

Chapter 12

Expression and Microarrays

Joaquín Dopazo and Fátima Al-Shahrour

Abstract

High throughput methodologies have increased by several orders of magnitude the amount of experimental microarray data available. Nevertheless, translating these data into useful biological knowledge remains a challenge. There is a risk of perceiving these methodologies as mere factories that produce never-ending quantities of data if a proper biological interpretation is not provided.

Methods of interpreting these data are continuously evolving. Typically, a simple two-step approach has been used, in which genes of interest are first selected based on thresholds for the experimental values, and then enrichment in biologically relevant terms in the annotations of these genes is analyzed in a second step. For various reasons, such methods are quite poor in terms of performance and new procedures inspired by systems biology that directly address sets of functionally related genes are currently under development.

Key words: Functional interpretation, functional genomics, multiple testing, gene ontology.

1. Introduction

Genes operate within the cell in an intricate network of interactions that is only recently starting to be envisaged (1–3). It is a widely accepted fact that co-expressed genes tend to play common roles in the cell (4, 5). In fact, this causal relationship has been used to predict gene function from patterns of co-expression (6, 7).

In this scenario, a clear necessity exists for methods and tools that can help to understand large-scale experiments (microarrays, proteomics, etc.) and formulate genome-scale hypotheses from a systems biology perspective (8). Dealing with genome-scale data in this context requires the use of functional annotations of the genes, but this step must be approached from within a systems

biology framework in which the collective properties of groups of genes are considered.

DNA microarray technology can be considered the dominant paradigm among genome-scale experimental methodologies. Although many different biological questions can be addressed through microarray experiments, three types of objectives are typically undertaken in this context: class comparison, class prediction, and class discovery (9). The two first objectives fall in the category of supervised methods and usually involve the application of tests to define differentially expressed genes, or the application of different procedures to predict class membership on the basis of the values observed for a number of key genes. Clustering methods belong to the last category, also known as unsupervised analysis because no previous information about the class structure of the data set is used.

The extensive use of microarray technology has fueled the development of functional annotation tools that essentially study the enrichment of functional terms in groups of genes defined by the experimental values. Examples of such terms with functional meaning are gene ontology (GO) (10), KEGG pathways (11), CisRed motifs (12), predictions of transcription factor binding sites (13), Interpro motifs (14), and others. Programs such as ontoexpress (15), FatiGO (16), GOMiner (17), and others, can be considered representatives of a family of methods designed for this purpose (18). These methods are used *a posteriori* over the genes of interest previously selected in a first step, in order to obtain some clues to the interpretation of the results of microarray experiments. Typical criteria for selection are differential expression (class comparison), co-expression (class discovery), or others. By means of this simple two-step approach, a reasonable biological interpretation of a microarray experiment can be reached. Nevertheless, this approach has a weak point: the list of genes of interest. This list is generally incomplete, because its definition is affected by many factors, including the method of analysis and the threshold imposed. In the case of class discovery analysis, the use of biological annotations has also been employed as a cluster validation criterion (19).

Thus, the difficulties for defining repeatable lists of genes of interest across laboratories and platforms even using common experimental and statistical methods (20) has led several groups to propose different approaches that aim to select genes, taking into account their functional properties. The Gene Set Enrichment Analysis (GSEA) (21, 22), although not free of criticisms (23), pioneered a family of methods devised to search for groups of functionally related genes with a coordinate (although not necessarily high) over- or under-expression across a list of genes ranked by differential expression coming from microarray experiments. Different tests have recently been proposed with this aim

for microarray data (24–30) and also for ESTs (31) and some of them are available in Web servers (32, 33). In particular, the FatiScan procedure (32, 33), which implements a segmentation test (24), can deal with ordered lists of genes independently from the type of data that originated them. This interesting property allows its application to other types of data apart from microarrays. Also recently, biological information (34, 35) or phenotypic information (36) has been used as a constitutive part of clustering algorithms in the case of class discovery (clustering) analysis.

