

# Chapter 19

## Functional Profiling Methods in Cancer

Joaquín Dopazo

### Summary

The introduction of new high-throughput methodologies such as DNA microarrays constitutes a major breakthrough in cancer research. The unprecedented amount of data produced by such technologies has opened new avenues for interrogating living systems although, at the same time, it has demanded the development of new data analytical methods as well as new strategies for testing hypotheses. A history of early successful applications in cancer boosted the use of microarrays and fostered further applications in other fields.

Keeping the pace with these technologies, bioinformatics offers new solutions for data analysis and, what is more important, permits the formulation of a new class of hypotheses inspired in systems biology, more oriented to pathways or, in general, to modules of functionally related genes. Although these analytical methodologies are new, some options are already available and are discussed in this chapter.

**Key words:** Functional profiling, Functional enrichment, Gene-set analysis, Pathway, Gene ontology, Systems biology, Microarray

---

### 1. Introduction

Among the battery of high-throughput methodologies that are revolutionizing cancer research, DNA microarrays can be considered the standard due to their popularity and characteristics. Although many different questions can be addressed through microarray experiments, there are usually three types of objective in this context: “class comparison,” “class prediction,” and “class discovery” (1, 2). The first two objectives usually involve the application of tests to define differentially expressed genes, or the use 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 in the study.

When strategies for microarray data analysis are considered from a historical perspective, an initial period can be distinguished in which almost all publications were related to reproducibility and sensitivity issues. Many classic microarray papers dating from the late 1990s were mainly proof-of-principle experiments (3, 4). Consequently, the methodological approaches used for analysis were mainly related to clustering and, in general, unsupervised approaches. This has caused a subsequent confusion with respect to the choice of the appropriate methodology for a proper data analysis, as noted by some authors (5). Later, sensitivity became a main concern as a natural reaction against very liberal interpretations of microarray experiments, such as the fold criteria, to select differentially expressed genes. It was soon obvious that genome-scale experiments should be carefully analyzed because many apparent associations happened merely by chance (6). In this scenario, methods for the adjustment of p-values, which are considered standard today, started to be extensively used (7, 8). The increasingly use of microarrays as predictors of clinical outcomes (9), despite not being free of criticisms (5), fueled the use of the methodology because of its practical implications. Comparative studies show that, although intra-platform reproducibility seems to be high, cross-platform and cross-laboratory coherence is still an issue (10). Another aspect that soon became of major importance was the interpretation of microarray experiments in terms of their biological implications, rather than restricting them to a mere comparison of lists of gene identifiers (11, 12). Thus, a number of methods that essentially search for the overrepresentation of functional modules within groups of genes previously defined in the experiment were developed. Examples of repositories widely used to define gene modules are Gene Ontology (GO) (13), KEGG pathways (14), or Biocarta (<http://www.biocarta.com>). Programs such as GOMiner (15), FatiGO (16), etc., can be considered representatives of a family of methods that use these gene module functional definitions to conjecture about the interpretation of the results of microarray experiments (17).

The difficulties for defining repeatable lists of genes of interest across laboratories and platforms even using common experimental and statistical methods (18) 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) (19, 20) has pioneered a family of methods devised not to find individual genes but to search for groups of functionally related genes with a coordinate (although not necessarily high) overexpression or underexpression across a list of genes ranked by differential expression between two classes, compared in microarray experiments. Different tests have recently been proposed for microarray data,

with this aim in mind (12, 21–25) and also for expressed sequence tags (ESTs) (26), and some of them are available on web servers (12, 27). In particular, the FatiScan procedure (12, 27) can deal with ordered lists of genes independently from the type of data that originated them. This interesting property allows for its application to a broad range of experimental designs (case–control, multiclass, survival, etc.) as well as to other type of high-throughput data apart from microarrays.

Thus, in addition to the conventional study of individual genes and proteins, genome-wide approaches based on high-throughput methodologies have helped to uncover fundamental principles of tumorigenesis, and increasing evidence points to cooperative, systems-level events as important factors to understand the mechanisms by which cancer gene products coordinately promote cellular transformation (28, 29). Moreover, modern trends in the pharmaceutical industry also point toward the use of functional genomics and systems biology-oriented studies (30, 31) as fundamental steps of the drug discovery pipeline.

