W472–W476 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl172

BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments

Fátima Al-Shahrour[1], Pablo Minguez[1], Joaquín Tárraga[1,2], David Montaner[1,2], Eva Alloza[1], Juan M. Vaquerizas[1], Lucía Conde[1], Christian Blaschke[3], Javier Vera[4] and Joaquín Dopazo[1,2,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Autopista del Saler 16, E-46013, Valencia, Spain, [2]Functional Genomics Node, INB, CIPF, Autopista del Saler 16, E-46013, Valencia, Spain, [3]Bioalma, Ronda de Poniente, 4, 2-C, E-28760, Tres Cantos, Madrid, Spain and [4]INB—BSC, Jordi Girona 29, Edifici Nexus II, E-08034 Barcelona, Spain

Received February 14, 2006; Revised and Accepted March 20, 2006

ABSTRACT

We present a new version of Babelomics, a complete suite of web tools for functional analysis of genome-scale experiments, with new and improved tools. New functionally relevant terms have been included such as CisRed motifs or bioentities obtained by text-mining procedures. An improved indexing has considerably speeded up several of the modules. An improved version of the FatiScan method for studying the coordinate behaviour of groups of functionally related genes is presented, along with a similar tool, the Gene Set Enrichment Analysis. Babelomics is now more oriented to test systems biology inspired hypotheses. Babelomics can be found at http://www.babelomics.org.

INTRODUCTION

Genes do not operate alone in the cell, but in a sophisticated network of interactions that we only recently start to envisage (1–3). It is a long recognized fact that co-expressing genes tend to be playing some common roles in the cell (4,5) and recently there are evidences that functionally related genes map close in the genome, even in higher eukaryotes (6,7). Complex traits, including diseases are starting to be considered from a systems biology perspective (8). Because of this, there is a clear necessity for methods and tools which can help to understand genome-scale experiments (microarrays, proteomics and the like) from a systems biology perspective. The proper interpretation of the experiments require functional annotation, but this annotation must be done in a systems biology context, in which the collective properties of groups of genes are taken into account. With the popularisation of DNA microarray technologies a number of methods arise to compare the enrichment in functional terms shown in groups of genes defined in the experiments. Programs such as ontoexpress (9) or FatiGO (10) are representatives of a family of methods designed for this purpose (11). A problem related to the management of genome-scale data followed by the inspection of thousands of functional terms is that a large number of associations will appear simply by chance (12,13). The multiple testing problem (14) was addressed for the first time by FatiGO (10) although now is a standard among these type of tools (11).

The extensive availability of functional annotations of a reasonable quality, specially facilitated by the universal adoption of the Gene Ontology (GO) (15) controlled vocabulary and other related initiatives such as KEGG (16), Interpro (17) and the like has improved enormously the accuracy of the above mentioned procedures of functional annotation. But beyond this, the extensive annotation permits to take conceptually different approaches to the analysis of genome-scale experiments more based on systems biology criteria. Thus, instead first selecting important genes (according to some criteria such as differential expression and the like) and then analysing them in terms of their biological roles, some authors proposed to directly analyse the behaviour of blocks of functionally related genes. The Gene Set Enrichment Analysis (GSEA) (18,19), the FatiScan (13) or the global test (20) constitute examples of this type of approach.

Suites such as Babelomics (21) or onto-tools (22), which gathers in an integrated environment different possibilities for functional annotation, will be more and more demanded in the future as the necessity of a more detailed interpretation of genome-scale experiments becomes more obvious.

Babelomics, named after the tale 'The Babel library' (23), a masterpiece by the famous Argentinean writer Jorge Luís Borges, has been running for more than one year and

To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: jdopazo@cipf.es

individual parts of it, such as the FatiGO tool (10), have been running for >3 years.

