Published online 20 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D1005–D1010***
doi:10.1093/nar/gkq1184

NCBI GEO: archive for functional genomics data sets—10 years on

Tanya Barrett*, Dennis B. Troup, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Rolf N. Muertter, Michelle Holko, Oluwabukunmi Ayanbule, Andrey Yefanov and Alexandra Soboleva

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892, USA

Received September 15, 2010; Revised November 1, 2010; Accepted November 3, 2010

ABSTRACT

A decade ago, the Gene Expression Omnibus (GEO) database was established at the National Center for Biotechnology Information (NCBI). The original objective of GEO was to serve as a public repository for high-throughput gene expression data generated mostly by microarray technology. However, the research community quickly applied microarrays to non-gene-expression studies, including examination of genome copy number variation and genome-wide profiling of DNA-binding proteins. Because the GEO database was designed with a flexible structure, it was possible to quickly adapt the repository to store these data types. More recently, as the microarray community switches to next-generation sequencing technologies, GEO has again adapted to host these data sets. Today, GEO stores over 20 000 microarray- and sequence-based functional genomics studies, and continues to handle the majority of direct high-throughput data submissions from the research community. Multiple mechanisms are provided to help users effectively search, browse, download and visualize the data at the level of individual genes or entire studies. This paper describes recent database enhancements, including new search and data representation tools, as well as a brief review of how the community uses GEO data. GEO is freely accessible at http://www.ncbi.nlm.nih.gov/geo/.

INTRODUCTION

In 2000, the Gene Expression Omnibus (GEO) database was established by the National Center for Biotechnology Information at the National Library of Medicine (1). GEO was originally built to archive the burgeoning volumes of high-throughput gene expression data beginning to be produced by the research community at that time. While GEO was originally established to host gene expression data, the database has evolved to host other data types, including comparative genomic analyses, chromatin immunoprecipitation profiling that characterizes genome–protein interactions, non-coding RNA profiling, SNP genotyping and genome methylation status analyses. Table 1 provides a breakdown of the study types and technologies hosted by GEO.

As microarrays became used routinely in almost every area of biomedical research, the Minimum Information About a Microarray Experiment (MIAME) guidelines were proposed (2). These guidelines outline the minimum information that should be included when describing a microarray experiment to ensure that the data can be interpreted by others. In 2002, the Nature journals announced that they supported this proposal and that authors were henceforth required to deposit microarray data in either the GEO or ArrayExpress (3) public repositories so that anyone could freely access and critically evaluate the data discussed in manuscripts (4). Many other journals similarly adopted this requirement. It is for this reason that GEO and ArrayExpress have experienced remarkable growth. Today, GEO archives approximately 20 000 studies comprising 500 000 samples, 33 billion individual abundance measurements, for over 1300 organisms, submitted by 8000 laboratories from around the world and supporting data for over 10 000 published manuscripts.

Since its inception, many aspects of the GEO database and operating procedures have undergone major revisions and development, including:

(i) database modifications to support evolving data types,

*To whom correspondence should be addressed. Tel: +1 301 402 8693; Fax: +1 301 480 0109; Email: barrett@ncbi.nlm.nih.gov

Table 1. List of study types and the number of *Series* records with those types

Study type	Number of *Series*
Expression profiling by array	17 812
Expression profiling by genome tiling array	303
Expression profiling by high throughput sequencing	131
Expression profiling by SAGE	206
Expression profiling by MPSS	21
Expression profiling by RT–PCR	25
Non-coding RNA profiling by array	341
Non-coding RNA profiling by genome tiling array	81
Non-coding RNA profiling by high throughput sequencing	233
Genome binding/occupancy profiling by array	70
Genome binding/occupancy profiling by genome tiling array	835
Genome binding/occupancy profiling by high throughput sequencing	238
Genome variation profiling by array	309
Genome variation profiling by genome tiling array	406
Genome variation profiling by SNP array	269
Methylation profiling by array	46
Methylation profiling by genome tiling array	115
Methylation profiling by high throughput sequencing	30
Protein profiling by protein array	31
SNP genotyping by SNP array	149

Users can retrieve studies of a particular type using the 'DataSet Type' field in the *GEO DataSets* query interface.