---

## 2. Materials

### **2.1. Definition of Gene Modules: Sources of Information**

Any functional analysis relies on the definition of gene modules related by biological properties of interest. Probably the most widely used source of definition of functional modules is the Gene Ontology (GO) catalog (13). GO represents the biological knowledge as a tree (more precisely as a directed acyclic graph [DAG], in which a node can have more than one parent) where functional terms near the root of the tree make reference to more general concepts while deeper functional terms near the leaves of the tree make reference to more specific concepts. If a gene is annotated to a given level then it is automatically considered to be annotated at all of the upper levels (all of the parent levels) up to the root. Because genes are annotated at different levels of the GO hierarchy, it is common to use this abstraction to choose a predefined level in the hierarchy instead of using directly the original levels of annotation of the genes (11, 32), which increases the power of the enrichment tests (11, 12, 33, 34).

The KEGG pathways database (14) or the Biocarta pathways (<http://www.biocarta.com>) are two extensively used sources of functional information. There are also databases that contain functional motifs mapped to proteins, such as the Interpro database (35) and many others.

In addition, other types of modules, such as transcriptional ones, can be defined as groups of genes under the same regulatory control.

Databases that collect regulatory motifs are available. Among the most popular are CisRed (36) and Transfac, which contains predictions of transcription factor binding sites (37). In addition, negative regulation mediated by microRNAs has recently gained relevance. The miRBase (38) contains putative gene targets of such microRNAs. Genes sharing one or more of these regulatory motifs can be considered a putative regulatory module.

Other ways of defining modules of different nature include the use of information obtained using text-mining procedures (39), chromosomal location (40, 41), protein–protein interactions, etc.

## **2.2. Bioinformatics Tools**

Beyond other technical or statistical considerations, the approximate level of acceptance of different gene-set analysis (GSA) methods among the scientific community is reported in **Tables 1** and **2**. **Table 1** presents an exhaustive list of bioinformatics tools available for functional profiling that implement tests for functional enrichment. Here the number of Scholar Google citations has been used as an approximate popularity index, given that it is reflecting the number of academic documents (mostly papers) citing a particular paper. Following this criterion, the most popular tools having more than 200 citations are EASE (42), DAVID (43), GOMiner (15), Babelomics/FatiGO (12, 16, 34), MAPP-Finder (44), GOStats (45), and Ontotools (46). In the case of GSA methods, **Table 2** shows that more than the 75% of the Scholar Google citations are monopolized by two tools: GSEA and Babelomics.

---

## **3. Methods**

### **3.1. Functional Enrichment Methods**

In the conventional approach for the functional annotation of microarray experiments, known as functional enrichment analysis, the functional interpretation of the data is performed in two steps: in a first step, genes of interest are selected using different procedures. In a subsequent step, the selected genes of interest are compared with a background (usually the rest of the genes) to find enrichment in any gene module. This comparison with the background is essential because an apparently high proportion of a given functional module could easily be nothing but a reflection of a high proportion of this particular module in the whole genome but not a proper enrichment. Actually, both enrichments and depletions of gene modules are potentially of interest. Therefore, unless there is a specific reason not to consider enrichment or depletion, two-sided tests are appropriate (47). This comparison between the selected genes and the background can be carried out

