
Cancer Genome Anatomy Project (National Cancer Institute)	http://www.cgap.nci.nih.gov
Children National Medical Center Microarray Center	http://www.microarray.cnmcresearch.org/pgadatatable.asp
Gene Expression Omnibus (National Center for Biotechnology Information)	http://www.ncbi.nih.gov/geo
Stanford Microarray Database	http://www.genome-www5.stanford.edu/MicroArray/SMD
Data standardization	
Microarray Gene Expression Data Society	http://www.mged.org
Major tools to match (and annotate) probes	
EnsMart (EMBL and Wellcome Trust Sanger Center)	http://www.ensembl.org/EnsMart
LocusLink (NCBI)	http://www.ncbi.nlm.nih.gov/LocusLink
RESOURCERER (The Institute for Genomic Research)	http://www.pga.tigr.org/tigr-scripts/magic/r1.pl
SOURCE (Stanford University)	http://www.source.stanford.edu/cgi-bin/sourceSearch
UniGene (NCBI)	http://www.ncbi.nih.gov/UniGene

method standardization and cross-validation of results obtained on the same cellular material.

4. Four case studies

Although quantitative information is lacking about the degree of correlation between platforms, several groups have been successful at integrating expression data sets.

Rhodes et al. [12] pooled four published and publicly available prostate cancer data sets generated by independent laboratories using Affymetrix® or dual-channel cDNA microarrays. The null hypothesis that differentially expressed genes differ among studies was tested for each combination of two, three, and four data sets. For each combination of data sets a gene-specific false discovery rate, i.e. an estimate of the probability of finding differential expression of that gene by chance, was computed for all genes present in all data sets in the combination. As expected, combination with more data sets led in general to increased confidence in differential expression estimates. About 90 genes were found differentially expressed with a false discovery rate under 0.05. Validating 90 genes with classical PCR-based protocols would be costly and cumbersome. The approach of Rhodes et al. [12] re-uses public data and is automated.

Ramaswamy et al. [34] searched predictors of metastasis in primary tumors by analyzing five published microarray data sets generated by independent laboratories on several Affymetrix® platforms, and the Rosetta® inkjet platform. They compared the expression profiles of 12 metastatic adenocarcinomas of diverse tissue origin with the profiles of 64 primary tumors. One hundred and twenty-eight genes distinguished metastatic from primary tumors. Remarkably, some primary tumors showed a expression profile characteristic of metastatic tumors, leading the investigators to the hypothesis that early signs of metastasis could be present in some primary tumors. To test this hypothesis, they took another data set from 62 stage I and II primary lung tumors and applied hierarchical clustering in the space of the 128 metastasis-specific genes, i.e. all other genes were ignored. Two main clusters highly correlated with the original primary vs. metastatic tumor distinction were found. Patients whose primary tumor bore the metastasis signature had a significantly shorter survival time. Further refinement of the model reduced the 128 gene signature to 17 genes. This simplified signature could reproduce the lung tumor data set results on other data sets: 21 prostate cancer samples, 78 small stage I breast carcinomas, and 60 medulloblastomas. Interestingly, the signature did not correlate with survival in 58 diffuse large B cell lymphoma samples, in line with the view that hematopoietic tumor cells use specific navigation mechanisms. The metastasis-specific signature was found from mRNA samples extracted from many cells – a result also supported by an earlier breast cancer study [35]. This finding challenges the widely held view that metastasis arises from rare cells with metastatic potential, and supports the view that many cells in primary tumors have this potential [36]. It is further shown that the signature applies to many types of cancers.