2. Methods

2.1. Threshold-Based Functional Analysis

The final aim of a typical genome-scale experiment is to find a molecular explanation for a given macroscopic observation (e.g., which pathways are affected by the deprivation of glucose in a cell, what biological processes differentiate a healthy control from a diseased case). The interpretation of genome-scale data is usually performed in two steps: In a first step genes of interest are selected (because they co-express in a cluster or they are significantly over- or under-expressed when two classes of experiments are compared), usually ignoring the fact that these genes are acting cooperatively in the cell and consequently their behaviors must be coupled to some extent (*see Note 1*). In this selection, stringent thresholds to reduce the false-positives ratio in the results are usually imposed. In a second step, the selected genes of interest are compared with the background (typically the rest of genes) in order to find enrichment in any functional term. This comparison to the background is required because otherwise the significance of a proportion (even if high) cannot be determined. The procedure is illustrated in **Fig. 12.1** for the interpretation of either co-expressing genes found by clustering (*see Fig. 12.1A*) or genes selected by differential expressing among two pre-defined classes of experiments (*see Fig. 12.1B*).

This comparison is made by means of the application of tests such as the hypergeometric, χ^2 , binomial, Fisher's exact test, and others. There are several available tools, reviewed in (18). Among these methods, the most popular ones (more cited in the literature) are Onto-express (15) (<http://vortex.cs.wayne.edu/ontocexpress/>) and FatiGO (16) (<http://www.fatigo.org>). These tools use various biological terms with functional meaning such as GO (10), KEGG pathways (11), etc.

Although this procedure is the natural choice for analyzing clusters of genes, its application to the interpretation of differential gene expression experiments causes an enormous loss of information because a large number of false-negatives are obtained in

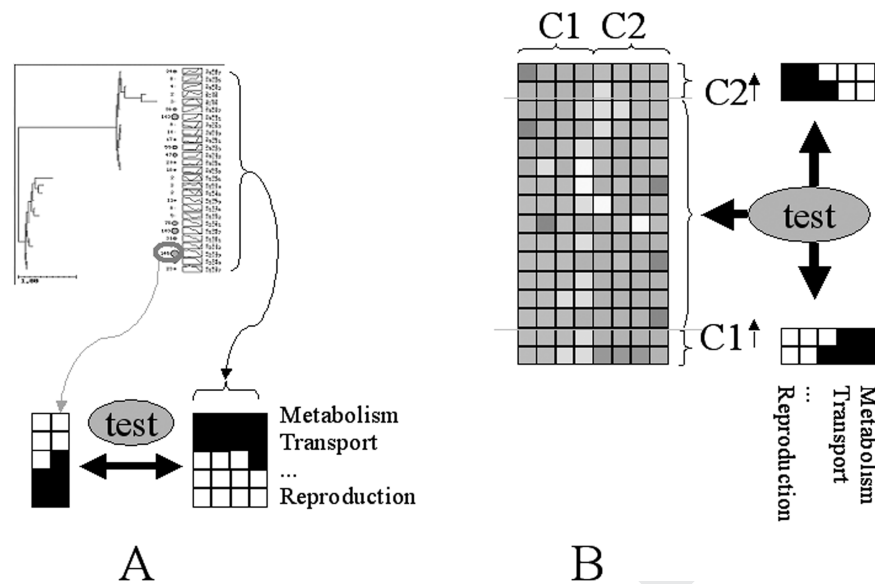


Fig. 12.1. Two-step procedure for the functional annotation of distinct microarray experiments. **A.** The unsupervised approach: functional interpretation of clusters of co-expressing genes. A cluster of genes is selected for functional analysis and the genes inside it are checked for significant enrichment of functional terms with respect to the background (the rest of genes). **B.** The supervised approach: functional annotation of genes differentially expressed between two classes (C1 and C2). The differentially expressed genes are checked for enrichment of functional terms with respect to the background, the genes not showing differential expression between (A) and (B).

order to preserve a low ratio of false-positives (and the noisier the data the worse is this effect).

