W210–W213 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* *Published online 16 May 2010*
doi:10.1093/nar/gkq388

Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling

Ignacio Medina[1,2], José Carbonell[1], Luis Pulido[1,2], Sara C. Madeira[3,4], Stefan Goetz[1,2], Ana Conesa[1], Joaquín Tárraga[1,5], Alberto Pascual-Montano[6], Ruben Nogales-Cadenas[6], Javier Santoyo[1,2], Francisco García[1,5], Martina Marbà[1,5], David Montaner[1] and Joaquín Dopazo[1,2,5,*]

[1]Bioinformatics Department, Centro de Investigación Príncipe Felipe (CIPF), Autopista del Saler 16, [2]CIBER de Enfermedades Raras (CIBERER), Valencia, Spain, [3]Knowledge Discovery and Bioinformatics (KDBIO) group, INESC-ID, Lisbon, Portugal, [4]Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, [5]Functional Genomics Node, INB, CIPF, Valencia, Spain and [6]National Center for Biotechnology-CSIC, Madrid, Spain

Received February 21, 2010; Revised April 16, 2010; Accepted April 24, 2010

ABSTRACT

Babelomics is a response to the growing necessity of integrating and analyzing different types of genomic data in an environment that allows an easy functional interpretation of the results. Babelomics includes a complete suite of methods for the analysis of gene expression data that include normalization (covering most commercial platforms), pre-processing, differential gene expression (case-controls, multiclass, survival or continuous values), predictors, clustering; large-scale genotyping assays (case controls and TDTs, and allows population stratification analysis and correction). All these genomic data analysis facilities are integrated and connected to multiple options for the functional interpretation of the experiments. Different methods of functional enrichment or gene set enrichment can be used to understand the functional basis of the experiment analyzed. Many sources of biological information, which include functional (GO, KEGG, Biocarta, Reactome, etc.), regulatory (Transfac, Jaspar, ORegAnno, miRNAs, etc.), text-mining or protein–protein interaction modules can be used for this purpose. Finally a tool for the *de novo* functional annotation of sequences has been included in the system. This provides support for the functional analysis of non-model species. Mirrors of Babelomics or command line execution of their individual components are now possible. Babelomics is available at http://www.babelomics.org.

INTRODUCTION

High-throughput technologies such as transcriptomics (microarrays) proteomics, large-scale genotyping [genome wide association studies (GWAS)], next generation sequencing, etc., produce huge amounts of data of unfeasible interpretation without the application of automatic procedures for functional profiling (1). The idea behind this new version of Babelomics is to integrate primary (normalization, calls, etc.) and secondary [signatures, predictors, associations, Transmission/disequilibrium tests (TDTs), clustering, etc.] analysis tools within an multiple-purpose platform that allows relating some of these genomic data and/or interpreting them by means of different functional enrichment or gene set methods. Such interpretation is made not only using functional definitions [GO (2), KEGG (3), Biocarta, Interpro (4), reactome (5)] but also regulatory information [Transfac (6), Jaspar (7), ORegAnno (8)] protein–protein interactions (9), text-mining module definitions (10) and the possibility of producing *de novo* annotations through the Blast2GO system (11).

Babelomics (12,13) as well as Gene Expression Pattern Analysis Suite (GEPAS) (14,15) have been uninterruptedly running for more than 8 years. Currently, Babelomics have an average of more than 200 experiments analyzed per day, respectively, (http://bioinfo.cipf.es/webstats/babelomics/awstats.babelomics.bioinfo.cipf.es.html), distributed among many different countries (http://bioinfo.cipf.es/

*To whom correspondence should be addressed. Tel: +34 96 328 96 80; Fax: +34 96 328 97 01; Email: jdopazo@cipf.es

toolsusage). Since their last release, a total of 65 280 anonymous users and 4521 registered users have used these tools.

In terms of technology, Babelomics has been reengineered, speeded up and transformed to web services. Babelomics has a new interface that allows the definition of persistent sessions and asynchronous use (a program can be left running and come back later to see the results) through a queue system. Moreover, the complete program can be installed locally and their modules can now be independently invoked as command line programs and can be integrated into analysis pipelines.