(ii) increasingly stringent submission requirements, concomitant with developing community standards like MIAME,
(iii) enhanced submission formats that ease the burden on submitters and promote well-annotated MIAME-compliant data and
(iv) improved indexing and analysis tools that help users more easily locate information relevant to their interests.

A timeline summarizing database growth and major developments over the last 10 years, including recent enhancements, is provided in Figure 1.

DATABASE STRUCTURE AND DATA FLOW

The overarching design and organization of GEO remains largely unchanged from previous descriptions (5). *Platform*, *Sample* and *Series* records continue to be the core submitter-supplied objects, while *DataSet* records represent curated studies that form the basis of GEO's advanced data display and analysis tools. Information describing the objects, their content and relationships is provided at http://www.ncbi.nlm.nih.gov/geo/info/overview.html.

Primary database: submitter-supplied *Platform*, *Sample* and *Series* records

A *Platform* record is composed of a summary description of the array or sequencer and, for array-based *Platforms*, a data table defining the array template. A *Sample* record is composed of a description of the biological material, the experimental protocols to which it was subjected and a data table containing abundance measurements for each feature on the corresponding *Platform* table. A *Series* record defines a set of related *Samples* considered to be part of a study and describes the overall study aim and design.

We continue to take advantage of the flexible nature of the primary database which facilitates capture of a wide variety of rapidly developing heterogeneous data types. The tabular data from original submissions are not fully normalized in the database but rather are stored as plain text tab-delimited tables. While these tables can in principle contain any number of rows and columns, a suite of validation checks performed at the time of upload ensures that they meet basic format and content requirements, which in turn facilitates extraction of core elements to the secondary database. The validation rules can be easily modified to quickly accommodate novel data types and evolving curation standards. Corresponding raw data files are linked from each record and stored on FTP servers. *Series* summaries, biological sample descriptions, treatments and protocol metadata are stored in specified fields within database tables and have appropriate relations and restrictions. Unique and stable accession numbers are assigned to each record.

Secondary databases: curated *DataSets* and *Profiles*

Although the submitter-supplied objects in the primary database are very heterogeneous with regards to content, style and level of metadata detail, the array-based expression data generally share a common set of elements that can be extracted and further processed. These include:

(i) sequence identity tracking information of each feature on the *Platform*,
(ii) normalized expression measurements and
(iii) text describing the biological source and experimental aim.

This information is extracted from the submitter-supplied records and organized into an upper-level object called a GEO *DataSet*. A *DataSet* represents a summarized collection of consistently processed experimentally related *Sample* records categorized according to experimental variables. GEO *Profiles* are derived from GEO *DataSets*. A GEO *Profile* consists of the expression measurements for an individual gene across all *Samples* in a *DataSet*. Genes in GEO *Profiles* are periodically re-annotated according to the latest information in Entrez Gene and UniGene, an important consideration given the dynamic nature of gene annotations. *DataSets* and *Profiles* are a means for transforming diverse styles of incoming data into a relatively standardized format upon which downstream data analysis and data display tools are built.

SUBMISSION PROCEDURES, FORMATS AND STANDARDS

Much effort continues to be invested into making the deposit process as simple as possible for submitters,