2.2. Threshold-Free Functional Analysis

From a systems biology perspective, this way of understanding the molecular basis of a genome-scale experiment is far from efficient. Methods inspired by systems biology focus on collective properties of genes. Functionally related genes need to carry out their roles simultaneously in the cell and, consequently, they are expected to display a coordinated expression. Actually, it is a long recognized fact that genes with similar overall expression often share similar functions (4, 37, 38). This observation is consistent with the hypothesis of modularly behaving gene programs, where sets of genes are activated in a coordinated way to carry out functions. Under this scenario, a different class of hypotheses, not based on genes but on blocks of functionally related genes, can be tested. Thus, lists of genes ranked by any biological criteria (e.g., differential expression when comparing cases and healthy controls) can be used to directly search for the distribution of blocks of functionally related genes across the list without imposing any arbitrary threshold. Any macroscopic observation that causes this ranked list of genes will be the consequence of cooperative action of genes that are part of functional classes,

pathways, etc. Consequently, each functional class “responsible” for the macroscopic observation will be found in the extremes of the ranking with highest probability. The previous imposition of a threshold based on the rank values that does not take into account the cooperation among genes is thus avoided under this perspective. **Fig. 12.2** illustrates this concept. Genes are arranged by differential expression between the classes C1 and C2. On the right part of the figure, labels for two different functional classes have been placed at the positions in the list where genes playing the corresponding roles are situated. Function A is completely unrelated to the experiment because it appears simultaneously over-expressed in class C1 and C2 and also in intermediate positions. Conversely, function B is predominantly performed by genes with high expression in class C2, but scarcely appears in class C1. This observation clearly points to function B as one of the molecular bases of the macroscopic observation made in the experiment. Instead of trying to select genes with extreme values

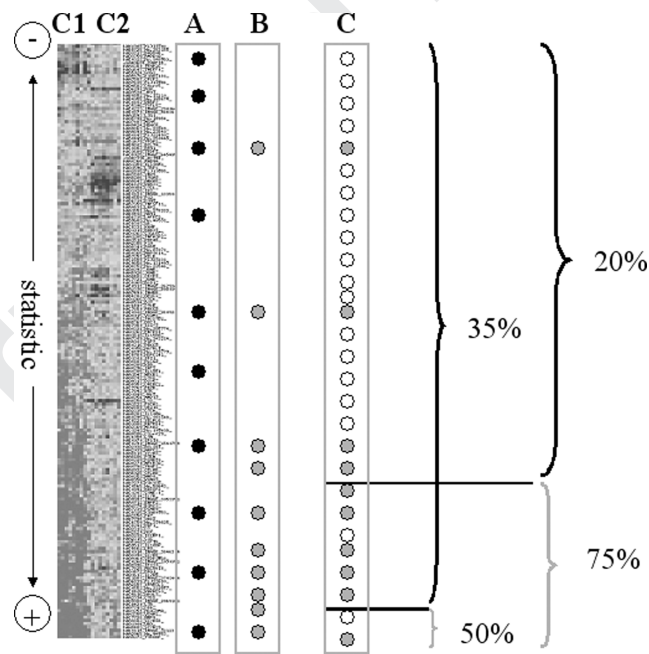


Fig. 12.2. Threshold-free procedure for the functional annotation of class comparison experiments. **A.** Functional label unrelated to the experiment from which the rank of genes was obtained. **B.** Functional label related to the experiment. **C.** Schematic representation of two partitions of the segmentation test. In the first partition, 35% of the genes in the upper segment are annotated with the term (C), whereas 50% of the genes in the lower segment are annotated with this term. This difference in percentages is not statistically significant. In the second partition, the 75% of the genes in the lower segment are annotated as (C), whereas only 20% are annotated as (C) in the upper partition. These differences in the proportions are high enough to be considered significant after the application of a Fisher’s exact test.

of differential expression, systems biology-inspired methods will directly search for blocks of functionally related genes significantly cumulated in the extremes of a ranked list of genes.

Other methods that have been proposed for this purpose, such as the GSEA (21, 22) or the SAFE (39) method, use a non-parametrical version of a Kolmogorov-Smirnov test. Other strategies are also possible, such as direct analysis of functional terms weighted with experimental data (28), Bayesian methods (29), or model-based methods (26). Conceptually simpler and quicker methods with similar accuracy have also been proposed, such as the parametrical counterpart of the GSEA, the PAGE (30) or the segmentation test, Fatiscan (24), which is discussed in the next section.

2.3. FatiScan: A Segmentation Test

A simple way of studying the asymmetrical distributions of blocks of genes across a list of them is to check if, in consecutive partitions, one of the parts is significantly enriched in any term with respect to the complementary part. Fig. 12.2C illustrates this concept with the representation of ordered genes in which gray circles represent those genes annotated with a particular biological term and open circles represent genes with any other annotation. In the first partition, 35% of the genes in the upper segment are annotated with the term of interest, whereas 50% of the genes in the lower segment are annotated with this term. This difference in percentages is not statistically significant. However, in the second partition, the differences in the proportions are high enough to be considered significant (75% vs. 20%): The vast majority of the genes with the annotation are on the lower part of the partition.