STRUCTURE OF THE PROGRAM

The program is organized into different sections that are described below. Two of them are related to data input and data management. The rest of them are different analysis options for data analysis. Novelties with respect to previous versions are indicated at the corresponding sections.

Upload data

This section is new. Upload process has been separated from the tools. Different parsers check the integrity and format of the data. The data accepted are: expression microarrays (one-channel Affymetrix and Agilent, and two-channel Agilent, Genepic and generic), Array-CGH (Agilent), generic data matrices of expression, ArrayCGH and SNPs, lists of identifiers (gene, transcript, protein, SNP, functional terms, ranks), lists of annotations (gene annotation, extended annotations) and some other data such as dendrogram descriptions (in Newick format; (http://evolution.genetics.washington.edu/phylip/newicktree.html), Blast results, protein–protein interaction data and genotype data (in standard PED and MAP formats; http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml). The program documentation contains details on the formats used.

To ease the process, the data format can be uploaded according to the format, as described above, or according the tool to be used.

Preprocessing

This section allows preprocessing the data loaded in the previous section. Almost all the options (except normalization of Affymetrix and two-color Agilent and genepix arrays) are new. All the array types loaded in the previous section can be normalized by different methods. Limma (16) and affy packages from Bioconductor (17) are implemented.

This section also includes an editor which allows easy addition and/or modification of labels and tags that will be further used as class or category labels for gene selection, prediction, etc.

Other preprocessing facilities for normalized data are available, such as log-transformations, replicate merging, missing value imputation, etc.

A converter of identifiers is available with more than 80 cross-referenced identifiers for genes, proteins, transcripts, microarray probesets, pathways, functional annotations and regulatory regions.

Finally, a facility for obtaining the existing annotations for lists of gene identifiers is available. All the possible annotations used in Babelomics (functional, regulatory, etc.) can be used for this purpose.

Expression

This section corresponds to the GEPAS (14,15) functionality. It includes tests for differential expression [two- or multi-class comparison, survival analysis, correlation to continuous parameters or time/dosage series analysis (18)], methods for class prediction (19) such as SVM (support vector machines), KNN (k-nearest neighbors), Random Forest and Naive Bayes, with different feature selection methods [differential expression, genetic algorithms, principal component analysis (PCA)] and clustering methods for both samples and genes implementing the algorithms UPGMA (unweighted pair group method with arithmetic mean), SOTA (self organizing tree algorithm) (20), *K*-means and SOM (self organizing maps). We have also included biclustering methods (21).

The users of GEPAS will appreciate the novelties here, which are 'limma' methods for one-, two- and multi-class comparison, new multiple testing correction methods (Benjamini–Hochberg (22), Benjamini–Yekutieli (23), Bonferroni, Holm (24) and Hochberg (25)) for differential gene expression. The predictor module has novelties such as the addition of random forest (26) and Naive Bayes methods, the use of different algorithms for feature selection (see above), new parameter tuning options and new representations of the results with receiver operating characteristic curves with new metrics [area under the curve, root mean squared error, Matthews correlation coefficient and accuracy]. We have also added biclustering (21) as a new clustering method.

Genomics

This section is completely new. It implements SNP-based **genotyping** (27). This module can deal with GWAS case–control studies and carries out chi-square, Fisher and linear or logistic model tests. For trios, the program can carry out TDT. Functionality is taken from the PLINK program (28). Again, the results of the test can be analyzed by the functional analysis module (see below) and the novel pathway-based analysis (PBA) strategies can be applied (29). Population stratification can be analyzed by identity-by-state (IBS) (28) and PLINK documentation (http://pngu.mgh.harvard.edu/~purcell/plink/strat.shtml). This is a simple but potentially powerful approach to population stratification, which can use the whole genome SNP data.

Functional analysis

This module inherits the functionality of the previous version of Babelomics (12,13) although many novelties have been included. Apart from the functional enrichment methods such as the popular FatiGO (30), and gene set analysis methods such as the segmentation test (31) or the logistic regression model (32) (a new addition), other

testing strategies have been added. Thus, the Genecodis method (33) that finds concurrent annotations in ranked lists of genes is one of the new methods included. It is also possible to carry our PBA in GWAS experiments by means of the novel module Gesbap (29). Gene modules defined using text-mining derived functional annotations related to medical terms and chemical compounds can also be used for gene-set analysis (GSA) (10). Also another module uses gene expression data already available in databases to define tissue-specific or phenotype-specific gene expression profiles. These can be used to check the similarity of a particular experiment to the standard profiles of healthy or diseased tissues (34). Finally, the possibility of finding significant subnetworks of protein–protein interactions associated to the genomic experiment analyzed is included in an additional module (9) as another novelty.