Expression data may also be integrated with other types of high-throughput data. Thousands of protein–protein interactions (PPI) have been detected in yeast with two-hybrid assays [37,38] and mass spectroscopy [39,40]. Little is known about the artifacts of these methods. Remarking that proteins can interact in a cell only if they are co-expressed, Kemmeren et al. [11] integrated these data with several published yeast gene expression data sets in order to increase confidence in PPI data. They first selected one two-hybrid data set and a set of expression profiles from one study and used them as a test bed to develop a method to correlate PPI and co-expression. Next, they applied this method to all the PPI data and 326 expression profiles compiled from five unrelated studies addressing diverse issues such as cell cycle, response to unfolded proteins, or response to pheromone treatment. A few hundred PPIs in the two-hybrid and mass spectroscopy data sets had been previously established through independent investigations. Depending on the data set and the selected confidence threshold, 54–71% of these PPI exhibited co-expression. This brings credence to the approach. Since protein expression can be controlled at the post-transcriptional level, a complete concordance is not expected. Out of 5342 two-hybrid-derived PPIs, 973 were also co-expressed, and may be considered functional with increased confidence. One of the protein partners had a functional annotation while the other did not in 328 of these 973 PPIs. Kemmeren et al. propose transferring the functional annotation of the known protein to the other, hypothesizing that two interacting proteins are most likely involved in the same biological process. To test this idea they deleted the gene encoding the non-annotated protein partner of an annotated protein for five PPIs.

In all cases the deletion resulted in the same phenotype as observed following deletion of the annotated partner, suggesting that annotation can be transferred.

Integration of gene expression data sets and genome maps has established the view that expressed genes are not randomly distributed along chromosomes, they form clusters [41–44]. In the arguably most quantitative study on this subject so far, Spellman and Rubin [44] compiled expression data from 88 unrelated experimental conditions in *Drosophila* assessed on 267 Affymetrix® microarrays from six independent laboratories, and mapped them to the *Drosophila* chromosome maps. They calculated the pair-wise correlation of expression of adjacent genes across the 88 conditions. These correlations were then averaged over windows sliding along the chromosome. Windows of different sizes were tested in order to measure the length of stretches of co-expressed genes. About 200 clusters of 10–30 co-regulated genes, representing 20% of all genes, span the *Drosophila* genome. Spellman and Rubin repeated their analysis on a data set from which homologous genes located in a same neighborhood in chromosome maps were removed. The number of clusters dropped to 176. Thus, recent gene duplication and function similarity between the resulting homologs is not a general explanation for the observed 200 clusters. Next, they mapped each gene to its Gene Ontology (GO) functional categories. The Gene Ontology [45] is a human-curated structured vocabulary which describes gene products in terms of chemical and biological functions, and of cellular location. Genes within a given cluster were not biased toward a same set of functional GO categories, suggesting that they are not functionally related. Calculation of correlations between cluster location and map of *Drosophila* polytene chromosomes failed to identify a relation with band morphology in polytene chromosomes. Taken together, these results led Spellman and Rubin to the provocative conclusion that clusters occur because transcription is sloppy: a by-product of the unfolding of the chromatin structure around a gene being expressed is the expression of its neighbors. This effect may contribute to illegitimate transcription. From the point of view of data integration, expression data were mapped on genome maps, the resulting cluster structure was then correlated with sequence data, GO categories, and chromosome band morphology. All these calculations were performed on a genome-wide scale.

5. Strategies to cope with noisy and heterogeneous data sets

The noise inherent in high-throughput methods and the heterogeneity of experimental protocols may obscure biologically relevant relations between data sets and preclude data integration. The studies reviewed above cope with these issues. In the following, we extract the general strategies that were used for doing so.

Rhodes et al. [12] focused on genes differentially expressed in several data sets. These data sets represent the same biological condition, prostate cancer, but rely on different platforms. Differential expression was assessed independently for each gene in each data set. The calculation grouping results from the various data sets operated on a statistical confidence measure, *p*-value, not on expression level. This strategy avoids direct comparisons of data sets and related cross-platform normalization issues.

As mentioned earlier, Ramaswamy et al. [34] identified from one data set a small set of genes distinguishing primary tumors with metastatic potential from primary tumors without such potential. They then used it to cluster data sets in order to assess its validity and generality across many types of cancers. Focusing on a small number of genes dramatically reduced the search space when performing clustering, making it more likely to find relevant cluster structures. In addition, data sets were not pooled together. Ramaswamy et al. [34] compared qualitatively the conceptual end-results of independent analysis of each data set, namely the applicability of the signature to the data. Reducing data, i.e. lowering resolution, is likely to have an averaging effect making integration less sensitive to noise.