The segmentation test used for threshold-free functional interpretation consists of the sequential application of the FatiGO (16) test to different partitions of an ordered list of genes. The FatiGO test uses a Fisher's exact test over a contingency table for finding significantly over- or under-represented biological terms when comparing the upper side with the lower side of the list, as defined by any partition. Previous results show that a number between 20 and 50 partitions often gives optimal results in terms of sensitivity and results recovered (24). Given that multiple terms (T) are tested in a predefined number of partitions (P), the unadjusted p -values for a total of $T \times P$ tests must be corrected. The widely accepted FDR (40) can be used for this purpose. Nevertheless, carrying out a total of $T \times P$ tests would correspond to the most conservative scenario, in a situation in which no *a priori* functional knowledge of the system is available. Usually many terms can initially be discarded from the analysis due to prior information or just by common sense.

The FatiScan test has two fundamental advantages when compared to alternative methods based on Kolmogorov-Smirnov or related tests. On one hand, this method does not require an

extreme non-uniform distribution of genes. It is able to find different types of asymmetries in the distribution of groups of genes across the list of data. On the other hand, and more importantly, this method does not depend on the original data from which the ranking of the list was derived. The significance of the test depends only on the ranking of the genes in the list and the strategy used for performing the partitions. This means that, in addition to DNA microarray data, this method can be applied to any type of genome-scale data in which a value can be obtained for each gene. FatiScan is available within the Babelomics package (32, 33) for functional interpretation of genome-scale experiments (<http://www.babelomics.org>).

2.4. Differential Gene Expression in Human Diabetes Samples

We have used data from a study of gene expression in human diabetes (21) in which a comparison between two classes (17 controls with normal tolerance to glucose versus 26 cases composed of 8 with impaired tolerance and 18 with type 2 diabetes mellitus, DM2) did not detect even a single gene differentially expressed. We ordered the genes according to their differential expression between cases and controls. A t-test, as implemented in the T-Rex tool from the GEPAS package (41–43) was used for this purpose (see Note 2). The value of the statistic was used as the ranking criteria for ordering the list. As in the original analysis (21), we were unable to find individual genes with a significant differential expression (differentially expressed genes with an adjusted p -value < 0.05).

A total of 50 partitions of the ranked list were analyzed with the FatiScan algorithm for over- or under-expression of KEGG pathways and GO terms. The following KEGG pathways were found to be significantly over-expressed in healthy controls vs. cases: *oxidative phosphorylation*, *ATP synthesis*, and *Ribosome*. Contrarily, *Insulin signalling pathway* was up-regulated in diseased cases. When GO terms were analyzed, we found as significantly up-regulated in healthy controls: *oxidative phosphorylation* (GO:0006119), *nucleotide biosynthesis* (GO:0009165) (biological process ontology), *NADH dehydrogenase (ubiquinone) activity* (GO:0008137), *nuclease activity* (GO:0004518) (molecular function ontology), and *mitochondrion* (GO:0005739) (cellular component ontology). Some of the terms were redundant with the KEGG pathways, although here we have also the *ubiquinone* class, which does not appear in KEGG. Since FatiScan implements more functional terms, we also analyzed Swissprot keywords and found *Ubiquinone*, *Ribosomal protein*, *Ribonucleoprotein*, *Mitochondrion*, and *Transit peptide* as over-expressed in healthy controls vs. cases. Other alternative methods give similar results. *Oxidative phosphorylation* and *mitochondrion* are found by GSEA (21), PAGE (30), and other statistics (27). *Nucleotide biosynthesis* can be assimilated to other datasets found by these three methods