Regarding the gene modules that can be used in order to produce a functional interpretation of the results, many possibilities are available, including functional definitions [GO (2), GOSlim, KEGG (3), Biocarta (http://www.biocarta.com/), Interpro (4), reactome (5)], regulatory information [Transfac (6), Jaspar (7), ORegAnno (8), miRNA target genes from miRBase (35)], protein–protein interactions (9) and text-mining module definitions (10). Additionally, the user can define their own gene modules by uploading them (see upload section) or by using the annotation tools available in this version of Babelomics (see Blast2GO below). Jaspar, ORegAnno, reactome and protein–protein interactions are new modules in this release.

Different filters can be used to test sub-selection of gene modules, excluding in this way superfluous tests that only will result in a reduction of the statistical power of the method used.

Finally, the popular Blast2GO (11) module can be used to produce annotations of genes of non-model organisms. Such annotations can be stored and further used to analyze genomic experiments.

Utilities

Several utilities are available related to facilitate the annotation of the genes or to produce different visualizations of the data. Thus, the new modules for annotation and identification conversion already described in the preprocessing section can also be found here. Several new utilities to produce graphical representation of the results (histograms and boxplots, cluster representations, PCA viewers and the GO hierarchy viewer) are also available. Supplementary Figure 1 shows some of such graphics.

Technical details

Babelomics is designed as a web application so it can work on any operating system: Windows, GNU/Linux and MacOS. Babelomics has been tested in many browsers including: Firefox 3.x, Safari 4.x, Chrome 2.x, Opera 9.x, IE 7.x. and IE 8.x. The code has almost entirely rewritten in Java (except some routines for normalization of microarrays that are in R), which constituted a speed up of almost 30x with respect to previous versions. All the modules are web services and can be used in command line. A convenient queue system has been implemented. Babelomics is running in a high-end cluster with 10 dedicated Intel XEON Quad-Core CPUs at 2.0 GHz (summing up a total of 40 cores) with a large amount of RAM (total 60 GB).

CONCLUSIONS

Today's Babelomics is a long-term project that started in 2001 with the publication of the clustering method SOTA (20) for microarray data analysis, followed by the popular FatiGO (30) for functional enrichment analysis, which are now a constituent part of Babelomics. Later, different methods (both functional enrichment and gene set enrichment) along with a number of gene module definitions were assembled as the prototype of the previous Babelomics (12,13) while different methods for gene expression data analysis (differential expression, predictors, clustering) give rise in parallel to the GEPAS (14,15) project. Both packages have become popular in their respective areas, being Babelomics the third most cited web-based tool for functional analysis (Supplementary Table 1) and GEPAS the most cited web tool for microarray data analysis (Supplementary Table 2).

This new version of Babelomics embeds the gene expression data analysis functionality of GEPAS within the functional profiling framework of the previous Babelomics and includes new modules for the analysis of genomic data (genotyping and genomic copy number alterations). Many new methods and new module definitions of different nature (functional, regulatory, phenotypic, etc.) have been included in this version.

The Babelomics project aims to provide the scientific community with an advanced set of methods for the integrated analysis of genomic data within the context of functional profiling analysis without renouncing to a user-friendly and intuitive use. As the Functional Genomics node of the Spanish Institute of Bioinformatics (INB; http://www.inab.org) and being part of the Spanish Network of Cancer (RTICC; http://www.rticcc.org) and the Network of Centres for Research in Rare Diseases (CIBERER, http://www.ciberer.es), we have a direct contact with researchers which provided us much of the feedback necessary to make of Babelomics a useful tool.

Although there are many tools for the functional profiling of high-throughput experiments (Supplementary Tables 1 and 2), Babelomics is a widely used tool which offers a combination of features and a degree of integration that makes it unique among other resources available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The CIBER de Enfermedades Raras and the INB are initiatives of the ISCIII.