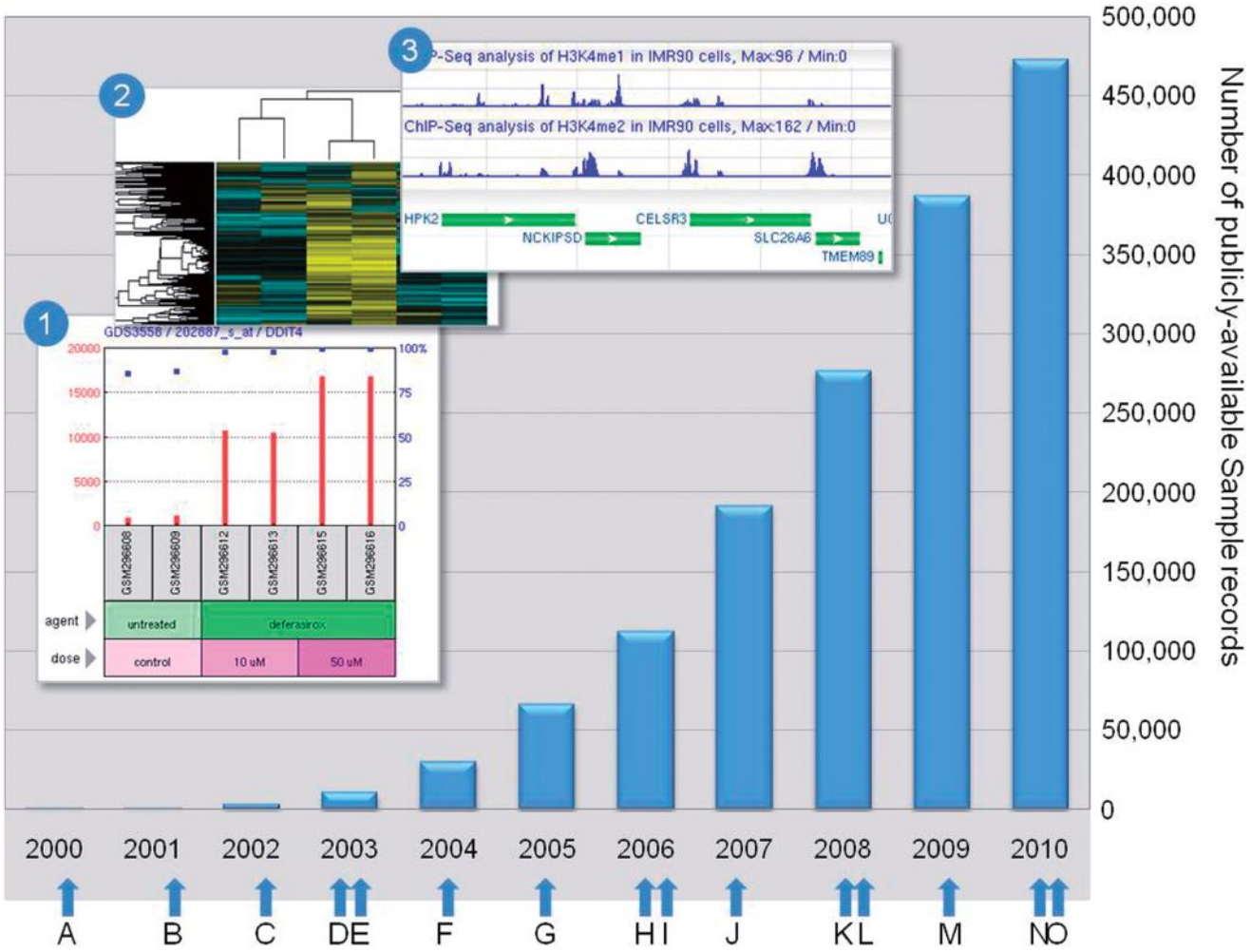

Figure 1. A timeline of GEO database growth, development and events. The chart represents accumulative growth of publicly-available *Sample* records from 2000 to September 2010. A further 80 000 Samples are currently held private until published, making a total of about 550 000 *Samples*. The current rate of submission and processing is over 10 000 *Samples* per month. (A) First data uploaded to database. (B) MIAME proposal is published, outlining the minimum information that should be included when describing a microarray experiment (2). (C) Nature journals announce requirement for microarray data deposit to public databases (4). (D) Reviewer access mechanism enabled, allowing anonymous confidential review of pre-published data. (E and inset 1) *GEO Profiles* database released, enabling search and visualization of individual gene expression charts. (F and inset 2) Interactive pre-computed cluster heatmaps released, allowing users to view and select regions of interesting gene expression patterns. (G) Major database modifications released aimed at better support of MIAME elements. (H) GEO increases enforcement of provision of raw data. (I) Bioconductor GEOquery package published, allowing GEO data to be imported into R environment (22). (J) GEOarchive spreadsheet submission format released, enabling rapid batch deposit of data. (K) All GEO *Series* records re-classified according to technology and experiment type making it simple to locate studies of a specific type; types are listed in Table 1. (L) Improvements to DataSet Browser and accompanying analysis tools panel implemented. (M and inset 3) First release of next-generation sequence tracks on NCBI's Sequence Viewer. These tracks were generated in support of the NIH Roadmap Epigenomics project, http://www.ncbi.nlm.nih.gov/geo/roadmap/epigenomics/. (N) Links generated to NCBI's new Epigenomics resource (7) which applies advanced curation and genome browser tracks for hundreds of next-generation sequence *Samples* derived from GEO. (O) Advanced Search tool released, helping users construct complex queries.