The studies of Kemmeren et al. [11] and Spellman and Rubin [44] relied on co-expression of genes over a compendium [46], i.e. a collection of expression profiles measured under diverse and unrelated biological conditions. Whether expression is influenced by the details of cell culture protocols or by interesting biological features is irrelevant as long as it reflects cell function. Although correlation calculations may still be obscured by differences in mRNA preparation and/or hybridization between data sets, it is resistant to differences in downstream procedures on living material.

Most importantly, the specifics of various platforms and laboratory procedures are much more likely to introduce differences between data sets than to introduce similarities. Thus, the cross-platform, cross-study, approach is conservative when it comes to discovering properties shared by data sets. All four studies presented above focus on shared properties: expression patterns shared by genes [11,44], genes differentially expressed common to data sets [12], and the universality of a signature of metastasis among cancers [34]. Because data integration is about elaborating synthetic views, it naturally brings focus on similarity between data sets. The resulting knowledge is more reliable because shared features are more likely to be robust with respect to noise and to the specifics of experimental procedures.

Acknowledgements: We thank David Venet for helpful comments on an earlier version of the paper. This work was supported by the European Union, Ministère de la Recherche Scientifique (PAI, Belgium), Fond National de la Recherche Scientifique, UCB-Pharma, Région Wallone, Fédération Contre le Cancer, and Fondation Weiner.

References

[1] Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Science 270, 484–487.
[2] Schena, M., Shalon, D., Davis, R.W. and Brown, P.O. (1995) Science 270, 467–470.
[3] Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H. and Brown, E.L. (1996) Nat. Biotechnol. 14, 1675–1680.
[4] Afshari, C.A. (2002) Endocrinology 143, 1983–1989.
[5] Gerhold, D.L., Jensen, R.V. and Gullans, S.R. (2002) Nat. Genet. 32 (Suppl.), 547–551.
[6] Banerjee, N. and Zhang, M.Q. (2002) Curr. Opin. Microbiol. 5, 313–317.
[7] Dhiman, N., Bonilla, R., O'Kane, D.J. and Poland, G.A. (2001) Vaccine 20, 22–30.
[8] Chung, C.H., Bernard, P.S. and Perou, C.M. (2002) Nat. Genet. 32 (Suppl.), 533–540.
[9] Quackenbush, J. (2002) Nat. Genet. 32 (Suppl.), 496–501.
[10] Slonim, D.K. (2002) Nat. Genet. 32 (Suppl.), 502–508.
[11] Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders,

R., Brazma, A. and Holstege, F.C. (2002) Mol. Cell 9, 1133–1143.

[12] Rhodes, D.R., Barrette, T.R., Rubin, M.A., Ghosh, D. and Chinnaiyan, A.M. (2002) Cancer Res. 62, 4427–4433.

[13] Kling, J. (2002) Scientist 16, 34.

[14] Fogarty, M. and Bahls, C. (2002) Scientist 16, 16.

[15] DeFrancesco, L. (2002) The Scientist Daily News, October 10th.

[16] Spellman, P.T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., Swiatek, M., Marks, W.L., Goncalves, J., Markel, S., Iordan, D., Shojatalab, M., Pizarro, A., White, J., Hubley, R., Deutsch, E., Senger, M., Aronow, B.J., Robinson, A., Bassett, D., Stoeckert Jr., C.J. and Brazma, A. (2002) Genome Biol. 3, RESEARCH0046.

[17] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J. and Vingron, M. (2001) Nat. Genet. 29, 365–371.

[18] Tsai, J., Sultana, S., Lee, Y., Pertea, G., Karamycheva, S., Antonescu, V., Cho, J., Parvizi, B., Cheung, F. and Quackenbush, J. (2001) Genome Biol. 2, software0002.1–0002.4.

[19] Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O. and Alizadeh, A.A. (2003) Nucleic Acids Res. 31, 219–223.