**Table 1**  
**Functional enrichment data analysis tools with at least ten Scholar Google citations**

Tool	Application type or URL for web servers	References	Citations <sup>a</sup>
EASE	Windows application	(42)	603
DAVID	<a href="http://www.DAVID.niaid.nih.gov">http://www.DAVID.niaid.nih.gov</a>	(43)	504
GOMiner	<a href="http://discover.nci.nih.gov/gominer/">http://discover.nci.nih.gov/gominer/</a>	(15, 55)	408
Babelomics	<a href="http://www.babelomics.org">http://www.babelomics.org</a>	(12, 16, 34, 50, 56)	402
MAPPFinder	<a href="http://www.GenMAPP.org">http://www.GenMAPP.org</a>	(44)	379
FatiGO	<a href="http://www.fatigo.org">http://www.fatigo.org</a>	(16)	341
GOSat	<a href="http://gostat.wehi.edu.au/">http://gostat.wehi.edu.au/</a>	(45)	249
Ontotools	<a href="http://vortex.cs.wayne.edu/ontoexpress/">http://vortex.cs.wayne.edu/ontoexpress/</a>	(32, 46, 57–59)	223
GOTM	<a href="http://genereg.ornl.gov/gotm/">http://genereg.ornl.gov/gotm/</a>	(60)	164
GO::TermFinder	Perl script	(61)	152
FunSpec	<a href="http://funspec.med.utoronto.ca">http://funspec.med.utoronto.ca</a>	(62)	100
GeneMerge	<a href="http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html">http://www.oeb.harvard.edu/hartl/lab/publications/GeneMerge.html</a>	(63)	96
FuncAssociate	<a href="http://llama.med.harvard.edu/Software.html">http://llama.med.harvard.edu/Software.html</a>	(64)	91
BINGO	Cytoscape plugin	(65)	75
GOToolBox	<a href="http://gin.univ-mrs.fr/GOToolBox">http://gin.univ-mrs.fr/GOToolBox</a>	(66)	74
GFINDER	<a href="http://www.medinfopoli.polimi.it/GFINDER/">http://www.medinfopoli.polimi.it/GFINDER/</a>	(67, 68)	49
WebGestalt	<a href="http://bioinfo.vanderbilt.edu/webgestalt/">http://bioinfo.vanderbilt.edu/webgestalt/</a>	(69)	46
GOSurfer	R package	(70)	45
CLENCH	Perl script	(71)	26
Pathway Explorer	<a href="https://pathwayexplorer.genome.tugraz.at/">https://pathwayexplorer.genome.tugraz.at/</a>	(72)	25
Ontology Traverser	R package	(73)	24
THEA	Java standalone	(74)	11
WebBayGO	<a href="http://blasto.iq.usp.br/~tkoide/BayGO/">http://blasto.iq.usp.br/~tkoide/BayGO/</a>	(75)	10
GOSat	R package	(76)	10

<sup>a</sup>Citations are taken from Scholar Google (by January 2008). Scholar Google is taken as an indirect estimation of the citation in papers but gives an idea on the impact in the scientific community

**Table 2**  
**Tools available for functional profiling by gene-set analysis with at least ten Scholar Google citations**

Tool	Application type or URL for web servers	References	Test	Citations <sup>a</sup>
GSEA	<a href="http://www.broad.mit.edu/gsea/">http://www.broad.mit.edu/gsea/</a>	(19, 20)	GS, C	1,013
Babelomics (FatiGO + FatiScan)	<a href="http://www.babelomics.org">http://www.babelomics.org</a>	(12, 16, 34, 50, 56)	FE/GS, C	402
FuncAssociate	<a href="http://llama.med.harvard.edu/Software.html">http://llama.med.harvard.edu/Software.html</a>	(64)	FE/GS, C	91
Global test	R package	(22)	GS, SC	89
PAGE	Python script	(25)	GS, C	42
ErmineJ	Java	(77)	GS, C	35
FatiScan	<a href="http://www.babelomics.org">http://www.babelomics.org</a>	(50)	GS, C	34
GO-mapper	Windows, Perl script	(24)	GS, C	33
SAFE	R package	(49)	GS, C	27
GOAL	<a href="http://microarrays.unife.it">http://microarrays.unife.it</a>	(78)	GS, C	25
Catmap	Perl script	(79)	GS, C	19
PLAGE	<a href="http://dulci.biostat.duke.edu/pathways/">http://dulci.biostat.duke.edu/pathways/</a>	(80)	GS, SC	18
GODist	Mathlab program	(81)	GS, SC	17
t-Profiler	<a href="http://www.t-profiler.org/">http://www.t-profiler.org/</a>	(82)	GS, C	12