while not compromising the quality of the information supplied. A spreadsheet-based format called GEOarchive has emerged as GEO's most popular deposit method for both array and next-generation sequence submissions. GEOarchive spreadsheets are easy to assemble, and enable rapid batch deposit of both large and small studies. Generic GEOarchive templates are provided, as well as templates tailored for specific data types. Web-based submission forms are also supported, as are plain text and XML formats. All these submission methods can capture all components of the MIAME checklist. The GEO curation team is available to support submitters deposit their data should the need arise.

All submissions are syntactically validated for correct document structure, organization and provision of basic MIAME elements. Curators also review the content of each submission, aiming to ensure that sufficient information has been supplied to allow general interpretation of the study. When any problems are identified, such as corrupt files or missing components, a curator will work with the submitter until the issues are resolved. Despite these checks, it should be noted that submitters have ultimate responsibility for the content and accuracy of

their records. If submitters notice errors or want to make adjustments to their records, they are free to do so at any time using online update tools.

Submitters can keep their records in private status until a manuscript describing the data is published. Submitters are invited to create a reviewer URL and include it with their manuscript when they send it to journals for review. This URL allows anonymous, confidential access to the private GEO records cited in the paper, enabling reviewers to download and critically evaluate the data.

Next-generation sequence data submissions

GEO accepts next-generation sequence data for studies that examine gene expression, genome–protein interactions, genome methylation and gene regulation. Submissions are organized similarly to array-based submissions and all standard GEO administrative procedures are applied, full details are provided at http://www.ncbi.nlm.nih.gov/geo/info/seq.html. GEO hosts the processed and analyzed sequence data, together with sample and study metadata; raw data files are brokered to and linked with NCBI's Sequence Read Archive (SRA) database (6), ensuring that these sequence data are integrated with NCBI's collection of sequence-specific resources.

METHODS TO EXPLORE AND RETRIEVE DATA

Given the huge volumes and wide scope of experiment types represented in GEO, it is crucial to provide robust tools that allow users to easily locate and access information relevant to their interests.

Query

Queries may be performed at the *GEO DataSets* and *GEO Profiles* search interfaces (http://www.ncbi.nlm.nih.gov/gds and http://www.ncbi.nlm.nih.gov/geoprofiles, respectively). A simple keyword search is often sufficient to locate relevant data. However, GEO data are extensively indexed under many separate fields, meaning that users can refine their search by constructing fielded queries. *Advanced Search* pages have recently been developed for the *GEO DataSets* and *GEO Profiles* databases. These pages greatly assist users to construct complex multi-part fielded search statements. Alternatively, users can write and execute their own search statements directly in the search boxes. A full listing of all indexed fields, with examples and tips, is provided at http://www.ncbi.nlm.nih.gov/geo/info/qqtutorial.html. Examples of indexed fields include DataSet Type, Gene Symbol, Gene Ontology (GO) terms, Chromosome and Base Position. The search fields are not restricted to descriptive metadata, they can also be used to locate genes flagged as having interesting gene expression patterns. For example, to find genes that exhibit differential expression with respect to age or development stage in mouse, users could query *GEO Profiles* with:

```
(age[Flag Information] OR development stage[Flag Information]) AND mouse[Organism].
```

Additional query approaches are provided on *GEO DataSet* records to help users locate genes of interest. These include a tool that allows the user to perform *t*-tests with varying levels of significance on selected sets of *Samples*, as well as pre-computed cluster heatmap images that may be queried for specific genes. Sequence-based queries may be performed using the GEO BLAST interface.

Browse

The *DataSet Browser* is located at http://www.ncbi.nlm.nih.gov/sites/GDSbrowser.