BIOLOGICAL INFORMATION USED FOR FUNCTIONAL ANNOTATION

Curated repositories

Different repositories of functionally relevant biological information are available and can be used for the functional annotation of genome-scale experiments. In this new release of Babelomics we have collected information from different repositories for several model organisms (*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana*), which has been cross-referenced using Ensembl (24) identifiers. The repositories used are as follows:

(i) *GO* is, probably, the most successful among the initiatives for the standardization of the nomenclature of biological processes, molecular functions and subcellular location, its three main ontologies (15). GO represents the biological knowledge as a tree (more precisely as a directed acyclic graph, DAG, in which a node can have more that one parent). Upper nodes represent more general concepts and as the DAG is traversed towards deeper levels, the definitions are more and more precise (e.g. cell cycle > regulation of cell cycle > positive regulation of cell cycle and so on) Since genes are annotated at different levels it is common to use the inclusive analysis (11,25) instead of using directly the annotation of the genes at the deepest level possible. In the inclusive analysis a level of abstraction is chosen and genes annotated at deeper levels are assigned to this level. This increments the efficiency of the test because there are less terms to test and more genes per term, but the selection of the level is arbitrary. We have implemented here the Nested Inclusive Analysis (NIA), in which the test is done recursively until the deepest level in which significance is obtained and only this last level is reported. In this way both variables: efficiency of the test and highest precision in the term found are optimized.

(ii) *InterPro* (17) is a database of protein families, domains and functional sites in which identifiable features (motifs) found in known proteins can be applied to guess about the possible functionality of unknown protein sequences.

(iii) The *SwissProt* (26) database contains for each entry a field called keywords which contain a controlled vocabulary of words, many of them (although not all) with functional meaning.

(iv) *KEGG pathways* (16) is a collection of manually drawn pathway maps representing our knowledge on the molecular interaction and reaction networks for Metabolism, Genetic Information Processing, Environmental Information Processing, Cellular Processes and Human Diseases (http://www.genome.jp/kegg/kegg2.html).

(v) *Transcription factor (TF)-binding sites* predicted using Transfac®. TFs are assigned to genes if the corresponding predicted TF-binding site (TFBS) for that TF if found in the 10 kb 5′ region of the gene. Search is carried out by the Match program (27), using only high quality matrices and with a cut-off to minimize false positives, from the Transfac database (28). TFBSs are only available for human and mouse.

(vi) *CisRed* (29) is a database for conserved regulatory elements predicted in promoter regions using multiple discovery methods applied to sequence sets that include corresponding sequence regions from vertebrates. Motif significance is estimated by comparison to randomized sequence sets that are adaptively derived from target sequence sets. In theory, all the Transfac® predictions should be a subset of these regulatory elements, but in practice the overlap is not complete. For this reason we still keep the Transfac® predictions. In addition, CisRed tables are only available for humans.

(vii) *Gene expression in tissues*: Two repositories containing information of gene expression in different tissues have been used:

(a) SAGE Tag libraries from the Cancer Genome Anatomy Project. A total of 279 human libraries that belong to 29 different tissues and 190 mouse libraries from 26 tissues have been used. The data were taken from http://cgap.nci.nih.gov/SAGE.

(b) Genomics Institute of the Novartis Foundation data. A total of 79 human tissues and 61 mouse tissues with normal histology were downloaded from http://wombat.gnf.org/index.html and used here.

Generation of annotations from the biomedical literature

The curated repositories above mentioned contain valuable information but a large amount of biomedical knowledge is still communicated in the old fashioned way of research publications. This information can only be extracted from the text with text-mining methods. Modern text-mining technology is still far away from 'understanding' human language (30) but some important advances have been made to extract some factual information with sufficient reliability from the scientific literature to be useful.

For the analysis of the biomedical literature precise identification of key entities of interest, such as genes, proteins, chemical compounds and disease names is crucial to index and retrieve relevant documents. As the biomedical language and vocabulary is of great complexity and changes constantly the identification of entities, commonly known as named entity recognition, is a cumbersome task.