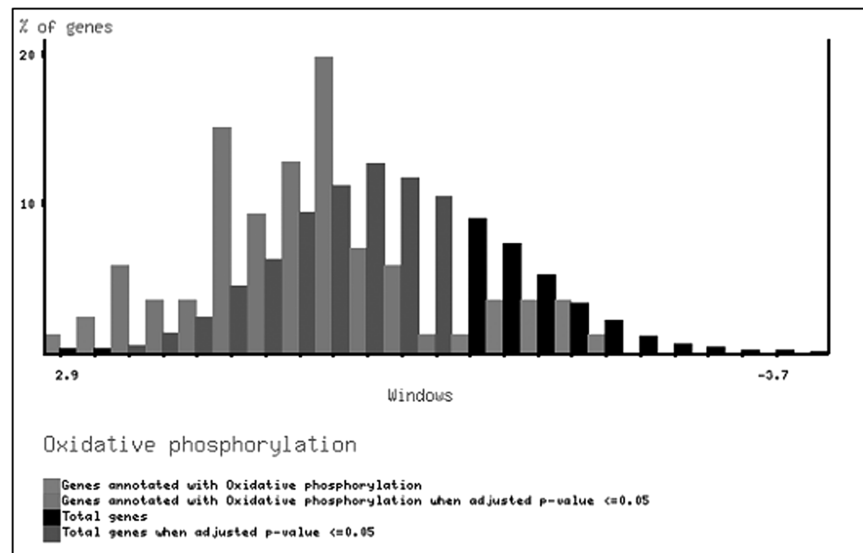


Fig. 12.3. Comparison between the background distribution of GO terms (*light gray bars*) and the distribution of *oxidative phosphorylation* GO term (*dark gray and black bars*). The last distribution is clearly shifted towards highest values of the t statistic (horizontal axis), corresponding to high expression in healthy controls. The transition of colors black to gray makes reference to the values of the t -statistic for which the partitions were found to be significant.

(21, 27, 30) based on a set of functional categories developed by (22). The rest of the terms were only found by FatiScan.

If, for example, the distribution of the GO term *oxidative phosphorylation* is compared to the background distribution of GO terms (Fig. 12.3) a clear trend to the over-expression of the complete pathway with high values of the t -statistic, corresponding to genes over-expressed in healthy controls, can be clearly observed.

4. Notes



1. Despite the fact that microarray technologies allow data on the behavior of all the genes of complete genomes to be obtained, hypotheses are still tested as if the problem consisted of thousands of independent observations. The tests applied involve only single genes and ignore all the available knowledge cumulated in recent years on their cooperative behavior, which is stored in several repositories (GO, KEGG, protein interaction databases, etc.). Since this process involves carrying out a large number of tests, severe corrections must be imposed to reduce the number of false-positives. Later, the genes selected using these procedures were functionally

analyzed in a second step. Much information is lost in both steps during this process. Recently, however, methods have been proposed that take into account the properties of the genes and address different biological questions not in a gene-centric manner, but in a function-centric manner for class comparison (21, 22, 24–27, 29, 30), class assignment (44) and class discovery (34–36).

A systems biology approach to the analysis of microarray data will in the future tend to use more information on the cooperative properties of the genes beyond their simple functional description. Thus it is expected that a deeper knowledge of the interactome or the transcriptional network of the genomes will contribute to more realistic biological questions being addressed by microarray experiments, resulting in more complete and accurate answers.

2. Given the number of steps necessary for the proper analysis of a microarray experiment (normalization, the analysis itself, and the functional interpretation), integrated packages are preferable in order to avoid problems derived from the change of formats. Among the most complete packages available on the web is GEPAS (32, 41–43) that offers various options for normalization, supervised and unsupervised analysis (<http://www.gepas.org>) and is coupled to the Babelomics suite (32, 33) for functional interpretation of genome-scale experiments (<http://www.babelomics.org>), in which various tests for two-steps of threshold-free functional interpretation are implemented.

Acknowledgments

This work is supported by grants from MEC BIO2005-01078, NRC Canada-SEPOCT Spain, and Fundación Genoma España.

References

1. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.
2. Hallikas, O., Palin, K., Sinjushina, N., et al. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* 124, 47–59.
3. Rual, J. F., Venkatesan, K., Hao, T., et al. (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.
4. Lee, H. K., Hsu, A. K., Sajdak, J., et al. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res* 14, 1085–1094.
5. Stuart, J. M., Segal, E., Koller, D., et al. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255.