FUNDING

Funding for open access charge: Spanish Ministry of Science and Innovation (MICINN) (projects BIO BIO2008-04212 and CEN-2008-1002). Red Temática de Investigación Cooperativa en Cancer (RTICC, partial) (RD06/0020/1019); Instituto de Salud Carlos III (MICINN).

Conflict of interest statement. None declared.

REFERENCES

1. Dopazo,J. (2009) Formulating and testing hypotheses in functional genomics. *Artif. Intell. Med.*, **45**, 97–107.
2. 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.
3. Kanehisa,M., Araki,M., Goto,S., Hattori,M., Hirakawa,M., Itoh,M., Katayama,T., Kawashima,S., Okuda,S., Tokimatsu,T. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.
4. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Buillard,V., Cerutti,L., Copley,R. *et al.* (2007) New developments in the InterPro database. *Nucleic Acids Res.*, **35**, D224–D228.
5. Vastrik,I., D'Eustachio,P., Schmidt,E., Joshi-Tope,G., Gopinath,G., Croft,D., de Bono,B., Gillespie,M., Jassal,B., Lewis,S. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.
6. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K. *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, **34**, D108–D110.
7. Bryne,J.C., Valen,E., Tang,M.H., Marstrand,T., Winther,O., da Piedade,I., Krogh,A., Lenhard,B. and Sandelin,A. (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
8. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.
9. Minguez,P., Gotz,S., Montaner,D., Al-Shahrour,F. and Dopazo,J. (2009) SNOW, a web-based tool for the statistical analysis of protein-protein interaction networks. *Nucleic Acids Res.*, **37**, W109–W114.
10. 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.
11. Conesa,A., Gotz,S., Garcia-Gomez,J.M., Terol,J., Talon,M. and Robles,M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
12. Al-Shahrour,F., Carbonell,J., Minguez,P., Goetz,S., Conesa,A., Tarraga,J., Medina,I., Alloza,E., Montaner,D. and Dopazo,J. (2008) Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic Acids Res.*, **36**, W341–W346.
13. 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.
14. 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.
15. Montaner,D., Tarraga,J., Huerta-Cepas,J., Burguet,J., Vaquerizas,J.M., Conde,L., Minguez,P., Vera,J., Mukherjee,S., Valls,J. *et al.* (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Res.*, **34**, W486–W491.
16. Smyth,G. (2005) In Gentleman,R., Carey,V., Dudoit,S., Irizarry,R. and Huber,W. (eds), *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. Springer, New York, pp. 397–420.
17. Gentleman,R.C., Carey,V.J., Bates,D.M., Bolstad,B., Dettling,M., Dudoit,S., Ellis,B., Gautier,L., Ge,Y., Gentry,J. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
18. Conesa,A., Nueda,M.J., Ferrer,A. and Talon,M. (2006) maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments. *Bioinformatics*, **22**, 1096–1102.
19. Medina,I., Montaner,D., Tarraga,J. and Dopazo,J. (2007) Prophet, a web-based tool for class prediction using microarray data. *Bioinformatics*, **23**, 390–391.
20. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
21. Madeira,S.C. and Oliveira,A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput. Biol. Bioinform.*, **1**, 24–45.
22. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.
23. Benjamini,Y. and Yekutieli,D. (2001) The control of false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.
24. Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, **6**, 65–70.
25. Hochberg,Y. (1988) A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, **75**, 800–802.
26. Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
27. McCarthy,M.I., Abecasis,G.R., Cardon,L.R., Goldstein,D.B., Little,J., Ioannidis,J.P. and Hirschhorn,J.N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
28. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
29. Medina,I., Montaner,D., Bonifaci,N., Pujana,M.A., Carbonell,J., Tarraga,J., Al-Shahrour,F. and Dopazo,J. (2009) Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res.*, **37**, W340–W344.
30. 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.
31. 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.
32. Sartor,M.A., Leikauf,G.D. and Medvedovic,M. (2008) LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, **25**, 211–217.
33. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
34. Sweet-Cordero,A., Mukherjee,S., Subramanian,A., You,H., Roix,J.J., Ladd-Acosta,C., Mesirov,J., Golub,T.R. and Jacks,T. (2005) An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nat. Genet.*, **37**, 48–55.
35. 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.