Type of test: *GS* gene set; *C* Competitive; *FE* functional enrichment; *SC* self-contained

<sup>a</sup>Citations are taken from Scholar Google (by January 2008). Scholar Google is taken as an indirect estimation of the citation in papers but gives an idea on the impact in the scientific community

by means of the application of different tests, such as the hypergeometric, Fisher's exact test  $\chi^2$  and binomial, which are considered to give similar results (47). Because many tests are conducted to check all the gene modules, adjustment for multiple testing, such as false discovery rate (FDR) (7) or others, must be used.

### 3.2. Gene-Set Analysis Methods

The interpretation of a genome-scale experiment using the two-steps functional enrichment approach is far from being optimal given that the thresholds imposed in the first step assuming independence preclude the detection of many gene modules. Methods directly inspired in systems biology focus on collective properties

of the genes more than on individual gene expression values. Modules of genes related by common functionality, regulation, or other interesting biological properties will simultaneously fulfill their roles in the cell and, consequently, they are expected to display a coordinated expression.

In its simplest formulation, the GSA method uses a rank of values derived from the experiment analyzed. Mootha et al. (19) ranked the genes according to their differential expression when two predefined classes (diabetic versus healthy controls) were compared by means of any appropriate statistical test (48). The position of the genes (that cooperatively act to define modules) within this ranked list is related to its participation in the trait studied in the experiment. Consequently, each module that is a causative agent of the differences between the classes compared will be found in the extremes of the ranked list with highest probability. Thus, instead of testing differential activities of genes, which implicitly assumes independent behavior (an aspect often ignored by the researchers applying the test), and later searching for enrichment in gene modules among the selected genes, GSA directly tests for gene modules significantly cumulated in the extremes of a ranked list of genes. In this way, artificial previous thresholds, which inadvertently change the meaning of our hypothesis testing schema, is avoided.

Different methods have been proposed for this purpose, such as the GSEA (19, 20) or the SAFE (49) methods, which use a nonparametrical version of a Kolmogorov–Smirnov test. Other strategies proposed are the direct analysis of functional terms weighted with experimental data (24) or model-based methods (22). Methods with similar accuracy, although conceptually simpler and quicker, have also been proposed, for instance, the parametrical counterpart of the GSEA, the PAGE (25), or the segmentation test, Fatiscan (50).

### **3.3. Functional Profiling in Array–CGH Experiments**

Genetic alterations, such as losses (deletions), gains (amplifications), or losses of heterozygosity (LOH) of genetic material that affect certain regions of the genome, have been shown to be the basis of many types of cancer (51). New technologies such as array–CGH, along with the use of expression arrays, offer for the first time the opportunity to accurately characterize the alterations in genomic copy number and the dependence of gene expression on the alterations (52). Despite the obvious fact that such alterations affect a large number of genes, most of the research is still focused in finding only one or a few genes responsible for a disease or a trait and ignores the chromosomal context (52). In particular, the putative impact that the local distribution of functions could have in the symptomatology of diseases that harbor copy number alterations or, in general, could have in gene regulation and/or silencing is largely unexplored. Actually, only a few attempts of

analyzing copy number alterations in terms of gains or losses of whole or parts of gene teams have been made to date (40, 41). Programs such as ISACGH (41) detect copy number alterations using conventional algorithms and allow a functional enrichment analysis of the regions with detected alterations.

### **3.4. Gene-Set Analysis in Genotyping**

Another field in which a gene set-based approach could be very useful is genotyping. Association and linkage studies with chips with increasingly density result in a frustrating effect of decreasing the power of the tests, because of the strict corrections that must be applied to the tests. Most genetic disorders have a complex inheritance and can be considered the combined result of variants in many genes, each contributing only weak effects to the disease. Given that, in any disorder, most of the disease genes will be involved in only a few different molecular pathways, the knowledge of the relationships (functional, regulatory, interactions, etc.) between the genes can help in the assessment of possible candidates (which may reside in different loci) with a joint basis for the disease etiology. The use of different gene module definitions (GO, KEGG, protein interactions and coexpression) in an integrated network was recently applied to interrelate positional candidate genes from different disease loci and then to test 96 heritable disorders in the Online Mendelian Inheritance in Man database (53). This gene set-based strategy resulted in a 2.8-fold increase over random selection.