For the detection of genes, proteins and diseases a combination of dictionaries (e.g. EntrezGene or UniProt for genes and proteins and UMLS for diseases), heuristics based on hand crafted rules and statistical measures are used. Chemical compounds are extracted based on morphological criteria (using knowledge about chemical nomenclature) and dictionaries of common names for chemicals.

Here, relationships between different biomedical entities that are calculated based on co-occurrences in sentences (a co-occurrence is when two entities appear in the same sentence) were used. The calculation is based on how unlikely it is to observe a certain level of co-occurrences to happen

by chance (31). The more unlikely the observed event, the stronger the relation between the entities is valued by the system. Using this approximation, gene association networks can be created, not specifying the precise relationships between the genes but organizing the literature in a way that makes exploration a lot easier.

The data used here were taken from the almaKnowledgeServer (http://aks.bioalma.com).

ENSEMBL INDEX

In order to maintain this huge system of gene annotations an universal index has been adopted. A total of 179 tables of different biological annotations and gene identifiers for seven organisms have been linked to their Ensembl IDs. Although the use of an universal cross-reference has many advantages this is not free of problems. Any gene not annotated in Ensembl will be lost in the analysis. This, obviously will affect to a very small amount of genes and should not affect to any general functional conclusion obtained by analysing a large and significant number of genes.

STRATEGIES FOR ANNOTATION OF GENOME-SCALE EXPERIMENTS

Typical genome-scale experiments are annotated in two steps. Firstly, 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 and so on) and then the enrichment of any type of biologically relevant label in these genes is compared with the corresponding distribution of the label in the background (typically the rest of genes). There are different available tools, such as FatiGO (10) and others (11), that use GO terms (15) or different functional labels, such as KEGG pathways, SwissProt keywords and the like, available in packages such as the Babelomics suite (21). From a systems biology perspective, this way of annotating the experiments is far from being efficient. This has led several groups to propose a different approach based on directly selecting blocks of functionally related genes (13,19,20). The rationale of these new approaches relies on the fact that the final aim in a typical genome-scale experiment is finding a molecular explanation for a given macroscopic observation (e.g. which pathways are affected by the deprivation of glucose in a cell). In the two-steps approach described previously, genes with different behaviour are firstly selected, usually ignoring the fact that these genes are acting cooperatively in the cell and consequently their behaviours must be coupled to some extent. To achieve this, very stringent thresholds to reduce the false positives ratio in the results are usually imposed. Then, the lists so obtained are compared with the background as described above. This procedure causes a tremendous loss of information because a large number of false negatives are sacrificed in order to preserve a low ratio of false positives, and the nosier the data are, the worse this effect is. Systems biology oriented methods can use lists of genes arranged by any biological criteria (e.g. differential expression when comparing cases and healthy controls) and search for the distribution of blocks of functionally related genes across it. If a particular function is defining the arrangement it will cumulate towards the extremes of the arrangement. A nice example is the study of differential gene expression between diabetics cases and normal controls, where no one single gene was found to be differentially expressed (because of the noise of the system), but pathways such as oxidative phosphorilation were found to be significantly repressed in the diabetic cases (13,18).

GENE-BY-GENE SELECTION FOLLOWED BY FUNCTIONAL ANNOTATION

Babelomics implements different procedures for the functional annotation of sets of pre-selected of genes, based on any experimental measure. Since GEPAS (32–34) is connected to Babelomics it is straightforward to analyse relevant genes, which have been selected by differential expression, or because they are part of a class predictor, or they co-express in clusters and so on. As mentioned above, different biological labels have been used for testing functional enrichment when comparing the distribution of such labels between gene datasets of interest and their corresponding references or backgrounds. The following tools are available:

- *FatiGO+*. This tool constitutes the evolution of FatiGO (10). In addition to GO terms it can test simultaneously for KEGG pathways, Interpro motifs, SwissProt keywords, TFBSs and CisRed motifs. The distribution of any combination (or all) of the terms between two groups of genes can be simultaneously tested by means of a Fisher exact test. All the *P*-values are adjusted by FDR. It can also be used to test genes defined by chromosomal positions (thus integrating the functionality of the old GenomeGO module (34) which has now been discontinued). The functionality of the old modules FatiWise and TransFat (34) have been completely included here and, consequently both modules have been discontinued. For the case of GO terms, the NIA has been implemented. So, GO terms are automatically tested from level 3 to depth 9 and only the deepest significant term is reported for each branch.
- *FatiGO*. This tool has been in use for more than three years and has been described elsewhere (10,21,25). Owing to its popularity still remain as an independent module although much of its functionality is integrated in FatiGO+. FatiGO implements NIA too.
- *Tissues Mining Tool (TMT)*. This tool compares the pretabulated expression values of two lists of genes in a set of tissues (see above) and report the tisues in which the differences in expression of the genes of both lists are more extreme by using a *t*-test. The resultant *P*-values are adjusted by FDR. For details see (21).
- *MARMITE* (My Accurate Resource for Mining TExts). This is the equivalent to FatiGO+ using as biological information precomputed gene-bioentity co-occurrences obtained using the text-mining software almaKnowledgeServer (see above). MARMITE reports significant differences in the distribution of the scores gene-bioentity between the two lists compared using for this a Kolmogorov–Smirnov test. The module uses data of co-occurrences among human gene names (HUGO ids) and three bioentity categories: disease-associated words, chemical products and word

roots. As in the rest of tests of the modules of Babelomics, *P*-values are adjusted by FDR.

DIRECT ANNOTATION OF BLOCKS OF FUNCTIONALLY RELATED GENES

Babelomics implements two methods for functional annotation of genome-scale experiments which are based on the study of the behaviour of blocks of genes: FatiScan (13) and GSEA (19).

- GSEA test the coordinated over- or under-expression of sets of genes using a Kolmogorov–Smirnov test over a weighted summation. This allows to detect asymmetrical distributions of sets of genes (defined because they share some functional property) cumulated in the highest or lowest values of an arrangement of genes according its differential expression when two experimental conditions are compared (18,19). Significance is obtained by means of the permutation of the dataset of gene expression values. In the implementation presented here, more biological terms than in the original distribution (http://www.broad.mit.edu/gsea/index.html) can be used (GO, KEGG pathways, SwissProt keywords, Interpro motifs can be tested for seven organisms—see above—while TFBSs and CisRed motifs can be tested only for human).
- FatiScan implements a segmentation test which checks for asymmetrical distributions of biological labels associated to genes ranked in a list (13,21). Unique in this type of approaches, this test only needs the list of ordered genes and not the original data which generated the sorting. This means that can be applied to the study of the relationship of biological labels to any type of experiment whose outcome is an sorted list of genes. Since Babelomics is linked to GEPAS, genes sorted by differential expression between two experimental conditions can be studied, but also genes correlated to a clinical variable (such as the level of a metabolite) or even to survival (33,34). Moreover, other lists of genes ranked by any other experimental or theoretical criteria can be studied (e.g. genes arranged by physico-chemical properties, mutability, structural parameters and so on) in order to understand whether there is some biological feature (among the labels used) which is related to the experimental parameter studied.

CONCLUSIONS

Obtaining, for example, a list of genes differentially expressed between two experimental conditions is only half the way to the proper interpretation of a genome-scale experiment. The functional annotation of these genes is a key step that many times is not performed just because the lack of the appropriate tool. Babelomics can be considered one of the largest and most complete resources for the functional annotation of genome-scale experiments. It contains tools unique in its functionality. Moreover, the tight connection of Babelomics to the GEPAS package (32–34) makes of it an invaluable resource for the analysis of microarray data.

An effort for innovating the tools and the subjacent philosophy of the package, with the aim of providing the possibility of addressing the problem of the annotation from a systems biology perspective, has been made. Thus, a new tool that makes use of annotations extracted from Pubmed abstracts by means of text-mining procedures (the MARMITE) has been included. Moreover, in addition to modules for functional annotation of pre-selected sets of genes, such as FatiGO+, MARMITE or TMT, Babelomics includes a completely renewed version of FatiScan and the GSEA. These last modules allows finding blocks of functionally related genes with a coordinated behaviour in a genome-scale experiment.