6. van Noort, V., Snel, B., Huynen, M. A. (2003) Predicting gene function by conserved co-expression. *Trends Genet* 19, 238–242.
7. Mateos, A., Dopazo, J., Jansen, R., et al. (2002) Systematic learning of gene functional classes from DNA array expression data by using multilayer perceptrons. *Genome Res* 12, 1703–1715.
8. Westerhoff, H. V., Palsson, B. O. (2004) The evolution of molecular biology into systems biology. *Nat Biotechnol* 22, 1249–1252.
9. Golub, T. R., Slonim, D. K., Tamayo, P., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.
10. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25–29.
11. Kanehisa, M., Goto, S., Kawashima, S., et al. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32, D277–280.
12. Robertson, G., Bilenky, M., Lin, K., et al. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res* 34, D68–73.
13. Wingender, E., Chen, X., Hehl, R., et al. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28, 316–319.
14. Mulder, N. J., Apweiler, R., Attwood, T. K., et al. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res* 33, D201–205.
15. Draghici, S., Khatri, P., Martins, R. P., et al. (2003) Global functional profiling of gene expression. *Genomics* 81, 98–104.
16. Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20, 578–580.
17. Zeeberg, B. R., Feng, W., Wang, G., et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 4, R28.
18. Khatri, P., Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587–3595.
19. Bolshakova, N., Azuaje, F., Cunningham, P. (2005) A knowledge-driven approach to cluster validity assessment. *Bioinformatics* 21, 2546–2547.
20. Bammler, T., Beyer, R. P., Bhattacharya, S., et al. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* 2, 351–356.
21. Mootha, V. K., Lindgren, C. M., Eriksson, K. F., et al. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34, 267–273.
22. Subramanian, A., Tamayo, P., Mootha, V. K., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102, 15545–15550.
23. Damian, D., Gorfine, M. (2004) Statistical concerns about the GSEA procedure. *Nat Genet* 36, 663.
24. Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics* 21, 2988–2993.
25. Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., et al. (2005) Testing association of a pathway with survival using gene expression data. *Bioinformatics* 21, 1950–1957.
26. Goeman, J. J., van de Geer, S. A., de Kort, F., et al. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20, 93–99.
27. Tian, L., Greenberg, S. A., Kong, S. W., et al. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci U S A* 102, 13544–13549.
28. Smid, M., Dorssers, L. C. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics* 20, 2618–2625.
29. Vencio, R., Koide, T., Gomes, S., et al. (2006) BayGO: Bayesian analysis of ontology term enrichment in microarray data. *BMC Bioinformatics* 7, 86.
30. Kim, S. Y., Volsky, D. J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 6, 144.
31. Chen, Z., Wang, W., Ling, X. B., et al. (2006) GO-Diff: Mining functional differentiation between EST-based transcriptomes. *BMC Bioinformatics* 7, 72.
32. Al-Shahrour, F., Minguez, P., Tarraga, J., et al. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* 34, W472–476.
33. Al-Shahrour, F., Minguez, P., Vaquerizas, J. M., et al. (2005) BABELOMICS: a suite of web tools for functional annotation and

- analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res*, 33, W460–464.
- [Au1] 34. Huang, D., Pan, W. (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics* in press.
35. Pan, W. (2006) Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics* 22, 795–801.
36. Jia, Z., Xu, S. (2005) Clustering expressed genes on the basis of their association with a quantitative phenotype. *Genet Res* 86, 193–207.
37. Eisen, M. B., Spellman, P. T., Brown, P. O., et al. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863–14868.
38. Wolfe, C.J., Kohane, I. S., and Butte, A. J. (2005) Systematic survey reveals general applicability of “guilt-by-association” within gene coexpression networks. *BMC Bioinformatics* 6, 227.
39. Barry, W. T., Nobel, A. B., and Wright, F. A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 21, 1943–1949.
40. Benjamini, Y., Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29, 1165–1188.
41. Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., et al. (2003) GEPAS: A web-based resource for microarray gene expression data analysis. *Nucleic Acids Res* 31, 3461–3467.
42. Herrero, J., Vaquerizas, J. M., Al-Shahrour, F., et al. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res* 32, W485–491.
43. Vaquerizas, J. M., Conde, L., Yankilevich, P., et al. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res* 33, W616–620.
44. Lottaz, C., Spang, R. (2005) Molecular decomposition of complex clinical phenotypes using biologically structured analysis of microarray data. *Bioinformatics* 21, 1971–1978.

Author Query:

[Au1]: Please update?

Uncorrected Proof