### **3.5. Conclusion**

As research in cancer is increasingly benefited by the introduction of high-throughput technologies, new hypotheses, inspired in systems biology concepts, can be addressed and checked (54). Bioinformatics has become an essential tool not only as a mere instrument for managing the huge amount of data produced by these new technologies, but to implement a new generation of algorithms and concepts that are opening the doors to the understanding of cancer as a system (28, 29). Biomedicine is becoming more computational and research in cancer is pioneering this transformation.

---

## **Acknowledgments**

This work is supported by grants from the Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), ISCIII, projects BIO 2008-04212 from the Spanish Ministry of Education and Science and National Institute of Bioinformatics (<http://www.inab.org>), a platform of Genoma España. EA is supported by a fellowship for the FIS of the Spanish Ministry of Health (FI06/00027).

## References

1. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
2. Allison, D.B., Cui, X., Page, G.P. and Sabripour, M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet*, **7**, 55–65.
3. Perou, C.M., Jeffrey, S.S., van de Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C., et al. (1999) Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc Natl Acad Sci USA*, **96**, 9212–9217.
4. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
5. Simon, R., Radmacher, M.D., Dobbin, K. and McShane, L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst*, **95**, 14–18.
6. Ge, H., Walhout, A.J. and Vidal, M. (2003) Integrating ‘omic’ information: a bridge between genomics and systems biology. *Trends Genet*, **19**, 551–560.
7. Benjamini, Y. and Yekutieli, D. (2001) The control of false discovery rate in multiple testing under dependency. *Ann Stat*, **29**, 1165–1188.
8. Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci USA*, **100**, 9440–9445.
9. 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., et al. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
10. Moreau, Y., Aerts, S., De Moor, B., De Strooper, B. and Dabrowski, M. (2003) Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet*, **19**, 570–577.
11. Al-Shahrour, F. and Dopazo, J. (2005) In Azuaje, F. and Dopazo, J. (eds.), *Data analysis and visualization in genomics and proteomics*. Wiley, pp. 99–112.
12. Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L. and Dopazo, J. (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–W464.
13. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–29.
14. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**, D277–D280.
15. Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, **4**, R28.
16. Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
17. Khatri, P. and Draghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
18. Bammmler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A., Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., et al. (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods*, **2**, 351–356.
19. Mootha, V.K., Lindgren, C.M., Eriksson, K.F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., et al. (2003) PGC-1 $\alpha$ -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, **34**, 267–273.
20. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, **102**, 15545–15550.
21. Goeman, J.J., Oosting, J., Cleton-Jansen, A.M., Anninga, J.K. and van Houwelingen, H.C. (2005) Testing association of a pathway with survival using gene expression data. *Bioinformatics*, **21**, 1950–1957.
22. Goeman, J.J., van de Geer, S.A., de Kort, F. and van Houwelingen, H.C. (2004) A global test for groups of genes: testing association

