Published online 20 April 2015 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* ***W117–W121***
doi: 10.1093/nar/gkv384

Babelomics 5.0: functional interpretation for new generations of genomic data

Roberto Alonso[1,2], Francisco Salavert[1,3], Francisco Garcia-Garcia[1], Jose Carbonell-Caballero[1], Marta Bleda[4], Luz Garcia-Alonso[1], Alba Sanchis-Juan[5], Daniel Perez-Gil[5], Pablo Marin-Garcia[5], Ruben Sanchez[1,6], Cankut Cubuk[1], Marta R. Hidalgo[1], Alicia Amadoz[1], Rosa D. Hernansaiz-Ballesteros[1], Alejandro Alemán[1,3], Joaquin Tarraga[1], David Montaner[1], Ignacio Medina[7] and Joaquin Dopazo[1,2,3,6,*]

[1]Computational Genomics Department, Centro de Investigación Príncipe Felipe (CIPF), Valencia, 46012, Spain, [2]Computational Genomics Chair, Bull-CIPF, Valencia, 46012, Spain, [3]Bioinformatics of Rare Diseases (BIER), CIBER de Enfermedades Raras (CIBERER), Valencia, 46012, Spain, [4]Department of Medicine, University of Cambridge, School of Clinical Medicine, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK, [5]Fundación Investigación Clínico de Valencia-INCLIVA, Valencia, 46010, Spain, [6]Functional Genomics Node, (INB) at CIPF, Valencia, 46012, Spain and [7]HPC Services, University of Cambridge, Cambridge, CB3 0RB UK

Received January 31, 2015; Revised April 9, 2015; Accepted April 11, 2015

ABSTRACT

Babelomics has been running for more than one decade offering a user-friendly interface for the functional analysis of gene expression and genomic data. Here we present its fifth release, which includes support for Next Generation Sequencing data including gene expression (RNA-seq), exome or genome resequencing. Babelomics has simplified its interface, being now more intuitive. Improved visualization options, such as a genome viewer as well as an interactive network viewer, have been implemented. New technical enhancements at both, client and server sides, makes the user experience faster and more dynamic. Babelomics offers user-friendly access to a full range of methods that cover: (i) primary data analysis, (ii) a variety of tests for different experimental designs and (iii) different enrichment and network analysis algorithms for the interpretation of the results of such tests in the proper functional context. In addition to the public server, local copies of Babelomics can be downloaded and installed. Babelomics is freely available at: http://www.babelomics.org.

INTRODUCTION

Babelomics is an integrative web-based platform for the functional analysis of transcriptomic and genomic data. Named after the tale 'The Babel library' (1), a masterpiece by the famous Argentinean writer Jorge Luís Borges that describes an infinite library containing all the possible books, Babelomics has been running for more than 10 years, becoming rapidly a classic in the field of functional analysis. Its first version, published in 2005 (2), consisted of a collection of methods for functional enrichment analysis (3,4) based on different biologically relevant terms (GO; Gene Ontology, KEGG, etc.). Since then, Babelomics has released new versions that incorporated transcriptomics primary data analysis methods from the GEPAS (5–9) (a web tool discontinued by the end of 2012). The functionality of these new versions was complemented with more functional analyses, such as network analysis (10,11), or text-mining (12). Also, the possible data types were expanded to single nucleotide polymorphisms (SNPs) and thus Genome Wide Association Analysis (GWAS) could be carried out in Babelomics (13–15). In terms of software, Babelomics has evolved by adopting increasingly efficient web technologies. Thus, from the plain HTML of the initial versions (2,14), Babelomics was re-engineered to use SOAP web services and Web 2.0 technology features, such as AJAX in the 2008 release (13). Later, in the 2010 release (15), the backend was rewritten in Java while an extensive use of JavaScript at the client side was made. The continuous adoption of new

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 9632 9701; Email: jdopazo@cipf.es
Present addresses:
Luz Garcia-Alonso, European Bioinformatics Institute (EMBL-EBI), European Molecular Biology Laboratory, Wellcome Trust Genome Campus, Hinxton, UK.
Alicia Amadoz, Regulation of Gene Expression Laboratory, CIMA, Pamplona, Spain.
Rosa D. Hernansaiz-Ballesteros, Faculty of Life Sciences & Medicine, King's College London, London, UK.

technologies, such as the HTML5 standard and RESTful web services, has enabled the design of successive interfaces that allow asynchronous use, as well as the management of projects, jobs and user accounts.