[20] Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A. and Wagner, L. (2003) Nucleic Acids Res. 31, 28–33.

[21] Pruitt, K.D. and Maglott, D.R. (2001) Nucleic Acids Res. 29, 137–140.

[22] Holloway, A.J., van Laar, R.K., Tothill, R.W. and Bowtell, D.D. (2002) Nat. Genet. 32 (Suppl.), 481–489.

[23] Tu, Y., Stolovitzky, G. and Klein, U. (2002) Proc. Natl. Acad. Sci. USA 99, 14031–14036.

[24] Yue, H., Eastman, P.S., Wang, B.B., Minor, J., Doctolero, M.H., Nuttall, R.L., Stack, R., Becker, J.W., Montgomery, J.R., Vainer, M. and Johnston, R. (2001) Nucleic Acids Res. 29, e41.

[25] Yang, I.V., Chen, E., Hasseman, J.P., Liang, W., Frank, B.C., Wang, S., Sharov, V., Saeed, A.I., White, J., Li, J., Lee, N.H., Yeatman, T.J. and Quackenbush, J. (2002) Genome Biol. 3, research0062.

[26] Naef, F., Hacker, C.R., Patil, N. and Magnasco, M. (2002) Genome Biol. 3, RESEARCH0018.

[27] Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett Jr., D.E., Hieter, P., Vogelstein, B. and Kinzler, K.W. (1997) Cell 88, 243–251.

[28] Blackshaw, S., Kuo, W.P., Park, P.J., Tsujikawa, M., Gunnersen, J.M., Scott, H.S., Boon, W.M., Tan, S.S. and Cepko, C.L. (2003) Genome Biol. 4, R17.

[29] Kothapalli, R., Yoder, S.J., Mane, S. and Loughran Jr., T.P. (2002) BMC Bioinform. 3, 22.

[30] Ishii, M., Hashimoto, S., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T. and Aburatani, H. (2000) Genomics 68, 136–143.

[31] Evans, S.J., Datson, N.A., Kabbaj, M., Thompson, R.C., Vreugdenhil, E., De Kloet, E.R., Watson, S.J. and Akil, H. (2002) Eur. J. Neurosci. 16, 409–413.

[32] Yuen, T., Wurmbach, E., Pfeffer, R.L., Ebersole, B.J. and Sealfon, S.C. (2002) Nucleic Acids Res. 30, e48.

[33] Kuo, W.P., Jenssen, T.K., Butte, A.J., Ohno-Machado, L. and Kohane, I.S. (2002) Bioinformatics 18, 405–412.

[34] Ramaswamy, S., Ross, K.N., Lander, E.S. and Golub, T.R. (2003) Nat. Genet. 33, 49–54.

[35] van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R. and Friend, S.H. (2002) Nature 415, 530–536.

[36] Bernards, R. and Weinberg, R.A. (2002) Nature 418, 823.

[37] 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., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.M. (2000) Nature 403, 623–627.

[38] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001) Proc. Natl. Acad. Sci. USA 98, 4569–4574.

[39] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, 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., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D. and Tyers, M. (2002) Nature 415, 180–183.

[40] Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, 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) Nature 415, 141–147.

[41] Cohen, B.A., Mitra, R.D., Hughes, J.D. and Church, G.M. (2000) Nat. Genet. 26, 183–186.

[42] Caron, H., van Schaik, B., van der Mee, M., Baas, F., Riggins, G., van Sluis, P., Hermus, M.C., van Asperen, R., Boon, K., Voute, P.A., Heisterkamp, S., van Kampen, A. and Versteeg, R. (2001) Science 291, 1289–1292.

[43] Lercher, M.J., Urrutia, A.O. and Hurst, L.D. (2002) Nat. Genet. 31, 180–183.

[44] Spellman, P.T. and Rubin, G.M. (2002) J. Biol. 1, 5.

[45] The Gene Ontology Consortium (2001) Genome Res. 11, 1425–1433.

[46] Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. and Friend, S.H. (2000) Cell 102, 109–126.