- with a clinical outcome. *Bioinformatics*, **20**, 93–99.
23. Tian, L., Greenberg, S.A., Kong, S.W., Altshuler, J., Kohane, I.S. and Park, P.J. (2005) Discovering statistically significant pathways in expression profiling studies. *Proc Natl Acad Sci USA*, **102**, 13544–13549.
  24. Smid, M. and 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.
  25. Kim, S.Y. and Volsky, D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
  26. Chen, Z., Wang, W., Ling, X.B., Liu, J.J. and Chen, L. (2006) GO-Diff: mining functional differentiation between EST-based transcripts. *BMC Bioinformatics*, **7**, 72.
  27. Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J.M., Conde, L., Blaschke, C., Vera, J. and Dopazo, J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*, **34**, W472–W476.
  28. Khalil, I.G. and Hill, C. (2005) Systems biology for cancer. *Curr Opin Oncol*, **17**, 44–48.
  29. Kitano, H. (2004) Cancer as a robust system: implications for anticancer therapy. *Nat Rev Cancer*, **4**, 227–235.
  30. Butcher, E.C. (2005) Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov*, **4**, 461–467.
  31. Searls, D.B. (2005) Data integration: challenges for drug discovery. *Nat Rev Drug Discov*, **4**, 45–58.
  32. Khatri, P., Sellamuthu, S., Malhotra, P., Amin, K., Done, A. and Draghici, S. (2005) Recent additions and improvements to the Onto-Tools. *Nucleic Acids Res*, **33**, W762–W765.
  33. Al-Shahrour, F., Arbiza, L., Dopazo, H., Huerta-Cepas, J., Minguez, P., Montaner, D. and Dopazo, J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.
  34. Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J.M., Conde, L., Blaschke, C., Vera, J. and Dopazo, J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res*, **34**, W472–W476.
  35. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., et al. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*, **33**, D201–D205.
  36. Robertson, G., Bilenky, M., Lin, K., He, A., Yuen, W., Dagpinar, M., Varhol, R., Teague, K., Griffith, O.L., Zhang, X., et al. (2006) cisRED: a database system for genome-scale computational discovery of regulatory elements. *Nucleic Acids Res*, **34**, D68–D73.
  37. Wingender, E., Chen, X., Hehl, R., Karas, H., Liebich, I., Matys, V., Meinhardt, T., Pruss, M., Reuter, I. and Schacherer, F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res*, **28**, 316–319.
  38. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, **34**, D140–D144.
  39. Minguez, P., Al-Shahrour, F., Montaner, D. and Dopazo, J. (2007) Functional profiling of microarray experiments using text-mining derived bioentities. *Bioinformatics*, **23**, 3098–3099.
  40. Conde, L., Montaner, D., Burguet-Castell, J., Tarraga, J., Al-Shahrour, F. and Dopazo, J. (2007) Functional profiling and gene expression analysis of chromosomal copy number alterations. *Bioinformatics*, **1**, 432–435.
  41. Conde, L., Montaner, D., Burguet-Castell, J., Tarraga, J., Medina, I., Al-Shahrour, F. and Dopazo, J. (2007) ISACGH: a web-based environment for the analysis of Array CGH and gene expression which includes functional profiling. *Nucleic Acids Res*, **35**, W81–W85.
  42. Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C. and Lempicki, R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol*, **4**, R70.
  43. Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C. and Lempicki, R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol*, **4**, P3.
  44. Doniger, S.W., Salomonis, N., Dahlquist, K.D., Vranizan, K., Lawlor, S.C. and Conklin, B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*, **4**, R7.
  45. Beissbarth, T. and Speed, T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
  46. Khatri, P., Bhavsar, P., Bawa, G. and Draghici, S. (2004) Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments. *Nucleic Acids Res*, **32**, W449–W456.
  47. Rivals, I., Personnaz, L., Taing, L. and Potier, M.C. (2007) Enrichment or depletion of a

- GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.
48. Cui, X. and Churchill, G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4**, 210.
  49. 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.
  50. Al-Shahrour, F., Diaz-Uriarte, R. and Dopazo, J. (2005) Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. *Bioinformatics*, **21**, 2988–2993.
  51. Albertson, D.G. and Pinkel, D. (2003) Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet*, **12**(Spec No 2), R145–R152.
  52. Pinkel, D. and Albertson, D.G. (2005) Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, **37**(Suppl), S11–S17.
  53. Franke, L., van Bakel, H., Fokkens, L., de Jong, E.D., Egmont-Petersen, M. and Wijmenga, C. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, **78**, 1011–1025.
  54. Kitano, H. (2002) Computational systems biology. *Nature*, **420**, 206–210.
  55. Zeeberg, B.R., Qin, H., Narasimhan, S., Sunshine, M., Cao, H., Kane, D.W., Reimers, M., Stephens, R.M., Bryant, D., Burt, S.K., et al. (2005) High-throughput GoMiner, an ‘industrial-strength’ integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*, **6**, 168.
  56. Al-Shahrour, F., Minguez, P., Tarraga, J., Medina, I., Alloza, E., Montaner, D. and Dopazo, J. (2007) FatiGO+: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res*, **35**, W91–W96.
  57. Draghici, S., Khatri, P., Bhavsar, P., Shah, A., Krawetz, S.A. and Tainsky, M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*, **31**, 3775–3781.
  58. Khatri, P., Desai, V., Tarca, A.L., Sellamuthu, S., Wildman, D.E., Romero, R. and Draghici, S. (2006) New Onto-Tools: Promoter-Express, nsSNPCounter and Onto-Translate. *Nucleic Acids Res*, **34**, W626–W631.
  59. Khatri, P., Voichita, C., Kattan, K., Ansari, N., Khatri, A., Georgescu, C., Tarca, A.L. and Draghici, S. (2007) Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res*, **35**, W206–W211.
  60. Zhang, B., Schmoyer, D., Kirov, S. and Snoddy, J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
  61. Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M. and Sherlock, G. (2004) GO::TermFinder – open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
  62. Robinson, M.D., Grigull, J., Mohammad, N. and Hughes, T.R. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.
  63. Castillo-Davis, C.I. and Hartl, D.L. (2003) GeneMerge – post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.
  64. Berriz, G.F., King, O.D., Bryant, B., Sander, C. and Roth, F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
  65. Maere, S., Heymans, K. and Kuiper, M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.
  66. Martin, D., Brun, C., Remy, E., Mouren, P., Thieffry, D. and Jacq, B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol*, **5**, R101.
  67. Masseroli, M., Galati, O. and Pinciroli, F. (2005) GFINDER: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res*, **33**, W717–W723.
  68. Masseroli, M., Martucci, D. and Pinciroli, F. (2004) GFINDER: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res*, **32**, W293–W300.
  69. Zhang, B., Kirov, S. and Snoddy, J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res*, **33**, W741–W748.
  70. Zhong, S., Storch, K.F., Lipan, O., Kao, M.C., Weitz, C.J. and Wong, W.H. (2004) GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl Bioinformatics*, **3**, 261–264.

71. Shah, N.H. and Fedoroff, N.V. (2004) CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, **20**, 1196–1197.
72. Mlecnik, B., Scheideler, M., Hackl, H., Hartler, J., Sanchez-Cabo, F. and Trajanoski, Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res*, **33**, W633–W637.
73. Young, A., Whitehouse, N., Cho, J. and Shaw, C. (2005) OntologyTraverser: an R package for GO analysis. *Bioinformatics*, **21**, 275–276.
74. Pasquier, C., Girardot, F., Jevardat de Fombelle, K. and Christen, R. (2004) THEA: ontology-driven analysis of microarray data. *Bioinformatics*, **20**, 2636–2643.
75. Vencio, R.Z., Koide, T., Gomes, S.L. and Pereira, C.A. (2006) BayGO: Bayesian analysis of ontology term enrichment in microarray data. *BMC Bioinformatics*, **7**, 86.
76. Falcon, S. and Gentleman, R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.
77. Lee, H.K., Braynen, W., Keshav, K. and Pavlidis, P. (2005) ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.
78. Volinia, S., Evangelisti, R., Francioso, F., Arcelli, D., Carella, M. and Gasparini, P. (2004) GOAL: automated Gene Ontology analysis of expression profiles. *Nucleic Acids Res*, **32**, W492–W499.
79. Breslin, T., Eden, P. and Krogh, M. (2004) Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, **5**, 193.
80. Tomfohr, J., Lu, J. and Kepler, T.B. (2005) Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, **6**, 225.
81. Ben-Shaul, Y., Bergman, H. and Soreq, H. (2005) Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression. *Bioinformatics*, **21**, 1129–1137.
82. Boorsma, A., Foat, B.C., Vis, D., Klis, F. and Bussemaker, H.J. (2005) T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res*, **33**, W592–W595.