Babelomics provides easy-to-use solutions for the most common scenarios of genomic and transcriptomics data analysis, offering the possibility of exploring the effects of alteration in gene expression levels or changes in genes sequences within a functional context (GO classes, interactome, etc.) Typical Babelomics users are laboratory researchers with no programming skills but have a hypothesis they want to test using their data. The use of Babelomics accelerates the discovery process in the laboratories and reduces the routine work load in the bioinformatics and biostatistics facilities thus contributing to the optimization of the whole research 'ecosystem'. Babelomics has also been extensively used in courses. Currently, an average of more than 200 experiments per day are analysed in Babelomics, (http://bioinfo.cipf.es/webstats/babelomics/awstats.babelomics.bioinfo.cipf.es.html), distributed across many different countries (http://bioinfo.cipf.es/toolsusage). The four Babelomics publications have received a total of 522 citations (recorded by 21 January 2015, according to Thomson Reuters' Web of Science, https://webofknowledge.com).

As a response to the changes in the technologies this new version of Babelomics includes the possibility of analysing genomic and transcriptomics data from new generation sequencing (NGS) experiments. Thus, in addition to conventional microarray data, RNA-seq data and massive resequencing data can now be uploaded and analysed. A detailed analysis of use during the last four years has been used for discontinuing a number of options that have been scarcely used. This has resulted in a simplified, more intuitive and more sustainable interface for Babelomics. New advanced options for visualization have been implemented. These include a network viewer for the representation of the results of network analysis, which allows users to customize the results and to produce high quality figures for papers. Also a genome browser, Genome Maps (16), which allows visualizing SNPs or variants in their genomic context, was included.

On the other hand, huge datasets from high-throughput technologies bring about new challenges for data analysis and visualization. To keep pace with this data revolution Babelomics web interface has been redesigned and rewritten using new web technologies. Also a radical restructuring has been done at the server side to speed it up and make the analysis faster and more efficient.

We have reduced the dependence on many external databases, which made the update of the relevant information difficult in previous versions. Babelomics now uses CellBase (17) as unique source of information. CellBase currently resides at the European Bioinformatics Institute (EBI) and it is updated on regular basis.

Summarizing, Babelomics 5.0 includes support for new genomic data from NGS experiments and new analysis options. From a technological point of view it includes new visual and web technologies that provide a more robust, fast and interactive interface.

BABELOMICS STRUCTURE

Babelomics is structured in four conceptually different parts represented in the main menu: *Processing*, *Expression*, *Genomics, Cancer* and *Functional*. In addition, the menu bar contains the data entry point, *Upload*, the *Jobs* manager and the question mark icon that contains the tutorial, contact and credits.

Babelomics relies on a series of powerful resources developed by us in the last years. Now, CellBase (17) provides all the functional information required in the different steps of analysis via highly optimized RESTful web services (see https://github.com/opencb/cellbase/wiki). Innovative visualization interfaces have been implemented in Babelomics. An interactive and highly efficient genome browser, Genome Maps (16), allows representing variants (or any other genomic feature) in its genomic context. Network analysis results are now visualized in a new interactive visual framework, CellMaps (https://github.com/opencb/cell-maps/wik), which can produce high-quality figures customized by the user.

Data upload and WorkSpace

Babelomics can be used either in anonymous mode or as a registered user. In anonymous mode, all the uploaded data and the results obtained (but not saved in the user's terminal) are lost at the end of the session. In registered user mode the options are the same, the only difference is that registered users can maintain the data and the results in the Babelomics workspace with a limit of 10 GB (that can be changed in local installations of Babelomics). Registration is free. The workspace structure has improved with respect to previous versions and has the familiar appearance and functionality of the typical file system. The upload option of the main menu brings about the WorkSpace, where data files can be uploaded. Data files can also be uploaded from within any analysis option of the menu. Different analysis options in the menu can have specific format requirements. As a general rule, data consist on raw sequencing (VCF or counts) or microarray (.CEL, etc.) files in the first steps (Preprocessing). In subsequent analysis steps, the files are taken properly formatted from the previous steps. Nevertheless, files can be preprocessed and analysed with other tools and uploaded at later analysis steps providing they are properly formatted.

Data processing

Microarray normalization contains the same options that Babelomics 4 offered, which includes support for Affymetrix normalization and both, one-channel and two-channel normalization for Agilent and Genepix.