ACKNOWLEDGEMENTS

This work is supported by grants from Fundació La Caixa, Fundación BBVA, MEC BIO2005-01078 and NRC Canada-SEPOCT Spain. The Functional Genomics node (INB) is supported by Genoma España. Funding to pay the Open Access publication charges for this article was provided by Genoma España.

Conflict of interest statement. None declared.

REFERENCES

1. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al*. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
2. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.
3. Hallikas,O., Palin,K., Sinjushina,N., Rautiainen,R., Partanen,J., Ukkonen,E. and Taipale,J. (2006) Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell*, **124**, 47–59.
4. Lee,H.K., Hsu,A.K., Sajdak,J., Qin,J. and Pavlidis,P. (2004) Coexpression analysis of human genes across many microarray data sets. *Genome Res*., **14**, 1085–1094.
5. Stuart,J.M., Segal,E., Koller,D. and Kim,S.K. (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
6. 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. *et al*. (2001) The human transcriptome map: clustering of highly expressed genes in chromosomal domains. *Science*, **291**, 1289–1292.
7. Hurst,L.D., Pal,C. and Lercher,M.J. (2004) The evolutionary dynamics of eukaryotic gene order. *Nature Rev. Genet*., **5**, 299–310.
8. Butcher,E.C., Berg,E.L. and Kunkel,E.J. (2004) Systems biology in drug discovery. *Nat. Biotechnol*., **22**, 1253–1259.
9. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
10. 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.
11. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
12. Ge,H., Walhout,A.J. and Vidal,M. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet*., **19**, 551–560.
13. 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.

14. Pollard,K.S., Dudoit,S. and van der Laan,M.J. (2004) *U.C. Berkeley Division of Biostatistics Working Paper Series*. Working Paper, 164.
15. 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. *Nature Genet.*, **25**, 25–29.
16. 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.
17. 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.
18. 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-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.
19. 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.
20. 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.
21. 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.
22. 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.
23. Borges,J. (2000) *Fictions.* Penguin Books Ltd, London.
24. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al.* (2005) Ensembl 2005. *Nucleic Acids Res.*, **33**, D447–D453.
25. Al-Shahrour,F. and Dopazo,J. (2005) In Azuaje,F. and Dopazo,J. (eds), *Data analysis and visualization in genomics and proteomics*. Wiley, West Sussex, UK, pp. 99–112.
26. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res.*, **34**, D187–D191.
27. Kel,A.E., Gossling,E., Reuter,I., Cheremushkin,E., Kel-Margoulis,O.V. and Wingender,E. (2003) MATCH: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.*, **31**, 3576–3579.
28. 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.
29. 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.
30. Erhardt,R.A.A., Schneider,R. and Blaschke,C. (2006) Introduction and status of text-mining techniques applied to biomedical text. *Drug Discov. Today*, in press.
31. Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.
32. Herrero,J., Al-Shahrour,F., Diaz-Uriarte,R., Mateos,A., Vaquerizas,J.M., Santoyo,J. and Dopazo,J. (2003) GEPAS: a web-based resource for microarray gene expression data analysis. *Nucleic Acids Res.*, **31**, 3461–3467.
33. Herrero,J., Vaquerizas,J.M., Al-Shahrour,F., Conde,L., Mateos,A., Diaz-Uriarte,J.S. and Dopazo,J. (2004) New challenges in gene expression data analysis and the extended GEPAS. *Nucleic Acids Res.*, **32**, W485–W491.
34. Vaquerizas,J.M., Conde,L., Yankilevich,P., Cabezon,A., Minguez,P., Diaz-Uriarte,R., Al-Shahrour,F., Herrero,J. and Dopazo,J. (2005) GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data. *Nucleic Acids Res.*, **33**, W616–W620.