Download

Complete GEO records and raw data files are freely available for bulk download from the GEO FTP site. High-throughput file transfer through Aspera Connect is also supported. Data are available in multiple formats including tab-delimited tables, plain text and XML. Raw data files are provided in native format where possible. Additional mechanisms are provided to enable download of more focused sets of data. For example, URLs can be constructed to retrieve metadata files for specific records and annotated gene expression values can be downloaded from *GEO Profiles* results pages. Download options are detailed at http://www.ncbi.nlm.nih.gov/geo/info/download.html.

Programmatic access

Programmatic access to metadata is supported through a suite of NCBI programs called the Entrez Programming Utilities. These tools allow users to develop pipelines that locate and download data of interest.

Data display

Various graphical renderings are provided to help users identify data of interest. *GEO Profile* charts portray the expression behavior of one gene across all *Samples* in a *DataSet*. The charts present several categories of information including expression measurement values, expression measurement rankings and an outline of the experimental design and variables. A user can look at a chart and quickly assess whether gene activity is shifting according to experimental conditions. Display of entire *DataSets* is facilitated through several types of precomputed cluster heatmaps. These clusters are interactive in that the user can search for specific genes, or highlight specific areas using a moveable selector box, and then download those regions as a text file or link to corresponding *GEO Profiles*. Next-generation sequence data presents new challenges for data display. Hundreds of GEO ChIP-seq and RNA-seq *Samples* have been selected for re-processing by NCBI's new Epigenomics resource (7) at http://www.ncbi.nlm.nih.gov/epigenomics. Once migrated to the Epigenomics resource, these data undergo advanced curation and are made available for review as precomputed tracks on genome browsers.

Links

Extensive internal and external links are applied to GEO data. Internal links help the user discover related gene expression *Profiles*. For example, *Profile Neighbor* links retrieve genes that show a similar gene expression pattern to the chosen *Profile* within a *DataSet*, *Chromosome Neighbor* links connect genes that are physically close to each other on the chromosome to help investigate gene expression neighborhoods, and *Sequence Neighbor* links gather *Profiles* from across all *DataSets* that are related by nucleotide sequence similarity. Inter-database links provide reciprocal connections to corresponding information in other NCBI databases including Gene, PubMed and GenBank (8).

HOW DOES THE COMMUNITY USE GEO DATA?

As the number of GEO data sets increases, so does the opportunity to aggregate and use the data in a much broader context than that for which it was originally intended. GEO typically receives over 40 000 web hits and 10 000 bulk FTP downloads per day. A review of the literature reveals over 1000 publications in which researchers describe how they applied data they found in GEO to their own studies, see http://www.ncbi.nlm.nih.gov/geo/info/ucitations.html. Many of these publications report how GEO data were used to verify gene expression signatures of individual genes in specific cell types or under specific conditions, providing evidence to support conclusions in the authors' own study (9). Other authors incorporate GEO data into their own analyses, such as using GEO data as controls or training sets within their own study (10). The huge volumes of data present bioinformaticians with a wealth of test material for developing new statistical algorithms (11) or analysis strategies, including modeling gene networks, predicting gene function and regulation (12) and identifying disease and toxicity predictors (13). Others have attempted to define systems-wide relationships and maps between GEO expression data and other data types such as gene pathways (14,15) and ontologies (16). Many subject-specific and/or added-value databases have been created using GEO data, targeting specific audiences, or presenting the data in alternative formats (17–19). Despite the heterogeneity of data types and experimental designs, many users have been able to perform powerful meta-analyses across thousands of independently-submitted GEO *Samples* (20,21). It can be anticipated that as public high-throughput functional genomic data sets continue to accumulate, third-party users will exploit these data in ever-more innovative ways.

CONCLUSIONS

Over the past decade, NCBI's GEO database has served as the major public archive for high-throughput microarray- and sequence-based functional genomic data sets. Aside from archiving, cross-linking and making vast amounts of data freely available for download, GEO also provides several user-friendly web-based tools and strategies to assist users query and analyze these data.

While GEO represents a unifying resource for thousands of valuable public functional genomics studies, significant challenges remain in achieving maximum benefit from these data. Those challenges include enabling better integration and cross-comparison of diverse data sets, data types and data resources and correlating the data with phenotype information. GEO will continue to be developed and refined towards achieving these goals, as well as improving the accessibility and usability of the data for as broad an audience as possible.