Regarding RNA-seq normalization, we have included an automatic decision rule to suggest the most suitable normalization method, depending on the potential biases detected in the data. The main factors that can originate biases are: library depth (irrelevant for samples from the same library), gene length and extreme differences in mRNA abundances. Thus, if a clear mRNA composition bias is detected, *Trimmed Means of M-values* (TMM) normalization (18) is

recommended, otherwise *Reads Per Kilobase per Million* (RPKM) method (19) is preferable. Noiseq (20) is used to carry out bias detection and the result produces the preselection of the normalization method (that obviously can be changed by the user).

In addition, there is an improved option for attribute edition that allows editing variable and label names in the data. Another option allows several transformations over the data matrices (including normalization, logarithm transformation, missing value imputation, etc.)

Expression data analysis

Typical expression data analyses include unsupervised analysis (clustering) and supervised analysis (differential expression or classifiers). These microarray data analyses are the same as in Babelomics 4. Actually, previous Babelomics versions implemented new clustering methods especially devised for clustering large datasets, such as the SOTA (21) and pioneered the implementation of classifiers in web tools (22).

The supervised analyses can be carried out with RNA-seq data as well. The method used for differential expression in RNA-seq data is different from the tests used for microarrays, given the different statistical distributions followed by both data types. RNA-seq counts are transformed with the *Voom* method (23) that allows subsequent linear analysis using *limma* (24). As in the case of microarray differential expression tests, different multiple-test correction methods are available.

Genomic data analysis

This module aims to give support to simple case/control or transmission disequilibrium test (TDT) experiments in Genome Wide Association Studies (GWAS) in Babelomics 4. The popular PLINK software (25) is used to carry out the tests. The results include a Manhattan plot, a list of SNPs and, below, a new graphical interface, provided by an embedded version of the genome viewer Genome Maps (16), that allows exploring significant SNPs or variants in its genomics context. In this new version we have also included one extensively used burden test for the analysis of sequence data, the Combined Multivariate and Collapsing (CMC) method (26). In particular, the regions defined here are genes (given that the most common NGS data is still produced by exome sequencing), that can be further analysed in the *Functional* data analysis module.

Cancer

In the last years, cancer genomics has experienced a data generation revolution. The completion of two large international initiatives, the Cancer Genome Atlas (27) (http://cancergenome.nih.gov/) and the International Cancer Genome Consortium (28) (https://icgc.org/) has made available a huge amount of genomic data. Thus, whole exome and genome sequencing of cancer samples is becoming mainstream. Here we have integrated two popular tools specifically devised for the analysis of cancer genomic sequences. One of them is Oncodrive-FM (29), which computes a metric of functional impact using three well-known methods, SIFT (30), PolyPhen (31) and MutationAssessor (32). This metric is used to detect potential cancer driver genes by studying how the functional impact of variants found across several tumour samples deviates from a null distribution. The other tool, OncodriveCLUST (33), aims to identify genes undergoing mutations that tend to be clustered instead of being evenly distributed within them. This method is designed to exploit the observation that mutations in cancer genes, especially oncogenes, often cluster in particular positions of the protein. Both methods allow detecting genes of potential relevance in cancer, within variant files (in VCF format), which can be further analysed in its functional context in the *Functional* data analysis module.

Functional data analysis

The differential aspect of Babelomics with respect to other similar tools is that any result in terms of (often not very informative) lists of genes with *P*-values obtained in any of the analysis modules above described (*Expression*, *Genomics* and *Cancer*) can internally be submitted to the *Functional* data analysis module where they can be interpreted within different functional contexts. A simple way to assess the possible functional roles played by a list of genes consists on studying the distribution of functional annotations associated to these genes. GO (34) is the most extensively used source of functional annotations for genes. *Single Enrichment* methods study over-representations on any of these GO terms in the resulting lists obtained in previous analysis modules (e.g. differentially expressed genes, genes containing SNPs or variants associated to the disease, etc.) The popular *FatiGO* algorithm (3), already present in early Babelomics versions, implements the single enrichment method also in this release. Similarly, the *Gene Set Enrichment* algorithm, now common to both, *Expression* (4) and *Genomic* (35) data, is included in the new Babelomics as well. In a similar way, lists of genes can be interpreted in the context of the interactome. Thus, *Network Enrichment* analysis finds the largest significant network that can be formed with the genes contained in a list (10). On the other hand, *Gene set network enrichment* analysis uses an ordered list of genes to find the network significantly associated to the highest values of the list (typically the lowest *P*-values of a test from any of the analysis modules) (11). Gene Set Enrichment methods are known to be more sensitive than Single Enrichment methods (36,37). At present, eight species representative of the main organism models are supported: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae*. More species will be supported in the future, although users can upload their own annotations for other species.

A new visual framework, CellMaps (https://github.com/opencb/cell-maps/wik), is provided in this new Babelomics version. This framework provides a smart and interactive representation of the resulting networks that allows users to produce customized high-quality figures, ready for the publisher.

BEHIND THE SCENE: THE BABELOMICS SERVER AND THE CLIENT

On the server side, some algorithms have been rewritten in C/C++ in order to take now advantage of the modern multi-core CPU capabilities of the hardware in which Babelomics runs. Also, the results of some analysis have been indexed to speed up queries from the web site. On-fly-time GZIP conversion has also improved data transfer for both downloads and uploads. Web services have been re-implemented using RESTful web services, which are lighter and have a low latency. CellBase (17), the database where Babelomics extracts information from, has been redesigned and transformed into a NoSQL database. This provides a high-performance and scalable solution to cope with increasingly complex and bigger queries. CellBase was originally born as an internal relational database to store biological data for Babelomics and was later released as independent project (17). During the last year, Babelomics and EBI developers have worked together to migrate CellBase to MongoDB, one of the most successful NoSQL databases. MongoDB is a high-performance and scalable document-oriented database, that makes easy to add big and complex data and provides a rich API to execute complex queries.

The client has been completely rewritten using new technologies and standards. Babelomics is now entirely implemented in HTML5 and makes use of other technologies such as Scalable Vector Graphics (SVG) for visualization or IndexedDB for caching information and minimizing the queries to the server. As part of this effort Babelomics has significantly contributed to develop Genome Maps and CellMaps, which have been integrated to offer high quality visualization capabilities to Babelomics users. Also, new Web Components standard is being used now. This allows the community to reuse most of the visual components developed. Due to the intensive use of cutting-edge web technologies only modern web browsers are fully supported, these include Chrome 36+, Firefox 36+, Safari 8+ and Opera 24+.

Apart from using the public version of Babelomics, the code is open and freely available for local installation. The code can be found in GitHub: https://github.com/babelomics/babelomics.

FUTURE PROSPECTS

The roadmap of Babelomics for the near future includes: more conventional tests for relatively common experimental designs not fully covered in this version, more context for the functional analysis and full integrative analysis of data and information. An obvious step ahead in functional analysis will be the inclusion of pathway analysis. We are working on the integration of models of signalling pathways (38,39) in future versions of Babelomics. Another important aspect is the regulatory information. Methods for inferring the regulatory circuit behind a transcriptomics experiment already developed (40) will also be included. Additionally, integrative analysis of complex experimental designs, in which genomic and transcriptomics data are simultaneously obtained, will be included. Finally, support for more species will soon be added. Since the functional information relies on CellBase (17), the inclusion of more species is straightforward.

CONCLUSIONS

Babelomics has evolved again and in this fifth version it incorporates next generation sequencing data, new analysis options and new technologies at both, server and client sides. It offers a user-friendly environment that provides a full range of solutions which include primary data analysis, followed by a bunch of test for different experimental designs and data types and completed with the possibility of testing the biological relevance of the results obtained within a functional context. One of the most distinctive features of Babelomics, the functional analysis of genomic data, is nowadays a critical aspect in data analysis. Given the multigenic nature of most traits, these can only be explained as the result of complex interactions between genes (41), a notion proposed more than a decade ago in the context of systems biology (42). Consequently, most diseases are better understood as failures of functional modules caused by different combinations of mutated genes rather than by unique mutation(s) in one single gene (43). The use of a Systems Biology perspective in the analysis of genomic data is leading to new approaches in biomedical research including diagnostics (44), drug discovery (45), as well as pharmacology and toxicology (46). The availability of a tool that offers user-friendly solutions to most of the conventional genomic analysis problems, complemented with an advanced functional analysis, within an environment that allows storing data and results, explains the success of Babelomics for more than a decade.

ACKNOWLEDGEMENTS

Part of this work has been carried out in the context of the HPC4G initiative (http://www.hpc4g.org) and the Bull-CIPF Chair for Computational Genomics. We are indebted to Nuria Lopez-Bigas for kindly providing us access to her Oncodrive tools.

FUNDING

Spanish Ministry of Economy and Competitiveness [BIO2011-27069], Conselleria d'Educacio of the Valencian Community [PROMETEOII/2014/025]; EU FP7-PEOPLE Project MLPM [316861]; Fundació la Marató TV3 [151/C/2013]. Funding for open access charge: Spanish Ministry of Economy and Competitiveness [BIO2011-27069].