FUNDING

Funding for open access charge: The Intramural Research Program of the National Institutes of Health, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
2. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.
3. Parkinson,H., Kapushesky,M., Kolesnikov,N., Rustici,G., Shojatalab,M., Abeygunawardena,N., Berube,H., Dylag,M., Emam,I., Farne,A. *et al.* (2009) ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.
4. Microarray standards at last. (2002) *Nature*, **419**, 323. Available at http://www.nature.com/nature/journal/v419/n6905/full/419323a.html.
5. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.
6. Shumway,M., Cochrane,G. and Sugawara,H. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.
7. Fingerman,I.M., McDaniel,L., Zhang,X., Ratzat,W., Hassan,T., Jiang,Z., Cohen,R.F. and Schuler,G.D. (2011) NCBI Epigenomics: a new public resource for exploring epigenomic datasets. *Nucleic Acids Res.*, **39**, D908–D912.
8. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
9. Sanchez-Navarro,I., Gamez-Pozo,A., Pinto,A., Hardisson,D., Madero,R., Lopez,R., San Jose,B., Zamora,P., Redondo,A., Feliu,J. *et al.* (2010) An 8-gene qRT-PCR-based gene expression score that has prognostic value in early breast cancer. *BMC Cancer*, **10**, 336.
10. Hou,J., Aerts,J., den Hamer,B., van Ijcken,W., den Bakker,M., Riegman,P., van der Leest,C., van der Spek,P., Foekens,J.A., Hoogsteden,H.C. *et al.* (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS ONE*, **5**, e10312.
11. Tuna,S. and Niranjan,M. (2010) Reducing the algorithmic variability in transcriptome-based inference. *Bioinformatics*, **26**, 1185–1191.

12. Gustafsson,M. and Hornquist,M. (2010) Gene expression prediction by soft integration and the elastic net-best performance of the DREAM3 gene expression challenge. *PLoS ONE*, **5**, e9134.
13. Patel,C.J. and Butte,A.J. (2010) Predicting environmental chemical factors associated with disease-related gene expression data. *BMC Med. Genomics*, **3**, 17.
14. Towfic,F., VanderPlas,S., Oliver,C.A., Couture,O., Tuggle,C.K., West Greenlee,M.H. and Honavar,V. (2010) Detection of gene orthology from gene co-expression and protein interaction networks. *BMC Bioinformatics*, **11(Suppl. 3)**, S7.
15. Duarte,N.C., Becker,S.A., Jamshidi,N., Thiele,I., Mo,M.L., Vo,T.D., Srivas,R. and Palsson,B.O. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc. Natl Acad. Sci. USA*, **104**, 1777–1782.
16. de Tayrac,M., Le,S., Aubry,M., Mosser,J. and Husson,F. (2009) Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: multiple Factor Analysis approach. *BMC Genomics*, **10**, 32.
17. Lopez,F., Textoris,J., Bergon,A., Didier,G., Remy,E., Granjeaud,S., Imbert,J., Nguyen,C. and Puthier,D. (2008) TranscriptomeBrowser: a powerful and flexible toolbox to explore productively the transcriptional landscape of the Gene Expression Omnibus database. *PLoS ONE*, **3**, e4001.
18. Yu,Y., Tu,K., Zheng,S., Li,Y., Ding,G., Ping,J. and Hao,P. (2009) GEOGLE: context mining tool for the correlation between gene expression and the phenotypic distinction. *BMC Bioinformatics*, **10**, 264.
19. Vazquez,M., Nogales-Cadenas,R., Arroyo,J., Botias,P., Garcia,R., Carazo,J.M., Tirado,F., Pascual-Montano,A. and Carmona-Saez,P. (2010) MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures. *Nucleic Acids Res.*, **38(Suppl.)**, W228–W232.
20. Huang,H., Liu,C.C. and Zhou,X.J. (2010) Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. *Proc. Natl Acad. Sci. USA*, **107**, 6823–6828.
21. Suthram,S., Dudley,J.T., Chiang,A.P., Chen,R., Hastie,T.J. and Butte,A.J. (2010) Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets. *PLoS Comput. Biol.*, **6**, e1000662.
22. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.