Conflict of interest statement. None declared.

REFERENCES

1. Borges,J.L. (2000) *Fictions*. Penguin Books Ltd., London.
2. 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.
3. 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.

4. 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.
5. 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.
6. 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.
7. 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.
8. Tarraga,J., Medina,I., Carbonell,J., Huerta-Cepas,J., Minguez,P., Alloza,E., Al-Shahrour,F., Vegas-Azcarate,S., Goetz,S., Escobar,P. *et al.* (2008) GEPAS, a web-based tool for microarray data analysis and interpretation. *Nucleic Acids Res.*, **36**, W308–W314.
9. 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.
10. 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.
11. Garcia-Alonso,L., Alonso,R., Vidal,E., Amadoz,A., de Maria,A., Minguez,P., Medina,I. and Dopazo,J. (2012) Discovering the hidden sub-network component in a ranked list of genes or proteins derived from genomic experiments. *Nucleic Acids Res.*, **40**, e158.
12. 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.
13. 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.
14. 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.
15. Medina,I., Carbonell,J., Pulido,L., Madeira,S.C., Goetz,S., Conesa,A., Tarraga,J., Pascual-Montano,A., Nogales-Cadenas,R., Santoyo,J. *et al.* (2010) Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.*, **38**, W210–W213.
16. Medina,I., Salavert,F., Sanchez,R., de Maria,A., Alonso,R., Escobar,P., Bleda,M. and Dopazo,J. (2013) Genome Maps, a new generation genome browser. *Nucleic Acids Res.*, **41**, W41–W46.
17. Bleda,M., Tarraga,J., de Maria,A., Salavert,F., Garcia-Alonso,L., Celma,M., Martin,A., Dopazo,J. and Medina,I. (2012) CellBase, a comprehensive collection of RESTful web services for retrieving relevant biological information from heterogeneous sources. *Nucleic Acids Res.*, **40**, W609–W614.
18. Robinson,M.D. and Oshlack,A. (2010) A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.*, **11**, R25.
19. Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
20. Tarazona,S., Garcia-Alcalde,F., Dopazo,J., Ferrer,A. and Conesa,A. (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–2223.
21. Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.
22. 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.
23. Law,C.W., Chen,Y., Shi,W. and Smyth,G.K. (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.
24. Ritchie,M., Phipson,B., Wu,D., Hu,Y., Law,C. and Shi,W. and GK, S. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, doi:10.1093/nar/gkv007.
25. 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.
26. Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.
27. The_Cancer_Genome_Atlas_Research_Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
28. Hudson,T.J., Anderson,W., Artez,A., Barker,A.D., Bell,C., Bernabe,R.R., Bhan,M.K., Calvo,F., Eerola,I., Gerhard,D.S. *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.
29. Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.
30. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
31. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.*, **30**, 3894–3900.
32. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
33. Tamborero,D., Gonzalez-Perez,A. and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
34. 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.
35. 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.
36. Dopazo,J. (2009) Formulating and testing hypotheses in functional genomics. *Artif. Intell. Med.*, **45**, 97–107.
37. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput. Biol.*, **8**, e1002375.
38. Sebastian-Leon,P., Carbonell,J., Salavert,F., Sanchez,R., Medina,I. and Dopazo,J. (2013) Inferring the functional effect of gene expression changes in signaling pathways. *Nucleic Acids Res.*, **41**, W213–W217.
39. Sebastian-Leon,P., Vidal,E., Minguez,P., Conesa,A., Tarazona,S., Amadoz,A., Armero,C., Salavert,F., Vidal-Puig,A., Montaner,D. *et al.* (2014) Understanding disease mechanisms with models of signaling pathway activities. *BMC Syst. Biol.*, **8**, 121.
40. Bleda,M., Medina,I., Alonso,R., De Maria,A., Salavert,F. and Dopazo,J. (2012) Inferring the regulatory network behind a gene expression experiment. *Nucleic Acids Res.*, **40**, W168–W172.
41. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
42. Kitano,H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662–1664.
43. Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.
44. Hood,L., Heath,J.R., Phelps,M.E. and Lin,B. (2004) Systems biology and new technologies enable predictive and preventative medicine. *Science*, **306**, 640–643.
45. Dopazo,J. (2014) Genomics and transcriptomics in drug discovery. *Drug Discov. Today*, **19**, 126–132.
46. Bai,J.P. and Abernethy,D.R. (2013) Systems pharmacology to predict drug toxicity: integration across levels of biological organization. *Annu. Rev. Pharmacol. Toxicol.*, **53**, 451–473.